Biometrics in a data stream context

Paulo Henrique Pisani

SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito:

Assinatura: _____

Paulo Henrique Pisani

Biometrics in a data stream context

Doctoral dissertation submitted to the Instituto de Ciências Matemáticas e de Computação - ICMC-USP, in partial fulfillment of the requirements for the degree of the Doctorate Program in Computer Science and Computational Mathematics. *FINAL VERSION*

Concentration Area: Computer Science and Computational Mathematics

Advisor: Prof. Dr. André Carlos Ponce de Leon Ferreira de Carvalho

USP – São Carlos
April 2017

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados fornecidos pelo(a) autor(a)

P674b Pisani, Paulo Henrique
 Biometrics in a data stream context / Paulo
 Henrique Pisani; orientador André Carlos Ponce de
 Leon Ferreira de Carvalho. -- São Carlos, 2017.
 206 p.

 Tese (Doutorado - Programa de Pós-Graduação em
 Ciências de Computação e Matemática Computacional) --
 Instituto de Ciências Matemáticas e de Computação,
 Universidade de São Paulo, 2017.

 1. Adaptive biometric systems. 2. Template
 update. 3. Data streams. 4. Keystroke dynamics. 5.
 Accelerometer biometrics. I. Carvalho, André Carlos
 Ponce de Leon Ferreira de, orient. II. Título.

Paulo Henrique Pisani

Biometria em um contexto de fluxo de dados

Tese apresentada ao Instituto de Ciências Matemáticas e de Computação - ICMC-USP, como parte dos requisitos para obtenção do título de Doutor em Ciências - Ciências de Computação e Matemática Computacional. *VERSÃO REVISADA*

Área de Concentração: Ciências de Computação e Matemática Computacional

Orientador: Prof. Dr. André Carlos Ponce de Leon Ferreira de Carvalho

USP – São Carlos
Abril de 2017

ACKNOWLEDGEMENTS

Firstly, I would like to thank everyone who contributed to this work. In particular, my parents, my sister, my family, professors, secretary staff, other employees from the Universidade de São Paulo and from the Universities I visited, friends and colleagues. I especially thank professor André Carlos Ponce de Leon Ferreira de Carvalho for the opportunity to work with him, for his contributions to this research and for his work as main advisor of this thesis. I also thank professors Ana Carolina Lorena, Romain Giot and Norman Poh for their contributions to the research carried out during this thesis and for the opportunity to work with them.

I would like to thank Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) and Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP - process 2012/25032-0) for the financial support. I also would like to thank LaBRI/Université de Bordeaux for the support to the Enhanced Template Update project.

ABSTRACT

The growing presence of the Internet in day-to-day tasks, along with the evolution of computational systems, contributed to increase data exposure. This scenario highlights the need for safer user authentication systems. An alternative to deal with this is by the use of biometric systems. However, biometric features may change over time, an issue that can affect the recognition performance due to an outdated biometric reference. This effect can be called as *template ageing* in the area of biometrics and as concept drift in machine learning. It raises the need to automatically adapt the biometric reference over time, a task performed by adaptive biometric systems. This thesis studied adaptive biometric systems considering biometrics in a data stream context. In this context, the test is performed on a biometric data stream, in which the query samples are presented one after another to the biometric system. An adaptive biometric system then has to classify each query and adapt the biometric reference. The decision to perform the adaptation is taken by the biometric system. Among the biometric modalities, this thesis focused on behavioural biometrics, particularly on keystroke dynamics and on accelerometer biometrics. Behavioural modalities tend to be subject to faster changes over time than physical modalities. Nevertheless, there were few studies dealing with adaptive biometric systems for behavioural modalities, highlighting a gap to be explored. Throughout the thesis, several aspects to enhance the design of adaptive biometric systems for behavioural modalities in a data stream context were discussed: proposal of adaptation strategies for the immune-based classification algorithm *Self-Detector*, combination of genuine and impostor models in the *Enhanced Template Update* framework and application of score normalization to adaptive biometric systems. Based on the investigation of these aspects, it was observed that the best choice for each studied aspect of the adaptive biometric systems can be different depending on the dataset and, furthermore, depending on the users in the dataset. The different user characteristics, including the way that the biometric features change over time, suggests that adaptation strategies should be chosen per user. This motivated the proposal of a modular adaptive biometric system, named *ModBioS*, which can choose each of these aspects per user. *ModBioS* is capable of generalizing several baselines and proposals into a single modular framework, along with the possibility of assigning different adaptation strategies per user. Experimental results showed that the modular adaptive biometric system can outperform several baseline systems, while opening a number of new opportunities for future work.

Keywords: Adaptive biometric systems, Template update, Data streams, Keystroke dynamics, Accelerometer biometrics.

RESUMO

A crescente presença da Internet nas tarefas do dia a dia, juntamente com a evolução dos sistemas computacionais, contribuiu para aumentar a exposição dos dados. Esse cenário evidencia a necessidade de sistemas de autenticação de usuários mais seguros. Uma alternativa para lidar com isso é pelo uso de sistemas biométricos. Contudo, características biométricas podem mudar com o tempo, o que pode afetar o desempenho de reconhecimento devido a uma referência biométrica desatualizada. Esse efeito pode ser chamado de *template ageing* na área de sistemas biométricos adaptativos ou de mudança de conceito em aprendizado de máquina. Isso levanta a necessidade de adaptar automaticamente a referência biométrica com o tempo, uma tarefa executada por sistemas biométricos adaptativos. Esta tese estudou sistemas biométricos adaptativos considerando biometria em um contexto de fluxo de dados. Neste contexto, o teste é executado em um fluxo de dados biométrico, em que as amostras de consulta são apresentadas uma após a outra para o sistema biométrico. Um sistema biométrico adaptativo deve então classificar cada consulta e adaptar a referência biométrica. A decisão de executar a adaptação é tomada pelo sistema biométrico. Dentre as modalidades biométricas, esta tese foca em biometria comportamental, em particular em dinâmica da digitação e em biometria por acelerômetro. Modalidades comportamentais tendem a ser sujeitas a mudanças mais rápidas do que modalidades físicas. Entretanto, havia poucos estudos lidando com sistemas biométricos adaptativos para modalidades comportamentais, destacando uma lacuna para ser explorada. Ao longo da tese, diversos aspectos para aprimorar o projeto de sistemas biométricos adaptativos para modalidades comportamentais em um contexto de fluxo de dados foram discutidos: proposta de estratégias de adaptação para o algoritmo de classificação imunológico *Self-Detector*, combinação de modelos genuíno e impostor no *framework* do *Enhanced Template Update* e aplicação de normalização de scores em sistemas biométricos adaptativos. Com base na investigação desses aspectos, foi observado que a melhor escolha para cada aspecto estudado dos sistemas biométricos adaptativos pode ser diferente dependendo do conjunto de dados e, além disso, dependendo dos usuários no conjunto de dados. As diferentes características dos usuários, incluindo a forma como as características biométricas mudam com o tempo, sugerem que as estratégias de adaptação deveriam ser escolhidas por usuário. Isso motivou a proposta de um sistema biométrico adaptativo modular, chamado *ModBioS*, que pode escolher cada um desses aspectos por usuário. O *ModBioS* é capaz de generalizar diversos sistemas *baseline* e propostas apresentadas nesta tese em um *framework* modular, juntamente com a possibilidade de atribuir estratégias de adaptação diferentes por usuário. Resultados experimentais mostraram que o sistema biométrico adaptativo modular pode superar diversos sistemas *baseline*, enquanto que abre um grande número de oportunidades para trabalhos futuros.

Palavras-chave: Sistemas biométricos adaptativos, Atualização de template, Fluxos de dados, Dinâmica da digitação, Biometria por acelerômetro.

LIST OF FIGURES

Figure 1 – Relationship among the chapters of the thesis.	30
Figure 2 – Biometric system.	34
Figure 3 – Adaptive biometric system.	35
Figure 4 – Adaptation using <i>Graph min-cut</i> : hypothetical example.	39
Figure 5 – Evaluation methodology in (ULUDAG; ROSS; JAIN, 2004).	44
Figure 6 – Evaluation methodology in (POH; KITTLER; RATTANI, 2014).	46
Figure 7 – Evaluation methodology in (RATTANI; MARCIALIS; ROLI, 2013a).	47
Figure 8 – Evaluation methodology in (PAGANO <i>et al.</i> , 2015).	48
Figure 9 – Evaluation methodology in (GIOT; ROSENBERGER; DORIZZI, 2012b).	49
Figure 10 – Overview of the user cross-validation for biometric data streams.	58
Figure 11 – Biometric data stream.	61
Figure 12 – Features extracted from the keystroke data.	65
Figure 13 – <i>Self-Detector</i> and its adaptive model.	76
Figure 14 – <i>Usage Control 2</i> - number of detectors over time (example).	83
Figure 15 – FNMR over time of <i>Usage Control</i> and baselines (keystroke dynamics).	88
Figure 16 – FNMR over time of <i>Usage Control</i> and baselines (accelerometer).	89
Figure 17 – FMR over time of <i>Usage Control</i> and baselines (keystroke dynamics).	90
Figure 18 – FMR over time of <i>Usage Control</i> and baselines (accelerometer).	91
Figure 19 – Correlation over time - keystroke dynamics (genuine samples only).	92
Figure 20 – Correlation over time - accelerometer biometrics (genuine samples only).	93
Figure 21 – <i>Enhanced Template Update</i> (ETU) Framework.	97
Figure 22 – ETU - <i>Positive Gallery Protection</i> (PGP).	98
Figure 23 – FNMR over time of ETU - Average from all users (keystroke dynamics).	106
Figure 24 – FNMR over time of ETU - Average from all users (accelerometer).	107
Figure 25 – FMR over time of ETU - Average from all users (keystroke dynamics).	109
Figure 26 – FMR over time of ETU - Average from all users (accelerometer).	110
Figure 27 – Sets to obtain the score normalization terms.	116
Figure 28 – Biometric systems using score normalization - relative performance.	119
Figure 29 – FMR and FNMR absolute performance difference per dataset.	121
Figure 30 – Relative Performance: score normalization vs adaptation.	123
Figure 31 – Score normalization performance per user - CMU.	125
Figure 32 – Score normalization performance per user - GREYC-Web (Logins).	126
Figure 33 – User biometric reference in <i>ModBioS</i>	132
Figure 34 – <i>Serial waterfall</i> data split in <i>ModBioS</i>	137

Figure 35 – Performance of combinations per user grouping/dataset.	140
Figure 36 – Performance of combinations per user - CMU.	142
Figure 37 – Performance of combinations per user - GREYC-Web (Logins).	143
Figure 38 – Correlations over time: comparing enrollment samples to all samples.	150
Figure 39 – FNMR over time of <i>ModBioS Fusion (Hybrid)</i> - Average from all users.	155
Figure 40 – FNMR over time of <i>Usage Control</i> and baselines - Average from all users.	155
Figure 41 – FMR over time of <i>ModBioS Fusion (Hybrid)</i> - Average from all users.	156
Figure 42 – FMR over time of <i>Usage Control</i> and baselines - Average from all users.	156
Figure 43 – Score normalization performance per user - GREYC.	190
Figure 44 – Score normalization performance per user - CMU.	191
Figure 45 – Score normalization performance per user - GREYC-Web (Logins).	192
Figure 46 – Score normalization performance per user - GREYC-Web (Passwords).	193
Figure 47 – Score normalization performance per user - McGill.	193
Figure 48 – Score normalization performance per user - WISDM 1.1.	194
Figure 49 – Score normalization performance per user - WISDM 2.0.	194
Figure 50 – Performance of combinations per user grouping - GREYC.	195
Figure 51 – Performance of combinations per user grouping - CMU.	196
Figure 52 – Performance of combinations per user grouping - GREYC-Web (Logins).	196
Figure 53 – Performance of combinations per user grouping - GREYC-Web (Passwords).	196
Figure 54 – Performance of combinations per user grouping - McGill.	197
Figure 55 – Performance of combinations per user grouping - WISDM 1.1.	197
Figure 56 – Performance of combinations per user grouping - WISDM 2.0.	197
Figure 57 – Performance of combinations per user - GREYC.	200
Figure 58 – Performance of combinations per user - CMU.	201
Figure 59 – Performance of combinations per user - GREYC-Web (Logins).	202
Figure 60 – Performance of combinations per user - GREYC-Web (Passwords).	203
Figure 61 – Performance of combinations per user - McGill.	204
Figure 62 – Performance of combinations per user - WISDM 1.1.	205
Figure 63 – Performance of combinations per user - WISDM 2.0.	206

LIST OF ALGORITHMS

Algorithm 1 – <i>Self-update</i>	37
Algorithm 2 – <i>Co-update</i>	38
Algorithm 3 – <i>Graph min-cut</i>	39
Algorithm 4 – <i>Growing window</i>	42
Algorithm 5 – <i>Sliding window</i>	43
Algorithm 6 – <i>Double Parallel (DB)</i>	43
Algorithm 7 – <i>Self-Detector</i>	70
Algorithm 8 – <i>M2005</i>	72
Algorithm 9 – <i>Self-Detector (Sliding and Growing)</i>	77
Algorithm 10 – <i>Usage Control and Usage Control R</i>	79
Algorithm 11 – <i>Usage Control S</i>	81
Algorithm 12 – <i>Usage Control 2</i>	82
Algorithm 13 – <i>Enhanced Template Update 0 (ETU 0)</i>	99
Algorithm 14 – <i>Enhanced Template Update 1 (ETU 1)</i>	100
Algorithm 15 – <i>Enhanced Template Update 2 (ETU 2)</i>	101
Algorithm 16 – <i>Enhanced Template Update 3 (ETU 3)</i>	101
Algorithm 17 – <i>ModBioS (single gallery)</i>	133
Algorithm 18 – <i>ModBioS (dual gallery)</i>	134
Algorithm 19 – <i>ModBioS Combiner (Grouped)</i>	136
Algorithm 20 – <i>ModBioS Combiner (User)</i>	136
Algorithm 21 – <i>ModBioS Combiner (Hybrid)</i>	136

LIST OF TABLES

Table 1	– Datasets that have been used to evaluate adaptive biometric systems.	52
Table 2	– Summary of keystroke dynamics datasets.	65
Table 3	– Summary of accelerometer-based gait biometrics datasets.	67
Table 4	– Summary of baseline biometric systems.	69
Table 5	– Results for <i>Usage Control</i> (all datasets).	84
Table 6	– Bayesian statistical test (balanced accuracy): <i>Usage Control</i>	86
Table 7	– <i>Enhanced Template Update</i> results - keystroke dynamics datasets.	103
Table 8	– <i>Enhanced Template Update</i> results - accelerometer datasets.	104
Table 9	– Bayesian statistical test (balanced accuracy): <i>ETU - Self-Detector</i>	105
Table 10	– Bayesian statistical test (balanced accuracy): <i>ETU - M2005</i>	105
Table 11	– Bayesian statistical test (FMR): <i>ETU - PGP</i>	111
Table 12	– Bayesian statistical test (balanced accuracy): score normalization.	122
Table 13	– Bayesian statistical test (balanced accuracy): score normalization vs adaptation.	124
Table 14	– Current module implementations.	137
Table 15	– Adaptive biometric systems generalized by <i>ModBioS</i>	139
Table 16	– Results for <i>ModBioS</i> using Grouped and User settings.	145
Table 17	– Results for <i>ModBioS</i> using <i>Hybrid</i> setting - keystroke dynamics datasets.	147
Table 18	– Results for <i>ModBioS</i> using <i>Hybrid</i> setting - accelerometer datasets.	148
Table 19	– Bayesian statistical test (balanced accuracy): <i>ModBioS - Fusion (Hybrid)</i>	149
Table 20	– Performance comparing <i>ModBioS</i> using <i>Hybrid</i> and <i>Random Combiner</i>	152
Table 21	– Bayesian statistical test (balanced accuracy): <i>Hybrid</i> Combiner vs random.	152
Table 22	– Balanced accuracy of baseline systems using the <i>ModBioS</i> tuning.	153
Table 23	– Bayesian statistical test (balanced accuracy): <i>Hybrid</i> vs tuned baselines.	154
Table 24	– Results - GREYC.	180
Table 25	– Results - CMU.	181
Table 26	– Results - GREYC-Web (Logins).	182
Table 27	– Results - GREYC-Web (Passwords).	183
Table 28	– Results - McGill.	183
Table 29	– Results - WISDM 1.1.	184
Table 30	– Results - WISDM 2.0.	184
Table 31	– Results using score normalization - GREYC.	185
Table 32	– Results using score normalization - CMU.	186
Table 33	– Results using score normalization - GREYC-Web (Logins).	186
Table 34	– Results using score normalization - GREYC-Web (Passwords).	187

Table 35 – Results using score normalization - McGill. 187

Table 36 – Results using score normalization - WISDM 1.1. 187

Table 37 – Results using score normalization - WISDM 2.0. 188

LIST OF ABBREVIATIONS AND ACRONYMS

BAcc	Balanced Accuracy
DB	Double Parallel
EER	Equal Error Rate
ETU	Enhanced Template Update
FAR	False Acceptance Rate
FIFO	First In First Out
FMR	False Match Rate
FNMR	False Non-Match Rate
FRR	False Rejection Rate
FTA	Failure to Acquire Rate
GUMR	Genuine Update Miss Rate
HMM	Hidden Markov Model
HTER	Half Total Error Rate
IDB	Improved Double Parallel
IUSR	Impostor Update Selection Rate
k-NN	k-Nearest Neighbour
LFU	Least Frequently Used
LRU	Least Recently Used
ModBioS	Modular Adaptive Biometric System
OCSVM	One-class Support Vector Machine
OMRk	Ordered Multiple Runs of k-Means
PGP	Positive Gallery Protection
SVM	Support Vector Machine

LIST OF SYMBOLS

$\mathcal{J}$	— Set of user indexes registered in a biometric system
j	— User index
$\mathcal{E}_j$	— Set of enrollment samples for the user j
ref_j	— Biometric reference of the user j
$\mathcal{R}$	— Set of biometric references from the registered users
$ref_j.T$	— User model in the biometric reference of the user j
$\mathbf{q}$	— Biometric query sample
θ_j^{verify}	— Parameters for the verification process for the user j
$label^p$	— Predicted label
$\mathcal{ID}$	— Set of recognized user indexes
$\theta^{identify}$	— Parameters for the identification process
$ref_{j(t)}$	— Biometric reference of the user j at instant t (usually before adaptation)
$\mathcal{A}$	— Set of biometric samples for adaptation
$ref_{j(t+1)}$	— Biometric reference of the user j at instant $t + 1$ (usually after adaptation)
θ_j^{adapt}	— Parameters for the adaptation process for the user j
$ref_j.GL$	— Gallery in the biometric reference of the user j
ds	— Biometric data stream
$\mathcal{D}$	— All users in a given dataset
$ref_j.GL^G$	— Genuine gallery in the biometric reference of the user j
$ref_j.T^G$	— Genuine user model in the biometric reference of the user j
$ref_j.GL^I$	— Impostor gallery in the biometric reference of the user j
$ref_j.T^I$	— Impostor user model in the biometric reference of the user j
μ_j^c	— Average of the cohort scores for the user j
σ_j^c	— Standard deviation of the cohort scores for the user j
$\mu_{I,j}^d$	— Average of the impostor scores for the user j
$\sigma_{I,j}^d$	— Standard deviation of the impostor scores for the user j
$\mu_{G,j}^d$	— Average of the genuine scores for the user j
μ_G^d	— Average of the genuine scores among all registered users
$S_{j,query}^c$	— Cohort scores for the user j considering a given query
S_j^G	— Genuine scores for the user j
S_j^I	— Impostor scores for the user j
$\mathcal{H}_j^I$	— All enrollment samples from the registered users i , where $i \neq j$

$ref_j.M$	— Modules in the biometric reference of the user j
$ref_j.M.U$	— Global modules in the biometric reference of the user j
$ref_j.M.G^1$	— Modules specific to model/gallery 1 in the biometric reference of the user j
$ref_j.M.G^2$	— Modules specific to model/gallery 2 in the biometric reference of the user j
$ref_j.T^1$	— User model/gallery 1 (<i>ModBioS</i>) in the biometric reference of the user j
$ref_j.T^2$	— User model/gallery 2 (<i>ModBioS</i>) in the biometric reference of the user j
$\mathcal{M}_{all}$	— List of all possible combinations of module implementations

CONTENTS

1	INTRODUCTION	25
1.1	Problem and hypotheses	26
1.2	Objective	29
1.3	Main contributions	29
1.4	Thesis chapters	29
2	ADAPTIVE BIOMETRIC SYSTEMS	31
2.1	Biometric systems and adaptation	32
2.2	Current adaptation strategies	36
2.2.1	<i>Offline adaptation</i>	36
2.2.2	<i>Online adaptation</i>	40
2.3	Evaluation	43
2.3.1	<i>Methodologies</i>	44
2.3.1.1	<i>Separate sets for adaptation and test</i>	44
2.3.1.2	<i>Joint set for adaptation and test</i>	46
2.3.2	<i>Metrics</i>	50
2.3.3	<i>Modalities and datasets</i>	52
2.4	Chapter remarks	53
3	EVALUATION METHODOLOGY	55
3.1	Biometrics in a data stream context	56
3.2	User cross-validation for biometric data streams	57
3.2.1	<i>Biometric data stream generation</i>	59
3.2.2	<i>Database-aware biometric systems and parameter setting</i>	61
3.3	Metrics	62
3.4	Biometric modalities and datasets	63
3.4.1	<i>Keystroke dynamics</i>	64
3.4.2	<i>Accelerometer-based gait biometrics</i>	66
3.5	Baseline biometric systems	68
3.5.1	<i>Self-Detector</i>	70
3.5.2	<i>M2005</i>	70
3.5.3	<i>OCSVM</i>	71
3.6	Statistical test	72
3.7	Chapter remarks	73

4	USAGE CONTROL FOR SELF-DETECTOR	75
4.1	Positive selection and adaptation strategies	76
4.1.1	<i>Usage Control: first version</i>	77
4.1.2	<i>Usage Control R</i>	80
4.1.3	<i>Usage Control S</i>	80
4.1.4	<i>Usage Control 2</i>	82
4.2	Experimental results	83
4.2.1	<i>Global results</i>	83
4.2.2	<i>Performance over time</i>	87
4.3	Chapter remarks	93
5	COMBINING GENUINE AND IMPOSTOR MODELS	95
5.1	Enhanced Template Update	96
5.1.1	<i>Positive Gallery Protection Check</i>	96
5.1.2	<i>ETU 0: Simple comparison of scores</i>	98
5.1.3	<i>ETU 1: Comparison of scores</i>	99
5.1.4	<i>ETU 2: k-NN like</i>	100
5.1.5	<i>ETU 3: Clustering impostor samples</i>	100
5.2	Experimental results	102
5.2.1	<i>Global results</i>	102
5.2.2	<i>Performance over time</i>	105
5.3	Chapter remarks	111
6	ADAPTATION USING SCORE NORMALIZATION	113
6.1	Score normalization	114
6.2	Computation of score normalization in the user cross-validation . .	115
6.3	Experimental results	118
6.3.1	<i>Adaptive biometric systems improved by score normalization</i> . . .	118
6.3.2	<i>Impact of score normalization versus adaptation</i>	122
6.3.3	<i>The best score normalization can be different among users</i>	125
6.4	Chapter remarks	126
7	MODULAR ADAPTIVE BIOMETRIC SYSTEM	129
7.1	ModBioS: modular adaptive biometric system	130
7.1.1	<i>Modules</i>	131
7.1.2	<i>Single gallery mode</i>	132
7.1.3	<i>Dual gallery mode (biometric fusion)</i>	133
7.2	Combiner	135
7.2.1	<i>Current module implementations</i>	137
7.2.2	<i>Generalized adaptive biometric systems</i>	138
7.3	Experimental results	139

7.3.1	<i>Is the best adaptation strategy different for distinct datasets?</i>	139
7.3.2	<i>Is the best adaptation strategy different among users (same dataset)?</i>	141
7.3.3	<i>How to choose the combination of modules?</i>	144
7.3.3.1	<i>Can both Grouped and User settings be combined?</i>	146
7.3.3.2	<i>Why the estimation on the enrollment data is not optimal?</i>	149
7.3.3.3	<i>Is the current choice of combinations any better than random choice?</i>	151
7.3.3.4	<i>Could the baselines benefit from the ModBioS parameter tuning?</i>	152
7.3.4	<i>Adaptive modular biometric system performance over time</i>	154
7.4	Extensions for the modular system	157
7.5	Chapter remarks	157
8	CONCLUSION	159
8.1	Contributions and results	160
8.2	Publications	164
8.3	Limitations	165
8.4	Future work	166
REFERENCES		169
APPENDIX A	EXPERIMENTAL RESULTS	179
APPENDIX B	EXPERIMENTAL RESULTS - SCORE NORMALIZATION	185
APPENDIX C	SCORE NORMALIZATION - PERFORMANCE PER USER	189
APPENDIX D	COMBINATIONS PERFORMANCE PER USER GROUPING	195
APPENDIX E	COMBINATIONS PERFORMANCE PER USER	199

INTRODUCTION

Nowadays, in light of the increasing use of digital identities, concerns regarding data exposure due to identity theft have increased. Identity theft occurs when someone acts as being someone else illegally (DUSERICK, 2004). These concerns emphasize the need of enhanced authentication mechanisms. Authentication is the process to verify the identity of a user through credentials (WINDLEY, 2005):

- what the user knows (e.g. password);
- what the user has (e.g. card, *token*);
- what the user is/does (e.g. fingerprint, iris, keystroke dynamics);
- a combination of the previous credentials.

The commonly used login and password combination may not provide enough security for some applications, since a password can be easily copied and even guessed when weak passwords are used (e.g. “123456”). Using a card or token can improve the security of a system, although the possession of the device may be compromised.

An alternative to using passwords is the use of biometrics to perform authentication. *Biometrics* is defined as the automatic recognition of individuals through their biological or behavioural traits (JAIN; NANDAKUMAR; ROSS, 2016). Several attributes have been used in the literature for this purpose, such as fingerprint, iris, keystroke dynamics, gait, etc. A *biometric system* is a pattern recognition system that acquires biometric data, extracts a feature set from these data, which is compared to the *biometric reference* stored in the biometric database (JAIN; ROSS; PRABHAKAR, 2004). The biometric reference is also referred to as template in the literature. There are two main processes in a biometric system: *enrollment*, when the biometric reference is obtained, and *test/recognition*, when the system is queried.

Biometric features should meet a number of requirements to be used in a biometric system. These requirements include universality, distinctiveness, collectability and permanence (JAIN; ROSS; PRABHAKAR, 2004). Permanence refers to the fact that the biometric features

will remain stable (not change) over time. However, recently, studies have shown that biometric features extracted from several traits may change over time (ROLI; DIDACI; MARCIALIS, 2008). Consequently, the biometric reference may not reflect the current user data anymore, an issue known as *template ageing* (JAIN; NANDAKUMAR; ROSS, 2016). As a result, the recognition performance of biometric systems may be affected. Throughout this thesis, the term performance refers to the recognition performance of a biometric system, also named as predictive performance in Machine Learning.

This thesis deals with *biometrics in a data stream context*. In Data Stream Mining, a data stream is defined as a potentially unlimited sequence of examples that are accessed sequentially, subject to concept drift (AGGARWAL *et al.*, 2003; BIFET *et al.*, 2009). The data stream here is the sequence of biometric samples¹ received by the biometric system. Data distribution from the genuine user is subject to changes over time, also referred to as *concept drift* (GAMA *et al.*, 2014; ŽLIOBAITĖ, 2010). In this context, the goal is to keep the biometric reference updated to changes observed in the data stream.

A straightforward way to update the biometric reference is to perform periodical enrollment sessions for all users in the system. However, this can imply in high operational costs and user annoyance. Another possibility is the use of an adaptive biometric system, which is able to automatically adapt the biometric reference over time (ROLI; DIDACI; MARCIALIS, 2008; POH; RATTANI; ROLI, 2012; RATTANI, 2015). In addition to the *enrollment* and *test/recognition* processes, these systems contain an additional process called *adaptation*. Adaptive biometric systems are the focus of this thesis.

1.1 Problem and hypotheses

The adaptive biometric systems area is relatively new and there are many questions still to be answered. This fact, combined to recent work reporting improvements of adaptive biometrics over non-adaptive approaches (GIOT; ROSENBERGER; DORIZZI, 2012b; RATTANI; MARCIALIS; ROLI, 2013a), highlights the importance of new studies in the area. It is also known that behavioural modalities are subject to a larger number of changes in a short term than physical ones (GIOT; ROSENBERGER; DORIZZI, 2012c). Nonetheless, most work done in the field has been concerned with physical modalities, like fingerprint and face. This work aims to reduce this gap by focusing on *behavioural modalities*.

All in all, the problem investigated in this thesis is the *design of adaptive biometric systems for behavioural modalities in a data stream context*. A biometric system in such a context needs to consider that the biometric system will receive biometric samples sequentially

¹ In Biometrics, the term *biometric sample* is usually adopted instead of the term *example*, commonly employed in the Machine Learning literature. This can be observed in several publications in the field of Biometrics, including standards, such as ISO/IEC 19784-1:2006 (ISO/IEC, 2006), and several papers (JAIN; NANDAKUMAR; ROSS, 2016; GIOT; EL-ABED; ROSENBERGER, 2012; RATTANI; MARCIALIS; ROLI, 2011a; POH; RATTANI; ROLI, 2012). This thesis adopts the same standard regarding the term *biometric sample*. Hence, a *biometric sample*, or simply a *sample*, is defined as a synonym of an *example* throughout the text.

and that the biometric features may change over time. Furthermore, these changes can occur in a short time span for the behavioural modalities.

This work focuses on keystroke dynamics (PISANI; LORENA, 2013) and accelerometer-based gait biometrics / accelerometer biometrics (SPRAGER; ZAZULA, 2009; KAGGLE, 2013). They have been chosen due to the availability of suitable datasets. A dataset needs to contain several samples for each user and ideally be acquired at different sessions, which greatly reduces the amount of datasets that are appropriate for studying adaptation. Moreover, most of the work in the area of adaptive biometric systems have dealt with physical modalities, indicating a gap for new research on behavioural modalities. These modalities usually imply in lower discriminative power than physical modalities (e.g. face, fingerprint, iris), introducing additional challenges to the biometric system. In addition, they tend to be subject to faster changes of the biometric features over time than physical ones, like face, for example (GIOT; ROSENBERGER; DORIZZI, 2012c), emphasizing the importance of further studies for behavioural biometric modalities in a data stream context.

Several aspects concerning the design of adaptive biometric are investigated in this work: adaptation strategy, combination of genuine and impostor models, score normalization and choice of adaptation rules per user via a modular framework. They are briefly described in this section, along with the hypotheses investigated in this thesis.

The immune-based *Self-Detector* algorithm has been successfully applied to keystroke dynamics, but in a scenario that did not consider the adaptation of the biometric reference (PISANI, 2012; PISANI; LORENA, 2015). Even though, that work suggested that the typing rhythm changes over time. However, is it possible to turn *Self-Detector* into an adaptive algorithm? *Self-Detector* is part of the positive selection of immune algorithms (STIBOR; TIMMIS, 2005). During the enrollment, this classification algorithm stores all samples as detectors. Afterwards, in the test phase, each query is compared to the detectors. If at least one detector matches the query, it is classified as genuine, otherwise, the query is classified as impostor. This immune algorithm can be considered to belong to the instance-based algorithms, which makes it easier to adapt its model over time (MCEWAN; HART, 2009; MENA-TORRES; AGUILAR-RUIZ, 2014). Instance-based algorithms store training examples in memory and perform classification of new examples using nearby examples in the memory (MITCHELL, 1997). The detector set can be understood as this memory for *Self-Detector*. One possibility to adapt the detector set would be to control its usage for matching, in order to keep only the most used detectors, leading to the first hypothesis. The research presented in this thesis proposed some alternatives in this line, as will be seen in Chapter 4.

- **H1:** Positive selection can become an adaptive class of immune algorithms by controlling the usage of the detectors, enabling its use for biometrics in a data stream context.

Furthermore, to the best of the author's knowledge, current adaptive biometric systems only adapt a model for the genuine user in the biometric reference. This also means that samples

classified as impostor are not considered, meaning that important information may be discarded. This brings the following question: would the use of an additional impostor model generated from samples classified as impostor be able to enhance the recognition performance? The second hypothesis comes from this reasoning.

- **H2:** The use of a negative/impostor model, in addition to the positive/genuine model, can improve the recognition performance of adaptive biometric systems.

Moreover, score normalization (POH; MERATI; KITTLER, 2009) has been used in biometrics to refine the final classification decision. This procedure often increases the class separability, which consequently improves the performance of biometric systems. Nevertheless, to the best of the author's knowledge, score normalization has not been used for biometrics in a data stream context. Hence, could score normalization be used to increase adaptive biometric systems performance by a better refinement of the final decision? It means that by using score normalization, a more effective threshold could also be chosen, which can in turn increase the amount of correctly accepted genuine samples for adaptation, while reducing the amount of impostor samples wrongly accepted for adaptation. It results in the third hypothesis.

- **H3:** Score normalization can improve the recognition performance of adaptive biometric systems.

Finally, as observed in previous studies (PISANI; LORENA; CARVALHO, 2015b), the best performing adaptation strategy for a dataset is not always the best strategy for another dataset. This may be caused by different *change patterns* among the datasets. A *change pattern* is defined here as the way that the biometric features change over time. In line with this reasoning, it has also been reported that characteristics, including the *change pattern*, of a user can be different from another user in the same dataset. The work from (PISANI; LORENA; CARVALHO, 2015b) has reported that the typing rhythm change can differ depending on the user. In (POH *et al.*, 2015), it was shown that the recognition performance changes in different fashions depending on the user (e.g. the performance decreases for some users over time, while it can also increase for others). Another study has also suggested that different groups of users may need modifications in the adaptation strategy and parameters (RATTANI; MARCIALIS; ROLI, 2009).

All these findings raise the following question: should a different adaptation strategy be chosen for each user in the same dataset to improve the recognition performance? A system able to select the most suitable adaptation strategy for each user would bring adaptive biometric systems to another level: the choice of adaptation rules per user. To the best of the author's knowledge, an adaptive biometric system with these capabilities has never been proposed. It motivates the next hypothesis.

- **H4:** An adaptive biometric system can be divided into modules, such that it can choose different module implementations for each user. Therefore, distinct adaptation rules can be assigned to each user.

This thesis hypothesizes that this new adaptive biometric system can be designed following a modular framework, in which each aspect of the system is divided into modules. By choosing the implementations of the modules, the adaptive behaviour of several current adaptive biometric systems could be reproduced. Within this framework, the choice of adaptation rules per user is equivalent to choosing the implementations of each module per user.

Assuming this modular system is able to generalize all baseline biometric systems, it would have the potential to obtain higher (or at least the same) recognition performance than that of all baselines can reach. Since the best adaptation strategy may differ from a user to another, this modular system would also potentially be able to obtain the best recognition performance on all datasets.

1.2 Objective

The main objective of this thesis is to *enhance the design of adaptive biometric systems for behavioural modalities in a data stream context to improve the recognition performance of these systems*. This is accomplished by proposing new adaptation strategies, use of an additional impostor model, application of score normalization and choice of adaptation rules per user.

In the end, a framework for a *modular adaptive biometric system* is proposed, which can generalize several baseline systems and, in the future, include all aspects of an adaptive biometric system investigated in this thesis. This modular system is able to choose the adaptation rules per user, bringing adaptation to another level: choice of adaptation rules per user.

1.3 Main contributions

The thesis deals with several aspects of an adaptive biometric system. The main contributions are presented next:

- Proposal of an evaluation methodology for *biometrics in a data stream context*;
- Improvement of immune-based *Self-Detector* to support adaptation of the detector set by controlling the usage of the detectors for matching;
- Combination of an impostor model with the genuine model in the biometric reference of an adaptive biometric system;
- Approach to apply score normalization for biometrics in a data stream context;
- Introduction of a *modular adaptive biometric system* which is able to generalize current adaptation strategies and choose distinct adaptation rules per user.

1.4 Thesis chapters

Throughout the thesis, several aspects of biometrics in a data stream context are discussed. An overview of the covered topics and their relationship is shown in Figure 1. First of

all, there are two base parts of the thesis: the review of the area of adaptive biometric systems (Chapter 2) and the proposed evaluation methodology, including a discussion on several questions regarding the design of experiments for biometrics in a data stream context (Chapter 3). Above these two base parts, three main aspects of the design of adaptive biometric systems are studied: the extension of the immune-based *Self-Detector* to support adaptation by controlling the usage of the detectors for matching (Chapter 4), the combination of genuine and impostor models in a new adaptation framework (Chapter 5) and the application of score normalization to biometrics in a data stream context (Chapter 6). Each of these chapters deal with a diverse aspect of adaptive biometric systems. Next, a *modular adaptive biometric system* is proposed, which is capable of generalizing several baselines and proposals into a single modular framework, along with the possibility of assigning different adaptation strategies per user (Chapter 7). Among the proposals in Chapters 4 to 6, only the *Usage Control* versions are generalized by the modular system (non-dashed arrow). The proposals in Chapters 5 and 6 can be part of the modular system in future implementations (dashed arrow). Finally, Chapter 8 presents the final conclusions, limitations and research directions for future work.

Figure 1 – Relationship among the chapters of the thesis.

ADAPTIVE BIOMETRIC SYSTEMS

Biometrics can be defined as the automatic recognition of individuals through their biological or behavioural traits (JAIN; NANDAKUMAR; ROSS, 2016). A number of traits have been used for this purpose, such as fingerprint, face, iris, keystroke dynamics, gait, etc. This thesis deals with biometric systems, which are pattern recognition systems that acquire biometric data, extract a feature set from these data, which is then compared to the *biometric reference* stored in the biometric database (JAIN; ROSS; PRABHAKAR, 2004). In the literature of the area, the biometric reference can also be referred to as template.

Biometric features should meet a number of requirements to be used in a biometric system. These requirements include universality, distinctiveness, collectability and permanence (JAIN; ROSS; PRABHAKAR, 2004):

- Universality: the biometric trait should be present in everyone;
- Distinctiveness: it should be possible to differ one person from another by the biometric feature;
- Collectability: it should be possible to measure the biometric trait quantitatively, to allow subsequent processing;
- Permanence: the biometric feature should remain stable (invariant) over time.

However, studies have shown recently that biometric features may change over time (ROLI; DIDACI; MARCIALIS, 2008). Consequently, the biometric reference may not reflect the current user data anymore, an issue known as *template ageing* (JAIN; NANDAKUMAR; ROSS, 2016). As a result, the recognition performance of biometric systems may be affected. In this thesis, the term performance refers to the recognition performance of a biometric system, also named as predictive performance in Machine Learning.

A straightforward way to update the biometric reference is to perform periodical enrollment sessions for all users in the system. However, this can imply in high operational costs and user annoyance. Another possibility is the use of an adaptive biometric system, which is

able to automatically adapt the biometric reference over time (ROLI; DIDACI; MARCIALIS, 2008; POH; RATTANI; ROLI, 2012; RATTANI, 2015). Adaptive biometric systems can adapt the biometric reference to either changes due to time (e.g. trait ageing) or due to changing conditions (e.g. different pose in a face recognition system) (POH; RATTANI; ROLI, 2012). Adaptation can also be employed to improve the biometric reference when few samples are used for enrollment (YU *et al.*, 2012).

This chapter provides an overview of the area of adaptive biometric systems, with special emphasis on adaptation strategies and evaluation methodology. The next sections are organized as follows: Section 2.1 introduces definitions and formalization adopted in the thesis; Section 2.2 describes several adaptation strategies proposed in the literature; Section 2.3 discusses aspects of the evaluation of adaptive biometric systems: methodologies, metrics and datasets; and, Section 2.4 presents the final remarks of the chapter.

2.1 Biometric systems and adaptation

A *biometric system* is a pattern recognition system that acquires biometric data, extracts a feature set from these data, which is then compared to the *biometric reference* stored in a biometric database (JAIN; ROSS; PRABHAKAR, 2004). There are two main processes in a biometric system: *enrollment* and *test/recognition*.

Given $\mathcal{J}$, a set of user indexes registered in a biometric system, the enrollment process for a user index $j \in \mathcal{J}$ receives a set of enrollment samples $\mathcal{E}_j$ and outputs the biometric reference ref_j . This is performed by the *enroll* function:

$$ref_j \leftarrow enroll(\mathcal{E}_j)$$

This biometric reference ref_j is then stored in a biometric database, which contains the set of biometric references $\mathcal{R} = \{ref_j \mid j \in \mathcal{J}\}$ from the registered users. The enrollment process is performed for all registered users (all users in the set $\mathcal{J}$). For example, these users could be the employees in a company or the students in a University. They can be considered as the genuine users of the biometric system.

In the context of this thesis, the enrollment is implemented by a *classification algorithm* through the function *train*, where $ref_j.T$ is the field in the biometric reference which stores the user model induced by the classification algorithm:

$$ref_j.T \leftarrow classificationAlgorithm.train(\mathcal{E}_j)$$

This thesis adopts a notation similar to object-oriented programming using the *dot* symbol (DEITEL; DEITEL, 2001). In this thesis, the *dot* is used to highlight that the field T belongs to ref_j in this case. In object-oriented programming, it could be understood as an attribute of

an object. Adaptive biometric systems need additional fields in the biometric reference, which are introduced throughout the thesis as the adaptation strategies are presented. The additional fields are also represented using the *dot* to emphasize they belong to the biometric reference. This *dot* notation is also used for the classification algorithm to make it clear that the function *train* is implemented by it. In object-oriented programming, this would refer to the methods of the objects.

The test/recognition process can operate in two different modes (JAIN; NANDAKUMAR; ROSS, 2016): *verification* and *identification*. In the *verification* mode, a biometric query sample $\mathbf{q}$ is compared to the biometric reference of a claimed user index j using parameters θ_j^{verify} . A query is a biometric sample acquired to perform recognition. This task is implemented by a classification algorithm, which outputs the *predicted* label $label^p$ for the biometric query: *genuine* or *impostor*. If the query is predicted as belonging to the claimed user, the label *genuine* is returned and, otherwise, the label *impostor* is returned. This is performed by the *test.verify* function:

$$label^p \leftarrow test.verify(ref_j, \mathbf{q}, \theta_j^{verify})$$

Several biometric systems output a *score* from the comparison of a sample query $\mathbf{q}$ to the biometric reference ref_j . From this *score*, the label $label^p$ is obtained by comparing it to a *decision threshold*. In this implementation, the *decision threshold* would be an element in the set of parameters θ_j^{verify} for verification. Other classification algorithms may require different parameters. Note that, in this thesis, the term *parameters* is used to refer to the *hyper-parameters* (PEDREGOSA, 2016) of the algorithms.

In the second mode, *identification*, a biometric query $\mathbf{q}$ is presented to the biometric system and it will output a set of recognized user indexes $\mathcal{ID}$ using the set of parameters $\theta^{identify}$, such that $\mathcal{ID} \subseteq \mathcal{J}$. Note that $\mathcal{ID}$ can be a null set $\{\}$, meaning that the query is classified as an *impostor*. The set of parameters $\theta^{identify}$ refers to the classification algorithm, similarly to the case of the verification mode. This is performed by the *test.identify* function:

$$\mathcal{ID} \leftarrow test.identify(\mathcal{R}, \mathbf{q}, \theta^{identify})$$

This thesis deals with verification mode as shown in Figure 2. In short, the test/recognition can be understood as a two-class classification task. More specifically, it is frequently implemented as a one-class classification task, since the enrollment samples belong only to the genuine user. A one-class classification task is a particular case of a two-class classification task in which all the training examples belong to the same class.

Figure 2 – Biometric system.

Similarly to the case of the *enrollment*, the verification is performed by a *classification algorithm* in this thesis. The output of the verification is obtained by the function *test* of the classification algorithm:

$$label^p \leftarrow classificationAlgorithm.test(ref_j.T, \mathbf{q}, \theta_j^{verify})$$

As previously discussed, several biometric systems also implement the verification by computing a similarity score between the query and the user model. In the context of this thesis, this is done by the function *getSimilarityScore*:

$$score \leftarrow classificationAlgorithm.getSimilarityScore(ref_j.T, \mathbf{q})$$

This function returns the similarity score between the user model $ref_j.T$ and the query $\mathbf{q}$. If the returned *score* is above a *decision threshold*, the predicted label is genuine and, otherwise, the label is impostor.

The features used for the biometric samples should meet a number of requirements to be used in a biometric system. These requirements include universality, distinctiveness, collectability and permanence (JAIN; ROSS; PRABHAKAR, 2004). Permanence refers to the fact that the biometric feature should remain stable (not change) over time. However, recent studies have shown that biometric features may change over time (ROLI; DIDACI; MARCIALIS, 2008). Consequently, the biometric reference may not reflect the current user profile, an issue known as *template ageing* (JAIN; NANDAKUMAR; ROSS, 2016).

As a result, the recognition performance of the biometric system may be affected. In light of this fact, adaptive biometric systems have been proposed. These systems are able to automatically adapt the biometric reference over time (ROLI; DIDACI; MARCIALIS, 2008; POH; RATTANI; ROLI, 2012) using an adaptation strategy as illustrated in Figure 3.

Figure 3 – Adaptive biometric system.

Apart from the *enrollment* and *test/recognition* processes, an adaptive biometric system also has an *adaptation* process. This process receives the current biometric reference $ref_{j(t)}$ at instant t and a set of biometric samples for adaptation $\mathcal{A}$ as input. The adaptation process then outputs the adapted biometric reference $ref_{j(t+1)}$. This is performed by the *adapt* function:

$$ref_{j(t+1)} \leftarrow adapt(ref_{j(t)}, \mathcal{A}, \theta_j^{adapt})$$

The set of samples used for adaptation is obtained during the system operation and contains samples that will be used for adapting the biometric reference. Usually, this set is composed by samples classified as genuine by the test/recognition process of the biometric system. Some studies only include samples classified as genuine with high confidence in this set (ROLI; MARCIALIS, 2006; GIOT; DORIZZI; ROSENBERGER, 2011). This can be implemented by the use of an additional *adaptation threshold*, which is more stringent than the *decision threshold*. This additional *adaptation threshold*, sometimes named as *update threshold*, would be an element in the set of parameters for adaptation θ_j^{adapt} . Depending on the adaptation strategy, different parameters may be required, as discussed in the next sections of this thesis.

The adaptation process is implemented by the *adaptation strategy*, which can be performed online or offline (POH; RATTANI; ROLI, 2012). Online adaptation occurs when adaptation is performed after each query sample is classified. Basically, the adaptation process is triggered every time the test/recognition is executed. Conversely, offline adaptation occurs when the biometric system does not perform adaptation after each classified query. In this case, the system waits to store a batch of samples in the set $\mathcal{A}$ and then starts the adaptation.

In summary, an adaptive biometric system is composed by a *classification algorithm* (test/recognition process) and an *adaptation strategy* (adaptation process), as shown in Figure 3. This thesis focuses on adaptation strategies for biometric systems. A key aspect of an adaptive biometric system is how the adaptation process is implemented. The next section presents several *adaptation strategies* that have been proposed in the literature.

2.2 Current adaptation strategies

This section presents several adaptation strategies that have been used to adapt the biometric reference. Most of them consider that the biometric reference is composed by several biometric samples/templates. This set of samples/templates is sometimes named as a gallery (GIOT; ROSENBERGER; DORIZZI, 2012b; BIGGIO *et al.*, 2012; RATTANI; MARCIALIS; ROLI, 2008; MHENNI *et al.*, 2016) and it is considered as a field of the biometric reference here $ref_j.GL$ ($gl_i \in ref_j.GL$). In line with this concept, adaptation can be briefly defined as the addition and removal of samples/templates from the gallery. Afterwards, the classification algorithm is re-trained on the adapted $ref_j.GL$ to obtain the model $ref_j.T$ for the test/recognition process. When an instance-based algorithm is used, $ref_j.T = ref_j.GL$.

A key assumption in current adaptation strategies is that all samples in the gallery are from the genuine user. This characterizes a *one-class* classification scenario in Machine Learning, where the user model (biometric reference) is trained only with genuine examples and, afterwards, has to classify whether the test examples (queries) belong or not to the genuine class.

In order to present the adaptation strategies, this section divides them into two main groups: *online* and *offline* adaptation. As discussed in the last section, online adaptation means that the biometric reference is updated as soon as each query is processed, while the offline adaptation waits to store a batch of samples before adapting the biometric reference.

2.2.1 Offline adaptation

In the offline mode, a set of biometric samples $\mathcal{A}$ is used as source for the adaptation process. During the evaluation of the biometric system, this set can be formed in different ways and, when presented to the adaptive biometric system, it processes the whole set and adapts the biometric reference. This set can contain just true genuine samples or both genuine and impostor samples, depending on the study. This is discussed in Section 2.3. Clearly, the adaptation process is easier when there are just true genuine samples in $\mathcal{A}$. If there is the possibility of the presence of impostor samples in $\mathcal{A}$, the adaptation strategy should avoid these samples in order to not include impostor patterns in the genuine biometric reference.

One of the first studies using adaptation in biometric systems was (ULUDAG; ROSS; JAIN, 2004), which deals with offline adaptation. The experiments were carried out on a fingerprint dataset acquired in two sessions, each containing 100 samples per user. First, using only data from the first session, 25 samples were used for enrollment (initial biometric reference), and the next 75 samples for test. Then, using the data from the second session, adaptation took place.

Two simple adaptation strategies were evaluated: BATCH-UPDATE and AUGMENT-UPDATE. In the former, all 25 samples of the initial biometric reference are replaced by the first 25 samples of the second session. Note that this study only uses true genuine samples for the

adaptation set $\mathcal{A}$. In the latter strategy, AUGMENT-UPDATE, the 25 samples from the second session are added to the initial set containing the 25 samples from the first session. Two template selection methods are applied over this adapted set: DEND and MDIST. DEND applies an agglomerative complete link clustering algorithm over the 50 samples to obtain a dendrogram. Then, K clusters are identified. For each of the K clusters, the sample with the lowest average distance to the other samples in the cluster is selected (the medoid). This results in a set with K samples. In MDIST, the output is also a set with K samples. However, MDIST adopts a simpler method. First, for each sample, the average distance to all other samples is computed. The K samples with lowest average distance are selected. One key benefit of using template selection is to avoid an endless increase of the set of samples used for the biometric reference, particularly for AUGMENT-UPDATE. According to the reported results, MDIST template selection yields higher recognition accuracy than DEND. Regarding the adaptation strategies, AUGMENT-UPDATE showed improved performance over BATCH-UPDATE, although both performed better than no adaptation.

In (ROLI; MARCIALIS, 2006), the authors proposed to use Self-training for adaptation, which is a semi-supervised learning method (ZHU, 2006). In this method, adaptation occurs by adding the queries classified as genuine to the biometric reference, as shown in the Algorithm 1. Usually, only those queries classified with high confidence are used for adaptation. This can be accomplished by adopting an additional *adaptationThreshold* (RATTANI *et al.*, 2009; RATTANI; MARCIALIS; ROLI, 2011b). When Self-training is applied to biometric systems, it is called *Self-update* in several papers (RATTANI; MARCIALIS; ROLI, 2013a; DIDACI; MARCIALIS; ROLI, 2014).

Self-update is a simple adaptation strategy that was initially evaluated under an offline adaptation scenario (ROLI; MARCIALIS, 2006). However, this strategy can also be applied in an online mode. In this case, the set of samples $\mathcal{A}$ is composed by the current query $\mathbf{q}$ only. If the similarity score between the query $\mathbf{q}$ and the gallery is above the *adaptationThreshold*, it is added to the gallery.

Algorithm 1: Self-update.

Input: $ref_{j(t)}$, $\mathcal{A}$, $\theta_j^{adapt} = \{adaptationThreshold\}$

Output: $ref_{j(t+1)}$

- 1 $\mathcal{A}' = \{a_i \in \mathcal{A} \mid classificationAlgorithm.getSimilarityScore(ref_{j(t)}.T, a_i) > adaptationThreshold\}$
 - 2 $ref_{j(t+1)} = ref_{j(t)}$
 - 3 $ref_{j(t+1)}.GL = ref_{j(t+1)}.GL \cup \mathcal{A}'$
 - 4 $ref_{j(t+1)}.T \leftarrow classificationAlgorithm.train(ref_{j(t+1)}.GL)$
 - 5 **return** $ref_{j(t+1)}$
-

An adaptation strategy related to *Self-update* is *Co-update* (ROLI; DIDACI; MARCIALIS, 2007; RATTANI; MARCIALIS; ROLI, 2013b), which is based on *Co-training* from semi-supervised learning (BLUM; MITCHELL, 1998; ZHU, 2006). *Co-update* considers biometric systems that deals with multi-biometrics, i.e., systems that handle more than one biometric modality at the same time (e.g. face and fingerprint). These systems can combine the results from both modalities to output the classification decision. For the adaptation process, there are two sets of samples, one for each modality: $\mathcal{A}^1$ and $\mathcal{A}^2$. If a sample a_i^1 from one modality is above its $adaptationThreshold^1$, the sample a_i^2 from the other modality is added to the adaptation set $\mathcal{A}'^2$ of that modality. This process is repeated for both modalities, so, if a sample a_i^2 from the modality 2 is above its $adaptationThreshold^2$, the sample a_i^1 from the other modality is added to the adaptation set $\mathcal{A}'^1$ of that modality. This adaptation strategy is described in the Algorithm 2.

Algorithm 2: *Co-update.*

Input: $ref_{j(t)} = \{ref_{j(t)}^1, ref_{j(t)}^2\}$, $\mathcal{A} = \{\mathcal{A}^1, \mathcal{A}^2\}$,
 $\theta_j^{adapt} = \{adaptationThreshold^1, adaptationThreshold^2\}$

Output: $ref_{j(t+1)}$

- 1 $\mathcal{A}'^1 = \{a_i^1 \in \mathcal{A}^1 \mid classificationAlgorithm.getSimilarityScore(ref_{j(t)}^2.T, a_i^1) > adaptationThreshold^2\}$
- 2 $\mathcal{A}'^2 = \{a_i^2 \in \mathcal{A}^2 \mid classificationAlgorithm.getSimilarityScore(ref_{j(t)}^1.T, a_i^2) > adaptationThreshold^1\}$
- 3 $ref_{j(t+1)}^1 = ref_{j(t)}^1$
- 4 $ref_{j(t+1)}^2 = ref_{j(t)}^2$
- 5 $ref_{j(t+1)}^1.GL = ref_{j(t+1)}^1.GL \cup \mathcal{A}'^1$
- 6 $ref_{j(t+1)}^2.GL = ref_{j(t+1)}^2.GL \cup \mathcal{A}'^2$
- 7 $ref_{j(t+1)}^1.T \leftarrow classificationAlgorithm.train(ref_{j(t+1)}^1.GL)$
- 8 $ref_{j(t+1)}^2.T \leftarrow classificationAlgorithm.train(ref_{j(t+1)}^2.GL)$
- 9 **return** $ref_{j(t+1)} = \{ref_{j(t+1)}^1, ref_{j(t+1)}^2\}$

Later, an adaptation strategy based on graphs was proposed by (RATTANI; MARCIALIS; ROLI, 2008; RATTANI; MARCIALIS; ROLI, 2013a). This strategy was designed for mono-modal biometrics and works as follows (Algorithm 3). First, the current gallery of samples $ref_j.GL$ is joined to the unlabeled set $\mathcal{A}$ and a graph based on k-Nearest Neighbour (k-NN) is generated from this joined set. Second, the graph is divided into two parts using the *max-flow/graph min-cut* algorithm (BLUM; CHAWLA, 2001). An hypothetical example is shown on Figure 4. Finally, the nodes belonging to the genuine part (*sourcePart*) are converted into a set of samples and kept as the new gallery and the classification algorithm is trained on the new gallery. The impostor part (*sinkPart*) is discarded.

A key aspect in this algorithm is how the graph is generated, which is based on k-NN. It is generated as described next:

- It assumes that there exists two nodes named as *genuine* and *impostor*. These nodes are needed for the *max-flow/graph min-cut* algorithm and they can be called *source* and *sink*, respectively.
- The genuine (source) node (**G** in Figure 4) is connected to all nodes belonging to the gallery with infinite weight.
- The unlabeled node that has the highest average distance to the nodes in the gallery is connected to the impostor (sink) node (**I** in Figure 4) with infinite weight. The distance is based on the score output of the classification algorithm.
- Finally, all nodes (gallery and unlabeled) are connected among themselves to the k closest nodes.

Algorithm 3: *Graph min-cut.*

Input: $ref_{j(t)}$, $\mathcal{A}$, $\theta_j^{adapt} = \{k\}$

Output: $ref_{j(t+1)}$

```

1  $\mathcal{A}' = ref_{j(t)}.GL \cup \mathcal{A}$ 
2  $graph \leftarrow kNNGraph(\mathcal{A}', k)$ 
3  $(sourcePart, sinkPart) \leftarrow graphMinCut(graph)$ 
4  $\mathcal{A}'' \leftarrow convertGraphToSet(sourcePart)$ 
5  $ref_{j(t+1)} = ref_{j(t)}$ 
6  $ref_{j(t+1)}.GL = \mathcal{A}''$ 
7  $ref_{j(t+1)}.T \leftarrow classificationAlgorithm.train(ref_{j(t+1)}.GL)$ 
8 return  $ref_{j(t+1)}$ 

```

Figure 4 – Adaptation using *Graph min-cut*: hypothetical example. In this figure, after the cut, the node a_2 is added to the gallery.

This adaptation strategy does not need an additional *adaptationThreshold* as seen in Algorithm 3. However, it adds another parameter, which is the k to draw the graph. By the procedure adopted to generate the graph and considering how *max-flow/graph min-cut* works, it can be observed that $ref_{j(t)}.GL \subseteq ref_{j(t+1)}.GL$. It happens because no sample already in the gallery can be removed, meaning that the gallery will keep increasing (or at least not change) after the adaptation procedure is finished. This implies that any impostor sample wrongly introduced into the gallery will not be removed. Actually, *Self-update* and *Co-update* also incur in the same problem.

According to (RATTANI; MARCIALIS; ROLI, 2013a), *Graph min-cut* obtained higher recognition performance than *Self-update*. An advantage of this adaptation strategy is that it can capture samples with higher intra-class variation when compared to the variations that *Self-update* can capture. Even if a genuine sample has low similarity to the biometric reference and would be discarded by the stringent *adaptationThreshold* of *Self-update*, it can be added by *Graph min-cut* if this sample stays in the genuine part of the cut.

Recently, a new adaptation strategy, named *Adapted Thresholds*, was proposed by (MHENNI *et al.*, 2016). This new strategy is based on the idea of making the threshold more stringent over time and works over current *Growing* and *Sliding* adaptation strategies (KANG; HWANG; CHO, 2007) (these two adaptation strategies are described in the next section). It was proposed in the context of keystroke dynamics and it assumes that the user has higher typing behaviour variability in the first moments, and that this variability decreases as user learns how to type a given expression. In fact, as reported by (MONTALVÃO *et al.*, 2015), the users tend to stabilize their typing rhythm for the same expression over time.

Another recent work was the application of *transfer learning* to adapt the user model. In (ÇEKER; UPADHYAYA, 2016), the authors studied three adaptive versions of support vector machines (SVMs), based on different proposals (YANG; YAN; HAUPTMANN, 2007; AYTAR; ZISSERMAN, 2011) from transfer learning (TAYLOR; STONE, 2009): Adaptive SVM, Deformable Adaptive SVM and Projective Model Transfer SVM. The general concept is that, given a user model trained on enrollment data, the SVM would adapt this model for a new set of labeled samples acquired later. The number of samples in this new set would be smaller than the amount used for training. According to the reported results, the transfer learning approaches can obtain higher recognition performance than a non-adaptive SVM re-trained on the new set of samples, particularly if the amount of new samples is small. In their experiments, the SVM required training samples from both genuine and impostor classes. Consequently, a set of labeled genuine and impostor samples are required for the adaptation.

2.2.2 Online adaptation

In the online mode, the adaptation process is launched at the same time the test/recognition process occurs, i.e., after each query is received by the biometric system. Thus, the adaptation strategy has as input the biometric reference and the query, rather than a set of

unlabeled samples as in the case of the offline adaptation. Nonetheless, most online adaptation strategies are based on adapting a gallery of biometric samples/templates, similarly to the offline adaptation strategies.

A study that deals with online adaptation is (SCHEIDAT; MAKRUSHIN; VIELHAUER, 2007). In this study, the set of samples used to obtain the biometric reference is managed using cache memory algorithms. The cache memory in this case is the gallery $ref_j.GL$. Several algorithms have been described in that technical report (although none was evaluated): first in first out (FIFO), least frequently used (LFU), least recently used (LRU) and extended replacement algorithm.

FIFO replaces the oldest sample of the gallery by each query classified as genuine. LFU keeps an usage counter for each sample of the gallery, which is increased every time the sample is the closest one to the query. When a new query is classified as genuine, the sample of the gallery with the lowest value of the usage counter is replaced by the query. However, this strategy may not be entirely suitable for adaptive biometric systems. For example, in case a sample is used too many times for a short period of time and, after a while, does not represent the current user data, it will not be easily replaced, since its usage counter would have a higher value than the other samples in the gallery. The authors suggested that this problem would be mitigated if the replacement were done periodically, but the authors did not give many details on how to implement it. Another related strategy, LRU, replaces the sample less used recently. In order to implement LRU, it would be necessary to store when each sample was used, which may be costly in terms of memory usage. The authors suggest to deal with this problem by using the *Clock algorithm*. This algorithm maintains a circular list and controls the recent usage of each sample by moving a pointer through the list. Finally, the extended replacement algorithm is presented, which assigns a relevance measure for the samples and use it to perform the replacement. This measure combines both LFU and LRU but, inherits some problems from LFU, like the difficulty to remove samples with high frequency of usage even if they do not represent the current user data. Although extended replacement algorithm reduces this problem, it is still present.

Later, (FRENI; MARCIALIS; ROLI, 2008) carried one of the first studies dealing with online adaptation for biometric systems. Six adaptation strategies were assessed: RANDOM, NAIVE, LFU, FIFO, MDIST, DEND. Adaptation is done by the replacement of samples in the gallery $ref_j.GL$ by queries classified as genuine with high confidence. These strategies work as follows. RANDOM just replaces a random sample of the gallery (it is a baseline). NAIVE replaces the sample of the gallery that is closest to the query. LFU and FIFO follows the same implementation previously described in (SCHEIDAT; MAKRUSHIN; VIELHAUER, 2007). MDIST and DEND, although having the same names as the template selection methods described in the last section (ULUDAG; ROSS; JAIN, 2004), are applied in (FRENI; MARCIALIS; ROLI, 2008) as replacement methods. In MDIST, $r = |ref_j.GL|$ gallery variations are obtained, each one replacing a different sample by the query. For each of the r

galleries, the average similarity score between the samples are computed. This average value is also computed for the current gallery, without the replacement. The gallery with the highest average score is selected and, if this average score is higher than that of the current gallery, this new gallery replaces $ref_j.GL$. The rationale of this strategy is to keep similar samples in the gallery. DEND works in the opposite way. Instead of selecting the gallery with the highest average score, it selects the gallery with the lowest score. The rationale of DEND is to keep higher intra-class variability in the gallery. According to the experiments carried out by the authors on fingerprint datasets, MDIST performed better than the other adaptation strategies.

In (KANG; HWANG; CHO, 2007), two adaptation strategies were evaluated using a keystroke dynamics dataset: *Growing window* and *Moving window*. The first strategy, *Growing window*, works similarly to *Self-update* (ROLI; MARCIALIS, 2006), but assuming that $adaptationThreshold = decisionThreshold$. As a result, all queries that are classified as genuine are added to the gallery $ref_j.GL$. The second strategy, *Moving window*, works as the FIFO adaptation strategy (SCHEIDAT; MAKRUSHIN; VIELHAUER, 2007). According to the reported results, both adaptation strategies can improve the performance of the biometric system when compared to a baseline system without adaptation. Although not entirely clear in the paper, the work from (KANG; HWANG; CHO, 2007) seems to have considered only true genuine samples for adaptation, as discussed in (GIOT; DORIZZI; ROSENBERGER, 2011). In (GIOT; ROSENBERGER; DORIZZI, 2012b), both adaptation strategies were applied in a semi-supervised scenario and were named as *Growing window* and *Sliding window*, which are described in Algorithms 4 and 5, respectively.

Algorithm 4: *Growing window*.

Input: $ref_{j(t)}$, $\mathcal{A} = \{\mathbf{q}\}$, $\theta_j^{adapt} = \{adaptationThreshold\}$
Output: $ref_{j(t+1)}$

- 1 $score \leftarrow classificationAlgorithm.getSimilarityScore(ref_{j(t)}.T, \mathbf{q})$
- 2 $ref_{j(t+1)} = ref_{j(t)}$
- 3 **if** $score > adaptationThreshold$ **then**
- 4 $ref_{j(t+1)}.GL = ref_{j(t+1)}.GL \cup \{\mathbf{q}\}$
- 5 $ref_{j(t+1)}.T \leftarrow classificationAlgorithm.train(ref_{j(t+1)}.GL)$
- 6 **end**
- 7 **return** $ref_{j(t+1)}$

Both adaptation strategies from (KANG; HWANG; CHO, 2007) were later extended in (GIOT; ROSENBERGER; DORIZZI, 2012b). The authors proposed some combinations of *growing window* and *moving window*, and the combination which attained the best recognition performance was *Double Parallel*. *Double Parallel* (DB), described in Algorithm 6, manages two galleries, one by *growing window* ($ref_j.GL^1$) and another by *moving window* ($ref_j.GL^2$). Consequently, there are two models for each user too: $ref_j.T^1$ and $ref_j.T^2$. When a query is presented to the biometric system, both models return a score. The average of these scores is

Algorithm 5: Sliding window.

Input: $ref_{j(t)}$, $\mathcal{A} = \{\mathbf{q}\}$, $\theta_j^{adapt} = \{adaptationThreshold\}$
Output: $ref_{j(t+1)}$

- 1 $score \leftarrow classificationAlgorithm.getSimilarityScore(ref_{j(t)}.T, \mathbf{q})$
- 2 $ref_{j(t+1)} = ref_{j(t)}$
- 3 **if** $score > adaptationThreshold$ **then**
- 4 $s_o \leftarrow oldestSample(ref_{j(t+1)}.GL)$
- 5 $ref_{j(t+1)}.GL = ref_{j(t+1)}.GL - \{s_o\}$
- 6 $ref_{j(t+1)}.GL = ref_{j(t+1)}.GL \cup \{\mathbf{q}\}$
- 7 $ref_{j(t+1)}.T \leftarrow classificationAlgorithm.train(ref_{j(t+1)}.GL)$
- 8 **end**
- 9 **return** $ref_{j(t+1)}$

the final score. Based on this score, the decision to classify a query as genuine or as impostor is done, as well as the decision to adapt the galleries.

Algorithm 6: Double Parallel (DB).

Input: $ref_{j(t)}$, $\mathcal{A} = \{\mathbf{q}\}$, $\theta_j^{adapt} = \{adaptationThreshold\}$
Output: $ref_{j(t+1)}$

- 1 $score1 \leftarrow classificationAlgorithm.getSimilarityScore(ref_{j(t)}.T^1, \mathbf{q})$
- 2 $score2 \leftarrow classificationAlgorithm.getSimilarityScore(ref_{j(t)}.T^2, \mathbf{q})$
- 3 $finalScore \leftarrow (score1 + score2)/2.0$
- 4 $ref_{j(t+1)} = ref_{j(t)}$
- 5 **if** $finalScore > adaptationThreshold$ **then**
- 6 $s_r \leftarrow oldestSample(ref_{j(t+1)}.GL^1)$
- 7 $ref_{j(t+1)}.GL^1 = ref_{j(t+1)}.GL^1 - \{s_r\}$
- 8 $ref_{j(t+1)}.GL^1 = ref_{j(t+1)}.GL^1 \cup \{\mathbf{q}\}$
- 9 $ref_{j(t+1)}.T^1 \leftarrow classificationAlgorithm.train(ref_{j(t+1)}.GL^1)$
- 10 $ref_{j(t+1)}.GL^2 = ref_{j(t+1)}.GL^2 \cup \{\mathbf{q}\}$
- 11 $ref_{j(t+1)}.T^2 \leftarrow classificationAlgorithm.train(ref_{j(t+1)}.GL^2)$
- 12 **end**
- 13 **return** $ref_{j(t+1)}$

2.3 Evaluation

The assessment of adaptive biometric systems performance is different from the assessment of standard non-adaptive biometric systems. This is mainly due to the need to consider the adaptation process. Several different methodologies have been used with this purpose. This section describes and discusses some of these methodologies adopted in previous work, along

with metrics and datasets used for the assessment of adaptive biometric systems.

2.3.1 Methodologies

In order to present the evaluation methodologies, this work divides them into two main groups: *separate sets for adaptation and test* and *joint set for adaptation and test*. A similar division is seen in (POH *et al.*, 2009). The former group contains methodologies that use a set of samples for adaptation and another disjoint set for test, while the latter group contains methodologies that share samples for both adaptation and test. A brief discussion on several methodologies within the two groups is presented in the next sections.

2.3.1.1 Separate sets for adaptation and test

First studies of adaptive biometric systems assumed that only true genuine biometric samples could be used for adaptation and, therefore, they assumed a supervised adaptation scenario (ULUDAG; ROSS; JAIN, 2004; ROLI; MARCIALIS, 2006; ROLI; DIDACI; MARCIALIS, 2007; KANG; HWANG; CHO, 2007). This assumption may not be true for most situations where it is needed that the biometric system adapts the biometric reference automatically, unaware of true labels.

In (ULUDAG; ROSS; JAIN, 2004), the authors evaluated template selection methods and they used a dataset with two sessions (100 samples per session). Each session was divided into two parts: TRAIN (first 25 samples) and TEST (last 75 samples). In the evaluation methodology, the first 25 samples (TRAIN1) were used for the initial biometric reference, which was then tested on TEST1. During the test, samples from other users were regarded as impostors. Afterwards, the biometric reference was adapted using TRAIN2 (first 25 samples of the second session) and tested on TEST2. Note that the adaptation only occurs using TRAIN2, which is a set that just contains true genuine samples, as shown in Figure 5.

Figure 5 – Evaluation methodology in (ULUDAG; ROSS; JAIN, 2004). The numbers refer to the session index.

Later, *Self-update* was investigated on face recognition (ROLI; MARCIALIS, 2006) and another work studied *Co-update* (ROLI; DIDACI; MARCIALIS, 2007). As stated by (POH;

RATTANI; ROLI, 2012; RATTANI; MARCIALIS; ROLI, 2013b), these papers on *Self-update* and *Co-update* did not consider impostor attack during the adaptation process. Another work that did not consider impostor attack is (KANG; HWANG; CHO, 2007). As pointed out by (GIOT; DORIZZI; ROSENBERGER, 2011) and (GIOT; ROSENBERGER; DORIZZI, 2012a), the work of (KANG; HWANG; CHO, 2007) dealt with a supervised scenario, where only true genuine samples are used for adaptation. In (KANG; HWANG; CHO, 2007), each user is enrolled using 10 samples and there are 75 genuine plus 75 impostor test samples for each of them. Although not entirely clear, the graphs from Figure 4 of the paper indicate that a separate set of samples was used for adaptation, and this set only contained true genuine samples, thus, characterizing a supervised adaptation.

Apart from not considering impostor attack, those first studies on adaptive biometric systems also adopted the *separate adapt-and-test* method, since the adaptation and test sets are disjoint, and samples used for adaptation are not used for test. Such a kind of method assumes that the biometric system stays a period only adapting the biometric reference (without performing test/recognition) and, afterwards, the biometric reference is fixed (not adapted) to perform recognition only. Some recent studies have also adopted the separate adapt-and-test approach (BIGGIO *et al.*, 2012; POH; KITTLER; RATTANI, 2014; ÇEKER; UPADHYAYA, 2016).

In (POH; KITTLER; RATTANI, 2014), the experiments were carried out on a dataset containing 12 sessions. In their experiments, session 1 was used for enrollment, then sessions 2 - 4 were employed for test. Next, session 5 was used for adaptation and sessions 6 - 8 were employed for test. Finally, session 9 was used for adaptation and sessions 10 - 12 were employed for test. An overview of the evaluation methodology is shown in Figure 6. Although this work used separate sets for adaptation and test, some impostor samples were included in the adaptation set. The adopted dataset has data under three different conditions: controlled (sessions 1-4), degraded (sessions 5-8) and adverse (sessions 9-12) (BAILLY-BAILLIÈRE *et al.*, 2003). As discussed in the beginning of this chapter, adaptive biometric systems can be used to adapt the biometric reference to either changes due to time (e.g. trait ageing) or due to changing conditions (e.g. different pose in a face recognition system). Some studies on physical biometric modalities seem to mainly deal with changing conditions instead of changes uniquely due to time, which is the case of that study.

Poisoning attacks to adaptive biometric systems are studied by (BIGGIO *et al.*, 2012). The main idea of these poisoning attacks is that an attacker could progressively introduce impostor samples in the adaptation process such that the target biometric reference is changed until it can better recognize an impostor and it may not be able to recognize the genuine user anymore. The work claims to be the first to raise such issue in the area of adaptive biometric systems. To evaluate the effect of such attacks, the work used a dataset for face recognition with 60 samples per user. A random subset of 10 samples was used for the initial enrollment and another subset of 10 samples was used for parameter tuning. The remaining 40 samples

Figure 6 – Evaluation methodology in (POH; KITTLER; RATTANI, 2014). The numbers refer to the session index.

were then part of the test. Then, a separate set of poisoning samples was used to adapt the biometric reference. However, the work only considered that the biometric reference is adapted with impostor patterns from the poisoning set. This may not correspond to a practical scenario, since both genuine and impostor samples can be used for adaptation. Consequently, the effect of poisoning could be reduced. The work from (BIGGIO *et al.*, 2012) also does not consider the chronological order of the genuine biometric samples, as it used random samples for the enrollment.

2.3.1.2 Joint set for adaptation and test

As illustrated in the last section, several studies have adopted the separate adapt-and-test method. However, this method does not make an optimal use of the available data, since the data used for adaptation are not used for test and vice-versa. This is a critical issue in the area as the availability of large datasets for studying adaptive biometric systems is limited. Another method consists of using unlabeled data for adaptation and using the same samples to test the recognition performance. It is known as *joint adapt-and-test* method (POH; RATTANI; ROLI, 2012). In addition to using more data for both adaptation and test, this method may also better reflect a practical scenario, where the system, once deployed, has to perform the recognition of all query samples and cannot stop the recognition to use some of the received samples for an adaptation period.

An important work in the area which proposed an evaluation methodology within the joint adapt-and-test is (RATTANI; MARCIALIS; ROLI, 2013b). A similar methodology was used in another work from the same authors in (RATTANI; MARCIALIS; ROLI, 2013a). The methodology was based on the DICE dataset, which has several sessions per user (10 samples per session), and is divided into three main parts (Figure 7):

- Part A (enrollment): this is when the enrollment is done. For such, the first two samples of the first session ($t = 1$) are used.
- Part B (adaptation): in this part, the adaptation is performed. For each user, an adapta-

tion set is formed, consisting of the samples from session t plus five random impostor samples. This is important since it considers that there may be impostor attacks during the adaptation. The first session used for adaptation is $t = 1$. However, in this particular case, the first two samples are discarded since they were used for enrollment, while, in the other sessions, all 10 samples are part of the adaptation set. This adaptation set is then presented to the adaptation strategy to perform adaptation. For *Self-update*, only those samples that reach a score above an update threshold are considered for adaptation, while the others are discarded.

- **Part C (test):** finally, the recently adapted biometric reference is used for test on the next session ($t + 1$). The first test session is $t = 2$, since the first adaptation session is $t = 1$. Samples from the test session from all other users are regarded as impostors. Note that the biometric reference is not adapted during the test. When the test is finished on session $t + 1$, Part B (adaptation) is launched again on the same session $t + 1$. The adapted biometric reference is then tested on session $t + 2$ and so on.

Figure 7 – Evaluation methodology in (RATTANI; MARCIALIS; ROLI, 2013a). The numbers refer to the session index $t \in [1; W]$ (there are W sessions in the dataset).

In this methodology, the last session is only used for test, while the very first session is just used for enrollment and adaptation. However, all other sessions are used for both adaptation and test, increasing the amount of samples for both processes. In a dataset with several sessions, such as the case of the DIEEE used in their experiments, this methodology allows the assessment of several adaptation cycles. Nevertheless, the division into sessions informs the adaptation strategy when adaptation should be launched and this information may not be available in a practical scenario.

Another evaluation methodology similar to (RATTANI; MARCIALIS; ROLI, 2013a) was adopted in (PAGANO *et al.*, 2015), though with some fundamental modifications. First, in order to standardize the performance evaluation, the samples from all users were divided into six batches. This was done because three datasets were considered in their experiments and the number of sessions was different among them. The fact that the evaluation methodology informs

when adaptation should be triggered also occurs in (PAGANO *et al.*, 2015), since adaptation is performed at the end of each batch. Second, after enrollment, the first process called is the test, followed by the adaptation. This implies that the remaining samples from the first batch are used for test and adaptation, while in the methodology from (RATTANI; MARCIALIS; ROLI, 2013a), these samples were used just for adaptation. Third, the amount of impostor samples is the same as the amount of genuine samples. They are also randomly selected, as in (RATTANI; MARCIALIS; ROLI, 2013a). However, the ratio of impostor samples in the adaptation set is 33% in (RATTANI; MARCIALIS; ROLI, 2013a) and 50% in (PAGANO *et al.*, 2015). An overview of the methodology adopted in (PAGANO *et al.*, 2015) is shown in Figure 8.

Figure 8 – Evaluation methodology in (PAGANO *et al.*, 2015). The numbers refer to the batch index (there are six batches).

In the methodology adopted in (RATTANI; MARCIALIS; ROLI, 2013a), not all impostor samples from the test are included in the adaptation set, but, instead, five random impostor samples are selected. Although the purpose of doing that is to better reflect a practical scenario, this method is not entirely sound. In a practical scenario, all query samples used for test should be part of the adaptation set. Unless the true label is provided, the system would not be able to select five impostor samples from the queries. Although not entirely clear in (PAGANO *et al.*, 2015), it seems that the same impostor samples used in the batch for test are part of the adaptation set.

This issue concerning the inclusion of impostor samples is addressed by the methodology adopted in (GIOT; ROSENBERGER; DORIZZI, 2013). A similar approach can be seen in (GIOT; ROSENBERGER; DORIZZI, 2012b). In this methodology, the samples from the first session are used for enrollment. The remaining sessions are then used for test and adaptation using *pools*. A *pool* is a sequence of queries, containing both genuine and impostor samples. The number of impostor samples in the *pool* is a function of the number of available genuine samples of the considered session. The number of impostor samples is defined according to the Equation 2.1, where N^I is the number of impostor samples, N^G is the number of genuine

samples available for the session and r is the impostor ratio.

$$N^I = N^G \times \frac{r}{1 - r} \quad (2.1)$$

After defining the number of impostor samples, the *pool* is created using all genuine samples from the session plus N^I random impostor samples. Genuine and impostor samples are randomly interleaved, but keeping the chronological order of the genuine samples. The *pool* is then presented, query by query, to the algorithm, which will perform the test on each query and also adapt the biometric reference using the query, depending on its rules (e.g. a more stringent threshold). By adopting this strategy, all sessions are used for test and adaptation, except the first one, which is dedicated for enrollment, as shown in Figure 9. This method also better reflects a practical scenario, as all queries are used to test and are also considered during the adaptation process.

Figure 9 – Evaluation methodology in (GIOT; ROSENBERGER; DORIZZI, 2012b). The numbers refer to the session index.

In order to simulate a scenario where the genuine user most frequently attempts to authenticate, the work from (GIOT; ROSENBERGER; DORIZZI, 2012b) adopted the impostor ratio $r = 30\%$. This methodology also used it to check the impact of changing the impostor ratio of the *pool* (GIOT; ROSENBERGER; DORIZZI, 2013). A higher number of impostor samples results in a higher likelihood of wrongly choosing an impostor sample for adaptation. According to results obtained by the authors, the recognition performance does decrease as the impostor ratio increases. However, even under this lower performance, the use of adaptation still resulted in higher performance than a system without adaptation.

This methodology adopted a different method to estimate the decision threshold and the adaptation threshold. The decision threshold is defined by the EER¹ (Equal Error Rate) computation, so it is not fixed over time. A discussion regarding the metrics and the use of EER is provided in the next section. The adaptation threshold, on the other hand, is obtained from the enrollment data. By using the well-known leave-one-out method on the enrollment samples, the threshold which obtains the EER is chosen as the adaptation threshold. This value

¹ EER is described in Section 2.3.2

is the *permissive threshold*. Another threshold is obtained when the FMR² (False Match Rate) is equals to 1%, using the same leave-one-out method. This additional threshold is named as *stringent threshold*. According to their experiments, the performance of the adaptive biometric system using the permissive threshold was higher than the system using the stringent threshold. However, the performance of the biometric system with a stringent threshold is less affected by changes in the ratio of impostor samples.

2.3.2 Metrics

The main metrics used for the evaluation of adaptive biometric systems are shown next (HIMAGA; KOU, 2008; Precise Biometrics, 2014; POH; KITTLER; RATTANI, 2014). These metrics have also been used to assess the recognition performance³ of non-adaptive biometric systems.

- FNMR (False Non-match Rate): rate of genuine *attempts* that were wrongly rejected (classified as impostor), as defined by Equation 2.2. A related metric is FRR (False Rejection Rate), which has almost the same definition of FNMR, but FRR also takes into account the FTA (Failure to Acquire Rate), as shown in Equation 2.3. FTA is the rate in which the system fails to obtain a biometric sample.

$$FNMR = \frac{\text{number of rejected genuine attempts}}{\text{number of genuine attempts}} \quad (2.2)$$

$$FRR = FTA + FNMR \times (1 - FTA) \quad (2.3)$$

- FMR (False Match Rate): rate of impostor *attempts* that were wrongly accepted (classified as genuine), as defined by Equation 2.4. A related metric is FAR (False Acceptance Rate), which has almost the same definition of FMR, but FAR also takes into account the FTA, as shown in Equation 2.5.

$$FMR = \frac{\text{number of accepted impostor attempts}}{\text{number of impostor attempts}} \quad (2.4)$$

$$FAR = FMR \times (1 - FTA) \quad (2.5)$$

- HTER (Half Total Error) and balanced accuracy: HTER is defined here as the average between FNMR and FMR (Equation 2.6). This metric is able to combine the results from both FNMR and FMR in a single value, which can simplify the study of the performance. Another way to view this rate is in terms of balanced accuracy (MASSO; VAISMAN,

² FMR is described in Section 2.3.2

³ Throughout this thesis, the term *performance* refers to the *recognition performance* of a biometric system, unless specified differently in the text. In Machine Learning, recognition performance is also named as *predictive performance*.

2010). The balanced accuracy in the context of this thesis is defined as shown in Equation 2.7.

$$HTER = \frac{FNMR + FMR}{2} \quad (2.6)$$

$$BAcc = 1 - HTER \quad (2.7)$$

- EER (Equal Error Rate): by changing the parameters⁴ of the classification algorithm (e.g. the decision threshold), FMR and FNMR can be either increased or decreased. Increasing FMR usually implies that FNMR is decreased and vice-versa. There is a certain point in which FMR reaches the same value of FNMR. This is the EER, that can be understood as a particular case of HTER, when $FMR = FNMR$.

In addition, the following metrics were proposed for adaptive biometric systems (GIOT; ROSENBERGER; DORIZZI, 2012b). Note that they assume that only genuine samples should be used for adaptation, which is the case for most adaptation strategies.

- IUSR (Impostor Update Selection Rate): rate of impostor samples involved in the adaptation process, as defined by Equation 2.8.

$$IUSR = \frac{\text{number of impostor samples involved in the adaptation process}}{\text{number of tested impostor samples}} \quad (2.8)$$

- GUMR (Genuine Update Miss Rate): rate of genuine samples not involved in the adaptation process, as defined by Equation 2.9.

$$GUMR = \frac{\text{number of genuine samples not involved in the adaptation process}}{\text{number of tested genuine samples}} \quad (2.9)$$

The presented metrics can be computed globally (an average of the performance in the complete experiment) or over time. The computation over time can be done by getting the average results per session. The work from (RATTANI; MARCIALIS; ROLI, 2011a) claims to be the first paper in the area to report results over time, instead of just obtaining it globally.

Still regarding those metrics, there are some important remarks. Most experiments consider datasets only with successful acquired biometric samples, so FTA is not measured. Thus, the recommended approach is to adopt the metric FNMR/FMR instead of FRR/FAR.

Several studies in the area report the results in terms of EER. However, as stated in (BENGIO; MARIÉTHOZ; KELLER, 2005), reporting results using this metric can be misleading. The main argument is that the EER can only be obtained by testing several parameter values

⁴ In this thesis, the term *parameters* is used to refer to the *hyper-parameters* of the algorithm.

on the test data, until false match equals false non-match. Nevertheless, such procedure to obtain the parameters may not be feasible in a practical scenario. Furthermore, when evaluating the performance over time (e.g. by session), the parameters (e.g. threshold) may differ from one test session to another, as highlighted by (GIOT; DORIZZI; ROSENBERGER, 2011).

In view of these problems, a more realistic procedure to evaluate adaptive biometric systems is to tune the parameters in the enrollment data and then apply the system with the tuned parameters in the test data. In this case, the results are reported in terms of FMR and FNMR for a given set of parameter values (e.g. threshold) obtained from the enrollment data only.

2.3.3 Modalities and datasets

A summary of several datasets that have been used to evaluate adaptive biometric systems performance is presented in Table 1. There are not many datasets available for the evaluation of adaptive systems. These datasets need to contain several samples per user, which should ideally be acquired at different sessions.

Table 1 – Datasets that have been used to evaluate adaptive biometric systems.

Datasets	# Users	Period/Sessions	Modalities
GREYC (GIOT; EL-ABED; ROSENBERGER, 2009)	100	2 months (5 sessions)	Keystroke dynamics
GREYC-Web (GIOT; EL-ABED; ROSENBERGER, 2012)	118	more than 1 year	Keystroke dynamics
CMU (KILLOURHY; MAXION, 2010)	51	8 sessions	Keystroke dynamics
AR Face Database (MARTINEZ; BENAVENTE, 1998)	116	14 days (2 sessions)	Face
Dataset from (BIGGIO <i>et al.</i> , 2012)	40	2 sessions	Face
BANCA 2D (BAILLY-BAILLIÈRE <i>et al.</i> , 2003)	52	12 sessions	Face
DIEE multi-modal (MARCIAS <i>et al.</i> , 2012)	49	1.5 years (10 sessions)	Face and fingerprint
FVC-2002 DB2 (MAIO <i>et al.</i> , 2002)	110 fingers	3 sessions	Fingerprint
Dataset from (ULUDAG; ROSS; JAIN, 2004)	50 fingers	aprox. 4 months (2 sessions)	Fingerprint
Fenker (FENKER; BOWYER, 2012)	322	aprox. 4 years	Iris
ELDASR (M. <i>et al.</i> , 2016)	50	20 samples per user	Voice

As seen in Table 1, most datasets are for physical biometric modalities, mainly face and fingerprint. This reflects the higher amount of studies on physical modalities in the area of adaptive biometric systems. There is a huge variability in terms of number of users and the period/sessions among the datasets. The higher the number of users and sessions, the higher is the reliability of the reported results.

The focus of the current thesis is on behavioural modalities, particularly keystroke dynamics and accelerometer-based gait biometrics. This thesis employed all keystroke dynamics datasets mentioned in Table 1. In addition, three other accelerometer datasets were part of the experiments (they are described in Section 3.4.2). To the best of the author's knowledge, the research from this thesis is the first to study adaptation strategies for accelerometer-based gait biometrics.

2.4 Chapter remarks

Authentication systems can be improved by employing biometric systems, which rely, among other aspects, on the permanence of the biometric features. However, recent studies have shown that this requirement may not hold for several biometric modalities and, therefore, the recognition performance of biometric systems can be affected. In order to deal with this issue, adaptive biometric systems have been proposed. These systems are able to adapt the biometric reference to take into account changes observed over time. This chapter presented an overview of the area of adaptive biometric systems, focusing on adaptation strategies and aspects concerning the performance evaluation. Since adaptive biometric systems is a relatively new area, there are several opportunities for research topics. This thesis focuses on adaptation for single modality, so it does not cover adaptation strategies that combine results from two modalities, like *Co-update*.

As discussed in the beginning of this chapter, adaptive biometric systems can be used to adapt the biometric reference to changes either due to *time* (e.g. trait ageing) or due to changing *conditions* (e.g. different pose in a face recognition system). Some studies on physical biometric modalities seem to mainly deal with changing conditions instead of changes due to time, impacting a key aspect of the proposed adaptation strategies: removal of outdated patterns from the gallery. For instance, an adaptation strategy that needs to expand the gallery to new *conditions* does not need to discard older samples. A key example shown in this chapter is *Growing/Self-update*. Conversely, an adaptation strategy that deals with changes due to *time* has to add the newest samples to the gallery as well as remove the outdated samples, since these outdated samples do not represent the current user data anymore. A key example of a strategy able to fit this case is *FIFO/Sliding*. Although not properly investigated so far, this thesis conjectures that behavioural biometric modalities are more subject to changes due to time than physical modalities (e.g. the user learned to type faster in the same condition/keyboard). This suggests that adaptation strategies for behavioural modalities should also be able to discard older samples from the gallery.

Still regarding adaptation strategies, this thesis focuses on some gaps found in the literature. First, controlling the usage of samples in the gallery was introduced in a technical report (SCHEIDAT; MAKRUSHIN; VIELHAUER, 2007), although the authors did not report any experiments. This thesis expands this concept and applies it to the *Self-Detector*, an immune-based classification algorithm which performed well for keystroke dynamics (PISANI, 2012; PISANI; LORENA, 2015). It is investigated in Chapter 4. Second, almost all adaptation strategies are solely based on a genuine model of the user and, consequently, adaptation consists of adapting a single genuine gallery. However, the use of an additional impostor gallery could improve the performance of the biometric system. This is explored in Chapter 5. Moreover, score normalization, known to be able to improve biometric systems performance by a refinement of the classification decision (POH; MERATI; KITTLER, 2009), has not been deeply explored

for adaptive biometric systems. A preliminary study on the use of score normalization for supervised adaptation to cope with different acquisition conditions can be seen in (POH *et al.*, 2010). However, score normalization has not been applied to adaptive biometric systems in a data stream context with unsupervised adaptation. This is studied in Chapter 6. Finally, to the best of the author's knowledge, all adaptive biometric systems proposed so far apply the same adaptation strategy to all users, even though characteristics, including the change pattern, may differ among the users registered in the biometric system. To this end, a biometric system able to choose the adaptation rules per user is investigated in Chapter 7.

Apart from the gaps regarding adaptation strategies, the evaluation methodologies that have been adopted in previous work are not entirely suitable to biometrics in a data stream context. This is due to several aspects, such as considering session division to perform adaptation (information that may not be available in a practical scenario), use of true labels for adaptation, presence of impostor attacks from both registered and unregistered users and evaluation metrics (e.g. EER). Next chapter discusses all of these aspects and presents an evaluation methodology for biometrics in a data stream context, which is proposed by the research carried for this thesis.

EVALUATION METHODOLOGY

As shown in the previous chapter, different evaluation methodologies have been used to assess the performance of adaptive biometric systems. However, none of them are entirely suitable to evaluate *biometrics in a data stream context*. The study of biometrics in this context considers a scenario where a potentially unlimited sequence of biometric samples is presented, one by one, to the biometric system. A biometric system needs to classify each sample as either genuine or impostor, then decide if the biometric reference should be adapted. The classification decision should be made before the next sample is received, although the decision to perform adaptation can be postponed. A key aspect in the data stream context considered in this thesis is that *the decision to perform adaptation is taken by the biometric system* and not determined by the evaluation methodology.

Evaluation methodologies presented in the previous chapter define when to perform adaptation in different ways, mainly due to the session-oriented approach. Several papers divide the dataset into sessions and use this additional information to decide when the biometric reference is adapted. Some studies only perform adaptation at the end of each session using the data acquired from the previous session. Other papers use disjoint sets for test and adaptation. Not only does such an approach tell the moments to perform adaptation, but it also does not make an optimal usage of the data. Instead of using a disjoint set for adaptation, a better usage of the data is to employ all test samples for adaptation too. This thesis adopts this principle.

There are even some papers that just use true genuine samples for adaptation, an assumption that could be considered unrealistic in several scenarios. More recent studies have considered that impostor samples may be wrongly used by the adaptation process. However, the source of impostor samples is another issue. These samples can be from known users of the biometric system (internal attacks) or from unknown users (external attacks). The evaluation methodology proposed during the research for this thesis considers both types of attacks.

Finally, a number of studies in the area have reported the results in terms of EER. As discussed in Section 2.3.2, this implies that the algorithm parameters (e.g. threshold) are adjusted

using the test data. When the result is reported per session, a different set of parameter values is chosen per session in order to obtain the EER. Reporting results using this method of parameter setting may be misleading, since the parameters need to be tuned before the classification in a practical scenario, i.e., at enrollment time. The experiments of this thesis tune the parameters using the enrollment samples and does not report results in terms of EER.

All in all, the previous arguments do not mean that previous evaluation methodologies are inappropriate to evaluate adaptive biometric systems. They can be effective in the context of their studies. However, they would not provide reliable results for the context of this thesis, which deals with *biometrics in a data stream context*. Hence, a new evaluation methodology, named *user cross-validation for biometric data streams* was proposed, as described in the next sections. These next sections are organized as follows: Section 3.1 presents definitions adopted in this chapter; Section 3.2 introduces the user cross-validation for biometric data streams methodology; Section 3.3 shows the metrics used to report the results and how they are computed to assess the performance over time; Section 3.4 describes the biometric modalities and datasets used in the experiments; Section 3.5 enumerates the baseline algorithms adopted in the experiments; Section 3.6 discusses the statistical test applied in this thesis; and, Section 3.7 presents the final remarks of the chapter.

3.1 Biometrics in a data stream context

A biometric system has two main phases: enrollment and test/recognition. The enrollment is done using the enrollment samples, which are usually the first samples acquired from the genuine user. Afterwards, the test/recognition phase takes place. *Biometrics in a data stream context* considers a scenario where a potentially unlimited sequence of biometric query samples is presented, one by one, to the biometric system.

This sequence of queries is named as a *biometric data stream* in this thesis (Equation 3.1). The definition of a biometric data stream is similar to the definition of a data stream in Machine learning (AGGARWAL *et al.*, 2003; BIFET *et al.*, 2009; GAMA, 2010). Each query sample in the sequence is a multidimensional vector. The queries in the biometric data stream are chronologically ordered, i.e., the query $\mathbf{q}_t$ is presented to the biometric system before the query $\mathbf{q}_{(t+1)}$. It is important to highlight that it does not mean that the queries appear at regular intervals, as a temporal sequence (HAN, 2006). Note that under a data stream context, there is no division into sessions, but rather, a single sequence of queries in a data stream.

$$ds = \mathbf{q}_1, \mathbf{q}_2, \mathbf{q}_3, \dots \quad (3.1)$$

The processing of data streams involves several aspects (GAMA, 2010). They are described in the context of biometrics as follows:

- Sequential access: the queries are accessed sequentially, in the order that they are received

by the biometric system. Accessing older queries can occur only if the queries are stored by the biometric system during its operation.

- Limited computational resources: as the data stream is potentially unlimited, it is not possible to store the complete stream in memory. Besides, it is desirable that algorithms working under a data stream context process each query before receiving the next one.
- Concept drift: the distribution that generates the biometric samples can change over time. This is related to the fact that the biometric samples from the genuine user may undergo changes, which can result in template ageing.

In the biometric data stream, the queries can be either genuine or impostor attempts. Considering a given *genuine* user j ($j \in \mathcal{J}$), genuine queries are biometric samples from this user. A genuine user can also be referred to as *reference* or *target* user. With reference to this target user, a biometric sample of a registered user $i \neq j$ from the registered set $\mathcal{J}$ is considered an impostor sample. This is considered an *internal* attack. Similarly, any sample that belongs to any non-registered user is also considered an impostor sample. However, in this case, it is an *external* attack. This relates to the problem of *open-set recognition* (LI; WECHSLER, 2005; SCHEIRER *et al.*, 2013), where part of the impostor users used for test are not present at enrollment time.

In short, a biometric data stream contains queries from both genuine and impostor users. This thesis also assumes that the biometric system never receives the true labels of the queries, since this information may not be obtained in a practical scenario. The biometric system has to classify each query and, at the same time, adapt the biometric reference over time according to its adaptation strategy. In this context, both *test* and *adaptation* processes can be merged into a single process, named *testAndAdapt*:

$$(label^p, ref_{j(t+1)}) \leftarrow testAndAdapt(ref_{j(t)}, \mathbf{q}, \theta_j^{verify}, \theta_j^{adapt})$$

Note that biometrics in a data stream context as defined here means that the set of samples for adaptation is $\mathcal{A} = \{\mathbf{q}\}$. Most baseline systems can implement this new process *testAndAdapt* by just calling the *test* and the *adaptation* process sequentially. However, the proposals developed in this thesis are presented as a single joint process *testAndAdapt*.

The next section presents an evaluation methodology for biometrics in a data stream context, which was proposed by the research carried out for this thesis.

3.2 User cross-validation for biometric data streams

The *user cross-validation for biometric data streams* methodology proposed by this research follows concepts similar to the k -fold cross-validation traditionally used in Machine learning (FLACH, 2012). However, as the name suggests, user cross-validation performs the sampling at the user index set $\mathcal{D}$, as opposed to the samples directly. A sample level k -fold

validation is not suitable in the current context because the order of the samples needs to be maintained to study the behaviour change. This subject-level cross-validation is used to divide which are the registered and the non-registered users. An overview of the user cross-validation for biometric data streams methodology can be seen in Figure 10. The set $\mathcal{D}$ contains user indexes from all users in a given dataset.

Figure 10 – Overview of the user cross-validation for biometric data streams. Adapted from (PISANI *et al.*, 2016).

First, the whole list of users is randomly divided into k sets of similar size. Thus, k experiments can subsequently be performed (k assumed the value 5 in the experiments for this

thesis). For each experiment, one set of users is employed as *non-registered* users and the remaining $k - 1$ sets form the *registered* users, forming different *user groupings* (they relate to the folds). The registered users are those known by the system and they form the $\mathcal{J}$ set. All k combinations of user sets are tested, so that every user is considered once as a non-registered user. Note that $\mathcal{J} \neq \mathcal{D}$, although $\mathcal{J} \subset \mathcal{D}$.

For each user in the $\mathcal{J}$ set, i.e. for each registered user, an execution is performed. An execution consists of two phases: enrollment and test+adaptation. The first N samples from each genuine user $j \in \mathcal{J}$ are employed to perform enrollment to obtain the initial biometric reference ref_j . As discussed in (RATTANI; MARCIALIS; ROLI, 2013b), the number of samples for enrollment should be fixed to allow a proper comparison among different algorithms. Additionally, behavioural biometric modalities, such as keystroke dynamics, usually need a higher number of enrollment samples to build the biometric reference than physical modalities. For instance, as discussed in (GIOT *et al.*, 2011), using less than 10 enrollment samples results in low performance while using around 40 samples seems to be the required amount. This number was adopted for the experiments of this thesis.

Afterwards, a biometric data stream is generated to assess the recognition performance. Each query in the data stream is firstly tested and, depending on the adaptation strategy, it can also be used for adaptation. This process of test and adaptation is repeated query by query, until all the data from the data stream are queried. The next section describes how this data stream is generated.

3.2.1 Biometric data stream generation

The data stream used in the experiments simulates a scenario where data samples arrive continuously, rather than arriving in sessions, although the actual data may have been collected at different sessions. Consequently, the decision whether to perform adaptation or not depends on the adaptation strategy. As no information on sessions is sent to the biometric system, this choice of a data stream better represents a practical scenario. The session information is actually a characteristic of the datasets used in the experiments and not really a concept present in practical implementations. Regarding the composition of the data stream, there are several aspects to be considered:

- **Chronological order:** As this thesis deals with adaptation to changes of the features over time, observing the chronological order is essential. This is done for both enrollment and test. First, enrollment is performed using biometric samples acquired earlier than those used to test. Second, the biometric samples of each user are presented in chronological order in the data stream for test+adaptation.
- **Ratio of genuine versus impostor samples:** The biometric data stream should contain both genuine and impostor queries to properly evaluate the biometric system performance. However, which ratio of genuine/impostor queries should be considered? A

balanced case (50% of samples for both classes) may not correctly reproduce a practical scenario. In an access control scenario, usually, the majority of authentication attempts are likely to come from the genuine user. Indeed, some previous studies have assumed this during the adaptation process (GIOT; ROSENBERGER; DORIZZI, 2012b; RATTANI; MARCIALIS; ROLI, 2013b). In (GIOT; ROSENBERGER; DORIZZI, 2012b), which deals with keystroke dynamics, the experiments adopted the ratio of 70% genuine samples / 30% impostor samples. This thesis adopts the same ratio to generate the biometric data stream.

- **Source of the impostor samples:** As stated previously, this study generates a data stream with 30% of impostor samples, but from where should they be obtained? There are two different sources: the registered users in the dataset (excluding the genuine one) or a separate dataset of unregistered users. The user cross-validation for biometric data streams methodology uses a combination of these two sources. As described in the previous section, two sets of users are obtained for each one of the k user groupings: *registered* and *non-registered* users. In the proposed methodology, when an impostor sample is selected for the data stream, there is a 50% chance of being drawn from a registered user and 50% of being from an unregistered user. As a result, internal and external attacks can be simulated in the system. This related to *open-set recognition* (LI; WECHSLER, 2005; SCHEIRER *et al.*, 2013), where part of the impostor users used for test are not present at enrollment time.

Another important issue is regarding the Doddington's Zoo phenomenon (DODDINGTON *et al.*, 1998). According to this phenomenon, four groups of users can affect the recognition performance of a biometric system: *sheeps* - users who can be easily recognized by the system; *goats* - users that are difficult to be recognized; *lambs* - users that are easily imitated; *wolves* - users that are good at imitating other users. Goats, lambs and wolves pose different problems to adaptive biometric systems: biometric reference from goats may not be successfully adapted, biometric reference of lambs are subject to the wrong inclusion of impostor data and wolves can introduce impostor data in the biometric references. In (RATTANI; MARCIALIS; ROLI, 2009), it was raised the issue that this phenomenon has to be considered during the design of adaptive biometric systems. However, according to their experiments, no significant performance degradation was observed. By using the user cross-validation for biometric data streams methodology, all users are employed as registered and unregistered users and the results are averaged, decreasing the impact of a possible bias due to the choice of a specific partition of users for the registered and the unregistered sets in view of this phenomenon.

Moreover, this thesis only studies *zero-effort attacks* (GIOT; ROSENBERGER; DORIZZI, 2013; POH; WONG; MARCIALIS, 2014) by using samples from different users as impostor attacks. Hence, advanced spoofing attacks are not studied in this thesis.

- **Sequence of samples:** There are several ways to present the sequence of queries. In an extreme case, all genuine queries could be presented first, followed by all impostor queries in the end (or vice versa, all impostors first, followed by all genuine queries). However, this would unlikely occur in a practical application scenario. Hence, the alternative adopted here is to randomly interleave genuine and impostor queries. The experiments are repeated 30 times to avoid any bias due to a particular random query interleave.
- **True labels:** Several studies on data streams (ŽLIOBAITĖ *et al.*, 2015) assume that the true label is given to the classifier some time after the prediction. It is a feasible assumption for some applications, such as stock market forecasting, for example. Nevertheless, for biometric systems, the true label is not given to the classifier in most implementations. Hence, in this thesis, the true label is *never* provided to the classifier during the experiments. Therefore, an adaptive biometric system should take the decision to adapt or not using a given sample without ever knowing its true label (the adaptation is unsupervised).

A hypothetical example of how a data stream is generated is shown in Figure 11. It presents the execution for a single user, where an enrollment set ($\mathcal{E}_j$) and a biometric data stream (ds) are provided. The enrollment data ($\mathcal{E}_j$) has N genuine samples and the biometric data stream (ds) has M genuine samples and K impostor samples. As it was considered that there are 30% of impostor samples in the data stream, then $K/(K+M) = 0.3$. Each individual query in ds is presented one after another. Genuine and impostor queries are randomly interleaved, as previously discussed.

Figure 11 – Biometric data stream. Adapted from (PISANI *et al.*, 2016). The enrollment data ($\mathcal{E}_j$) has N genuine samples. The biometric data stream (ds) contains M genuine samples and K impostor samples.

3.2.2 Database-aware biometric systems and parameter setting

This study assumes that the biometric system can share data among all registered users, similar to (BENGIO; MARIÉTHOZ, 2007). This approach is useful due to several aspects, mainly for obtaining impostor-like data. For instance, apart from the genuine user data, data from other registered users ($\mathcal{J}$) could be accessed as impostor data to improve the biometric reference. This additional data also helps to tune algorithm parameters. It is called as a *database aware biometric system* in this work. This assumption is used for the score normalization and the modular adaptive biometric system studied in Chapters 6 and 7, respectively.

However, the use of the database aware assumption inadvertently introduces an additional challenge during the experimental evaluation. Assuming a hypothetical scenario of a

dataset containing 100 users, a straightforward evaluation protocol would consider that the biometric system has access to data from all 100 users. Consequently, during enrollment, one user could adjust its biometric reference to protect itself from all other 99 users. Nonetheless, this experimental protocol would be biased, since all possible impostors seen during test are previously known by the system during the enrollment. This is avoided by the user cross-validation for biometric data streams methodology, since not all users from the dataset are part of the set $\mathcal{J}$. The simulation of external attacks by the proposed evaluation methodology (due to users not known to the system) becomes even more important when the system takes into account information from all registered users ($\mathcal{J}$), as in the case of some score normalization procedures and the modular system studied in Chapters 6 and 7.

As described in the next sections, the adaptation strategies adopted here work on top of non-adaptive classification algorithms. In order to properly compare the effect of the adaptation strategies, the parameter tuning is performed on the non-adaptive system and the adaptive counterparts assume the same parameter values. An approach related to the *database aware biometric system* is used to tune all the algorithms. The evaluation procedure using cross-validation for biometric data streams described in the previous sections is employed to evaluate several combinations of parameter values (similar to a *grid search* method). However, instead of using the complete dataset, only the enrollment samples from all users are employed for parameter tuning. The best combination of parameter values in terms of balanced accuracy is selected for the test on the complete dataset. In the current implementation, the parameter tuning is performed globally, i.e., all users assume the same parameter values. As the tuning is executed for the non-adaptive algorithms, the *adaptation threshold* is not tuned, so the experiments assume that this parameter of the adaptation strategies adopt the same value of the *decision threshold*.

Note that by adopting such a procedure to tune the parameters, the performance metrics FMR and FNMR are reported for a given set of parameter values, instead of using EER. This follows some guidelines present in (BENGIO; MARIÉTHOZ; KELLER, 2005), which argues that reporting results in terms of EER can be misleading. The premise for this statement is that EER is usually obtained by testing several threshold values on the test data (until FNMR equals to FMR), which may not be feasible in a practical application scenario. A more realistic procedure is to tune the parameters in the enrollment data, and then apply the obtained values in the test data.

3.3 Metrics

Common biometric metrics used in previous studies in the area are used here, such as FMR and FNMR (POH *et al.*, 2009; GIOT; ROSENBERGER; DORIZZI, 2012b; FENKER; BOWYER, 2012). Additionally, the balanced accuracy is used as a measure that combines both FMR and FNMR into a single value. These measures are properly defined in the last chapter on Section 2.3.2.

FMR and FNMR are also evaluated over time throughout the thesis. In order to perform such an evaluation, an improved version of the plot adopted in a previous paper (PISANI; LORENA; CARVALHO, 2015a) is used here. This improved version is presented in (PISANI *et al.*, 2016). In this plot, the FMR and FNMR are extracted from the biometric data stream using a window of size wds_{size} in steps of wds_{step} samples (the windows overlap). This thesis adopted $wds_{size} = 50$ and $wds_{step} = 10$, the same values from (PISANI *et al.*, 2016). The average values over all genuine users of the first user grouping of the *user cross-validation for biometric data streams* are reported. In this plot, if the users contain a different number of samples in the biometric data stream, the later parts of the plot cover the average of a lower number of users.

To overcome this problem, the plot also shows the interval based on the *standard error of the mean* (shaded area), as described in Equation 3.2 (CI_i : confidence interval), which makes use of the standard error SE calculated in Equation 3.3. In Equation 3.3, std_i is the standard deviation of the values at window i and $users_i$ is the number of users with available data at window i . This interval provides additional data to support the discussion of the results. An example of this plot is seen in Figure 15. The GREYC dataset (described later, in Section 3.4.1) is the single dataset not studied over time using this plot, since its biometric data streams have the lowest number of samples and it could be difficult to check the performance change over time.

$$CI_i = mean(measure)_i \pm 1.96 \times SE_i \quad (3.2)$$

$$SE_i = std_i / \sqrt{users_i} \quad (3.3)$$

3.4 Biometric modalities and datasets

This thesis focuses on behavioural biometric modalities, particularly on keystroke dynamics (PISANI; LORENA, 2013) and accelerometer-based gait biometrics / accelerometer biometrics (SPRAGER; ZAZULA, 2009; KAGGLE, 2013). These modalities have been chosen due to the availability of suitable datasets for this study. A dataset needs to contain several samples for each user and ideally be acquired at different sessions, which largely reduces the amount of datasets that are appropriate for studying adaptation. The next sections briefly describe each of the chosen biometric modalities, along with their datasets and features extracted to obtain the biometric samples. This thesis only makes use of publicly available datasets in order to allow the reproducibility of the reported results.

Moreover, most of the studies in the area of adaptive biometric systems have dealt with physical modalities, indicating a gap for new research on behavioural modalities. These modalities usually imply in lower discriminative power than physical modalities (e.g. face, fingerprint,

iris), introducing additional challenges to the biometric system. Furthermore, they tend to be subject to faster changes of the biometric features over time than physical modalities, like face, for example (GIOT; ROSENBERGER; DORIZZI, 2012c), emphasizing the importance of further studies for behavioural biometric modalities in a data stream context.

3.4.1 Keystroke dynamics

Keystroke dynamics is a behavioural biometric modality that recognizes individuals based on how they type on a keyboard. This modality has some advantages over commonly adopted alternatives, like fingerprint or iris recognition systems (HOSSEINZADEH; KRISHNAN, 2008; PEACOCK; KE; WILKERSON, 2004). First, keystroke dynamics does not require an additional sensor, since a common keyboard, usually present in all personal computers, is enough to acquire keystroke data. Second, keystroke dynamics recognition can be performed in background, while the user works on other applications in the computer. These advantages may contribute to a high acceptability of this modality.

As a behavioural modality, keystroke dynamics has a higher tendency to be subject to changes over time than physical modalities like fingerprint (GIOT; ROSENBERGER; DORIZZI, 2012c). In fact, the rhythm to type a password evolves over time and can be different in a short timespan. This is due to several reasons, which cannot always be controlled, such as: increased practice, changes on the environment, etc. These modifications increase the intraclass variability which, consequently, contributes to template ageing.

Four datasets were selected for the experiments with keystroke dynamics of this thesis, as shown in Table 2. As discussed in (GIOT; EL-ABED; ROSENBERGER, 2012), CMU and GREYC were regarded as the only datasets available that could provide statistical significant results for keystroke dynamics. A more recent dataset, GREYC-Web, released in that publication, was also used here. To the best of the author's knowledge, the datasets mentioned here are the only ones publicly available that have enough data for a study of keystroke dynamics model adaptation over time. A description of each dataset is presented next:

- **CMU** (KILLOURHY; MAXION, 2010): 51 users typed the password “.tie5Roanl” plus the *Enter* key 400 times in 8 sessions. Considering all users, a total of 20,400 biometric samples are available in this dataset.
- **GREYC** (GIOT; EL-ABED; ROSENBERGER, 2009): 100 users typed the expression “greyc laboratory” in at least 5 sessions, during a period of two months. Considering these 100 users, there are near 7,000 biometric samples available in GREYC dataset. However, this dataset has an average of 67.5 samples per user (only one user has more than 100 samples). Hence, although it has been used for studying adaptive biometric systems, there may not be enough samples to capture large typing changes in this dataset.
- **GREYC-Web** (GIOT; EL-ABED; ROSENBERGER, 2012): 118 users contributed to this dataset, some of them for more than 1 year. The updated version, available in the

authors website, was used here. This dataset has data for logins and passwords. Hence, this dataset can be divided into two datasets:

- **Logins:** for the transcription of the login (“laboratoire greyc”), the 35 users with at least 100 valid samples were considered. This results in more than 7,000 biometric samples.
- **Passwords:** for the transcription of the passwords (“SÉSAME”), the 29 users with at least 100 valid samples were considered. This results in more than 5,500 samples. To the best of the author’s knowledge, the research carried out during this thesis is the first work to use the passwords part of this dataset (PISANI *et al.*, 2016).

Table 2 – Summary of keystroke dynamics datasets: (GIOT; EL-ABED; ROSENBERGER, 2009), (KILLOURHY; MAXION, 2010) and (GIOT; EL-ABED; ROSENBERGER, 2012). The data on this table refers to the datasets after pre-processing.

	GREYC	CMU	GREYC-Web (L)	GREYC-Web (P)
# users	100	51	35	29
# samples (avg per user)	67.49	400	213.26	194.97
Expression	“greyc laboratory”	“tie5Roanl” + <i>Enter</i> key	“laboratoire greyc”	“SÉSAME”
# characters	16	11	17	6
Feature vector length	15	10	16	5

Apart from the *key* itself, the keyboard also provides the instants in which each *key* is pressed and released. From these data, a number of features can be extracted. This thesis used the feature *flight time* type 1 (TEH; TEOH; YUE, 2013), which is one of the most used features in previous keystroke dynamics studies (PISANI; LORENA, 2013). This feature is shown in Figure 12, which is the time difference between the instants when a *key* is released and the next *key* is pressed. Note that this feature can be a negative value, as shown between keys 2 and 3 in Figure 12. This can happen if the user presses the next key before releasing the previous one.

Figure 12 – Features extracted from the keystroke data.

After extracting the features, *rank transformation* was applied to them in order to improve the recognition performance of *Self-Detector*, as presented in (PISANI, 2012; PISANI; LORENA, 2015). This procedure basically associates a rank to each attribute of the feature vector and uses the ranked vector as the feature vector (the final vector is rescaled to $[0;1]$ range, dividing all values by the number of features). However, the *M2005* classification algorithm does not use rank transformation, since preliminary experiments have shown that raw flight

times result in higher accuracy than rank based data for this algorithm. These classification algorithms are described in Section 3.5.

3.4.2 Accelerometer-based gait biometrics

As the name suggests, accelerometer-based gait biometrics attempts to recognize users by accelerometer data. The term *accelerometer biometrics* was used in a recent competition to refer to the recognition of users by accelerometer data in *Kaggle* (KAGGLE, 2013). In the literature, this behavioural biometric modality is also referred to as cell phone-based biometrics (KWAPISZ; WEISS; MOORE, 2010). Accelerometer biometrics, when used to recognize users by their walking pattern, is closely related to *gait biometrics* (ZHANG; HU; WANG, 2011). In gait biometrics, data can be obtained from different sources, including visual data and sensors. The accelerometer is one type of sensor to obtain this data. This thesis focuses on gait biometrics using data obtained from accelerometer in mobile devices, like smartphones. Hence, this thesis adopts the terms *accelerometer-based gait biometrics* / *accelerometer biometrics*.

One of the first experiments to investigate the use of cell-phone accelerometers to recognize users was presented in (SPRAGER; ZAZULA, 2009), which considered users walking at three different speeds. Afterwards, (KWAPISZ; WEISS; MOORE, 2010) also studied how to perform user recognition for other activities too, like jogging, ascending and descending stairs. According to their results, it is also possible to successfully recognize users by smartphone accelerometer data from these activities. Later, in (NICKEL; WIRTL; BUSCH, 2012), three classification algorithms were evaluated to perform this task: *Hidden Markov Model* (HMM), *Support Vector Machines* (SVMs) and *k-Nearest Neighbour* (k-NN). According to their experimental results, k-NN obtained the best recognition performance, indicating that instance-based algorithms can obtain good performance in accelerometer biometrics. Another work also proposed a new distance metric, named *Cross Dynamic Time Warping Metric* (DERAWI; BOURS, 2013), which obtained the best predictive performance in their experiments.

Since accelerometer-based gait biometrics is a behavioural modality, it is expected to be subject to template ageing. The work of (MATOVSKI *et al.*, 2010) claims to be the first to consider the issue of *time* in gait biometrics, although they used visual data instead of accelerometer in their experiments. According to the reported results, users can still be recognized after nine months using a non-adaptive model. However, the authors did not study the application of adaptation strategies in their work.

Three datasets were selected for the accelerometer-based gait biometrics experiments, as shown in Table 3. The first dataset was collected by Jordan Frank at McGill University using the *Human Sense open-source Android data collection platform* (FRANK; MANNOR; PRECUP, 2010). This dataset contains raw accelerometer data from users walking for 15min in two different days. Four *HTC Nexus One* smartphones were used to acquire the data. In addition, two other datasets from WISDM (Wireless Sensor Data Mining) were used in this study. These datasets are *Activity Prediction* (KWAPISZ; WEISS; MOORE, 2011) and *Ac-*

titracker (LOCKHART *et al.*, 2011). The data for the first dataset were acquired in a controlled environment, while the second dataset has “real world data”, as stated in their web-site. Their newest versions of these datasets were employed in this thesis, as available at <http://www.cis.fordham.edu/wisdm/dataset.php>. These datasets contain data from users doing six different activities. Nonetheless, only the activity “walking” was considered, which is the focus in this thesis. This is the activity with the largest amount of data in the WISDM datasets.

Table 3 – Summary of accelerometer-based gait biometrics datasets: (FRANK; MANNOR; PRECUP, 2010), (KWAPISZ; WEISS; MOORE, 2011) and (LOCKHART *et al.*, 2011). The data on this table refers to the datasets after pre-processing.

	McGill	WISDM 1.1 (Activity Prediction)	WISDM 2.0 (Actitracker)
# users	20	33	131
# samples (avg per user)	1245.65	180.55	213.34
Activity	“walking”	“walking”	“walking”
Feature vector length	15	15	15

These datasets contain the acceleration forces measured for the three axes: x , y and z (one series of data for each axis). These series can be processed in different ways. A common strategy is to divide these series into frames of a predefined length. Afterwards, features are extracted from these frames. This thesis considers a frame equivalent to 2s of data, as in (PREECE *et al.*, 2009). Data were sampled at 20Hz in datasets in the WISDM dataset, therefore, every frame corresponds to 40 measures for each of the three axes. In the McGill dataset, the frequency was not reported, however, by looking at the data, it seems to be 25Hz (the majority of time differences between consecutive captures is 40ms). Consequently, every frame corresponds to 50 measures in the McGill dataset. To divide the frames, this work used only the order of the measures in the series per user, initially ignoring timestamps.

After a careful look at the datasets, particularly in the timestamps, some problems were found in the series. A possible reason is that these data come from an accelerometer device in a smartphone, which may not be as stable as a dedicated accelerometer device. Smartphones have several applications running at the same time, which might affect the process of logging accelerometer data (e.g. a heavy application may suddenly be launched and impact the device performance). In view of this fact, before processing the frame, those frames with problems on the timestamps were removed. These problems are described next:

- There is a negative (or zero) time difference between two consecutive measures. The smartphone may have been restarted in the middle of the frame;
- The frame has more than 10 times the expected length of data, although it has the target amount of measures (e.g. 40 in WISDM). This is equivalent to a frame being higher than 20s. The logging application may have been stopped and started again during the

smartphone operation, so the frame includes the end of one session and the beginning of another;

- The frame has less than 1s of data (or exactly 1s). An error during the logging may have occurred.

After removing these frames with problems, users with less than 100 valid frames were removed from the dataset, since they would also contain less than 100 biometric samples. The remaining frames are then interpolated using *cubic spline* (R implementation) and re-sampled at the correct frequency. Even the remaining frames presented problems like variable time differences between consecutive captures, justifying the decision to interpolate the data.

The features are extracted from the re-sampled frames. Different alternatives to extract the feature vector for accelerometer data are compared in (PREECE *et al.*, 2009). According to their results, the best performance was obtained by using features from the frequency domain, like using magnitudes resulted from a *Fast Fourier Transform* (FFT), an implementation of the *Discrete Fourier Transform* (DFT). Hence, this work applied FFT over each of the frames and used the magnitude of the first five components (R implementation) of each axis to generate the feature vector, as in (PREECE *et al.*, 2009). As a result, the feature vector is composed by 15 features (five for each axis).

3.5 Baseline biometric systems

The baseline algorithms were chosen according to what has been applied to each biometric modality in studies dealing with adaptive biometric systems. There are two types of baselines: the non-adaptive classification algorithms and the adaptation strategies. They are described in the next sections.

Several adaptation strategies previously applied to keystroke dynamics are used here as baselines: *Double Parallel* (GIOT; ROSENBERGER; DORIZZI, 2012b), *Growing* (KANG; HWANG; CHO, 2007; PISANI; LORENA; CARVALHO, 2015b), *Sliding* (KANG; HWANG; CHO, 2007; PISANI; LORENA; CARVALHO, 2015b) and *Adapted thresholds* (MHENNI *et al.*, 2016). *Double Parallel* and *Adapted thresholds* have been applied to keystroke dynamics only, so the experiments of this thesis also considers these two adaptation strategies just for keystroke dynamics. *Growing* and *Sliding*, on the other hand, were used for both keystroke dynamics and accelerometer-based gait biometrics.

For keystroke dynamics, the non-adaptive classification algorithm *M2005* (MAGALHÃES; REVETT; SANTOS, 2005) has been employed to adaptive biometric systems using the adaptation strategy *Double Parallel* (DB) (GIOT; ROSENBERGER; DORIZZI, 2012b). This strategy was extended later by the research of this thesis to update the *M2005* model incrementally. As a result, the issue of the endless increase of memory consumption was solved. This involved switching the usage of the median by the mean in *M2005* too. The extended adaptation strategy was named *Improved Double Parallel* (IDB).

Another classification algorithm employed in a similar context is *H2007* (HOCQUET; RAMEL; CARDOT, 2007), used with the *Adapted thresholds* strategy (MHENNI *et al.*, 2016). However, according to preliminary experiments under the *user cross-validation for biometric data streams* methodology, *Adapted thresholds* works better with *M2005* than with *H2007* in most cases. Hence, the results using *H2007* are not reported in this thesis. In addition, it was observed that this adaptation strategy may benefit from a less stringent threshold. Just for the case of this adaptation strategy, the thesis reports only the results which obtains the best balanced accuracy for this adaptation strategy, either the original threshold or the less stringent one. Note that *M2005* was designed for keystroke dynamics (e.g. it increases the score if two sequential flight times are matched), so it was not applied to accelerometer-based gait biometrics.

To the best of the author's knowledge, the research carried out for this thesis is the first to investigate the use of adaptive algorithms for accelerometer-based gait biometrics. Thus, there is no specific adaptive baseline in the literature to guide the choice of the classification algorithms. Since all the classification algorithms adopted in this research are one-class, another one-class algorithm was chosen: the one-class support vector machine (*OCSVM*) (SCHÖLKOPF *et al.*, 1999). This algorithm was chosen because SVM had already been applied to accelerometer-based gait biometrics (NICKEL; WIRTL; BUSCH, 2012), although not in the context of adaptation.

Finally, the immune-based *Self-Detector* (STIBOR; TIMMIS, 2005) is also used as a baseline non-adaptive classification algorithm. It had not been used for adaptive biometric systems, but the research for this thesis proposed several adaptation strategies for this algorithm. In addition, it has obtained good recognition performance for keystroke dynamics in previous work of the author without the use of adaptation (PISANI, 2012; PISANI; LORENA, 2015).

To sum up, the baseline biometric systems employed in this thesis are shown in Table 4. The next sections describe the non-adaptive classification algorithms adopted as baseline for the experiments. The adaptation strategies were described in Section 2.2.

Table 4 – Summary of baseline biometric systems. Adaptive biometric systems are indicated by the adaptation strategy between parenthesis like, for example, *Self-Detector (Sliding)*. Conversely, non-adaptive biometric systems do not use an adaptation strategy, hence, they do not contain a parenthesis in their names like, for example, *Self-Detector*.

Biometric System	Type	Modality	Classification algorithm	Adaptation strategy
<i>Self-Detector</i>	Non-adaptive	Both	Self-Detector	-
<i>Self-Detector (Growing)</i>	Adaptive	Both	Self-Detector	Growing/Self-update
<i>Self-Detector (Sliding)</i>	Adaptive	Both	Self-Detector	Sliding
<i>M2005</i>	Non-adaptive	Keystroke dynamics	M2005	-
<i>M2005 (DB)</i>	Adaptive	Keystroke dynamics	M2005	Double Parallel
<i>M2005 (IDB)</i>	Adaptive	Keystroke dynamics	M2005	I. Double Parallel
<i>M2005 (Adapted Thresholds - Growing)</i>	Adaptive	Keystroke dynamics	M2005	Adapted Thresholds with Growing
<i>M2005 (Adapted Thresholds - Sliding)</i>	Adaptive	Keystroke dynamics	M2005	Adapted Thresholds with Sliding
<i>OCSVM</i>	Non-adaptive	Accelerometer-based gait biometrics	OCSVM	-
<i>OCSVM (Growing)</i>	Adaptive	Accelerometer-based gait biometrics	OCSVM	Growing/Self-update
<i>OCSVM (Sliding)</i>	Adaptive	Accelerometer-based gait biometrics	OCSVM	Sliding

3.5.1 Self-Detector

Self-Detector is an immune-based algorithm (STIBOR; TIMMIS, 2005) that, by working like instance-based algorithms, makes it easier to adapt its model over time (MCEWAN; HART, 2009; MENA-TORRES; AGUILAR-RUIZ, 2014). The standard version shown in Algorithm 7 stores all genuine enrollment samples as detectors and assigns a radius to each of them. Afterwards, in the test/recognition, when a query is presented to the classification algorithm, it will be tested against all detectors. If at least one detector matches the query, it is classified as genuine and, otherwise, as impostor. In the adopted implementation, matching occurs if the distance between the detector and the query sample is smaller than the detector radius (also known as self-radius). This study defines the radius using the methodology described in Section 3.2.2. Distances are computed using the cosine distance as in (PISANI; LORENA, 2015), defined in Equation 3.4.

$$\text{cosine_dist}(\vec{x}, \vec{y}) = 1 - \frac{\sum_{i=1}^d x_i y_i}{\sqrt{\sum_{i=1}^d x_i^2 \sum_{i=1}^d y_i^2}} \quad (3.4)$$

Algorithm 7: Self-Detector.

Input: $\text{ref}_j.T, \mathbf{q}, \theta_j^{\text{verify}} = \{\text{selfRadius}\}$

Output: label^p

```

1 verified  $\leftarrow$  FALSE
2 foreach  $d_j$  in  $\text{ref}_j.T$  do
3   if  $\text{dist}(d_j, \mathbf{q}) < \text{selfRadius}$  then
4     verified  $\leftarrow$  TRUE
5     break
6   end
7 end
8 if verified then
9   return genuine
10 else
11   return impostor
12 end

```

The standard *Self-Detector* does not output a score. However, since a score is needed to use the score normalization procedures discussed in the Chapter 6, the correlation between the closest detector and the query (i.e. the maximum correlation value) is considered to be its score.

3.5.2 M2005

The work from (MAGALHÃES; REVETT; SANTOS, 2005) proposed a classification algorithm for keystroke dynamics, named here as *M2005*. This algorithm extracts a set of statis-

tics from the enrollment samples (mean, median and standard deviation) for each dimension of the feature vector and use them to perform classification.

When a query sample $\mathbf{q}$ is received by the algorithm, it checks whether each dimension i of the feature vector meets the following condition:

$$L(i) \leq \mathbf{q}(i) \leq U(i)$$

Where:

$$L(i) = \min(\text{mean}_i; \text{median}_i) \times (0.95 - \text{std}_i/\text{mean}_i)$$

$$U(i) = \max(\text{mean}_i; \text{median}_i) \times (1.05 + \text{std}_i/\text{mean}_i)$$

In this condition, $q(i)$ is the value of the dimension i in the given query $\mathbf{q}$ and $\text{mean}_i \equiv \text{mean}(\{e(i) \in \mathcal{E}_j(i)\})$, $\text{median}_i \equiv \text{median}(\{e(i) \in \mathcal{E}_j(i)\})$ and $\text{std}_i \equiv \text{std}(\{e(i) \in \mathcal{E}_j(i)\})$ are the mean, median and standard deviation, respectively, of the dimension i from the enrollment samples $\mathcal{E}_j$. For each dimension i of the query that satisfies this condition, the algorithm computes a *sum* according to the Algorithm 8. After checking all the dimensions of the query, the algorithm computes the final score according to the Equation 3.5, in which *max_sum* is defined as $1.0 + 1.5 \times (\text{dimension_count} - 1.0)$.

$$\text{Score} = \text{sum} / \text{max_sum} \quad (3.5)$$

Based on this score, classification is performed. If the obtained value is higher than a given *decisionThreshold*, the query is classified as genuine and, otherwise, as impostor. Note that *M2005* is not used for accelerometer-based gait biometrics, as it was designed for keystroke dynamics. For instance, *M2005* increases the score for consecutive feature/key matches, which is a reasonable procedure for keystroke dynamics, but may not be for FFT magnitude terms (as seen in accelerometer-based gait biometrics).

3.5.3 OCSVM

Support vector machines (VAPNIK, 1998) receive examples from all classes (genuine and impostor) and define a hyper-plane that separates the two classes with maximal margin. SVMs use kernels for this task in order to map the original data to a higher dimensional space (KRUENGKRAI; JARUSKULCHAI, 2003).

Later, SVMs were extended to support one-class problems, in which only examples from one class are used during the training. The one-class SVM (*OCSVM*) (SCHÖLKOPF *et al.*, 1999) defines a hyper-plane that stays as far as possible from the origin and keeps the trained class data on the opposite side of the origin. Another approach for one-class in SVMs is the *Support Vector Data Description* (SVDD) (TAX; DUIN, 2001), which defines the smallest hyper-sphere that can contain all data from the trained class.

Algorithm 8: M2005.

Input: $ref_j.T, \mathbf{q}, \theta_j^{verify} = \{decisionThreshold\}$
Output: $label^p$

```

1   $sum = 0$ 
2   $previousDimMatched \leftarrow FALSE$ 
3  foreach  $i$  in  $\{1, dimension\_count\}$  do
4      if  $ref_j.T.L(i) \leq \mathbf{q}(i) \leq ref_j.T.U(i)$  then
5          if  $previousDimMatched$  then
6               $sum = sum + 1.5$ 
7          else
8               $sum = sum + 1.0$ 
9          end
10          $previousDimMatched \leftarrow TRUE$ 
11     else
12          $previousDimMatched \leftarrow FALSE$ 
13     end
14 end
15  $max\_sum = 1.0 + 1.5 \times (dimension\_count - 1.0)$ 
16  $Score = sum / max\_sum$ 
17 if  $Score > decisionThreshold$  then
18     return genuine
19 else
20     return impostor
21 end

```

This thesis used the *OCSVM*, which was also applied to keystroke dynamics in (YU; CHO, 2003). The obtained classifier outputs the value +1 for the region in which most training examples are located and -1 for the remaining space. The parameter ν defines the fraction of *outliers* in the training examples. In the experiments, the implementation from the LIBSVM library was used (CHANG; LIN, 2011).

3.6 Statistical test

Applying a statistical test to check the significance of the obtained results is an important step in experimental research. In Machine Learning, for instance, it is very common to apply null hypothesis significance tests, like *Friedman* or *Wilcoxon signed rank* tests (DEMŠAR, 2006). However, the use of null hypothesis significance tests in Machine Learning to check whether the performance a given algorithm is significantly better than that of another has been criticized by (CORANI *et al.*, 2016; BENA VOLI *et al.*, 2016; DRUMMOND, 2006; DEMŠAR, 2008). It has been argued that these tests have a number of drawbacks, like obtaining statistical significance in those tests do not necessarily imply in practical significance. One of the stronger arguments is that null hypothesis significance tests simply do not answer the following question: which is the probability of the null and of the alternative hypothesis, given the observed data? Instead, those tests compute the probability of obtaining the observed or a larger

difference between the algorithms if the null hypothesis were true. Hence, this is different than answering the probability that the performance of one algorithm is better than that of another, given the observed experimental results. Moreover, the mean rank based tests (e.g. *Friedman*, *Nemenyi*) have also been criticized (BENAVOLI; CORANI; MANGILI, 2016) since the output of these tests depend on the pool of algorithms. For example, the performance difference between algorithms A and B can be considered significant depending on the performance of the remaining algorithms assessed in the pool.

The authors in (CORANI *et al.*, 2016; BENAVOLI *et al.*, 2016) proposed to use a Bayesian modeling in the statistical test to overcome these problems. The proposed test, named Bayesian Hierarchical test, is used in all experiments of this thesis to check the experimental results. The code provided by the authors is used here too: <<https://github.com/BayesianTestsML/tutorial/tree/master/hierarchical>>. In particular, the latest version of the code (commit 2c1ea90 from 16/January/2017). Based on this code, the random seeds were fixed, as described in (PISANI; CARVALHO, 2016). All the probabilities reported in this thesis are rounded. As a consequence of this rounding process, sometimes the reported values do not satisfy the following condition: $p(left) + p(rope) + p(right) = 100\%$. However, it does not affect any of the conclusions drawn throughout the thesis.

The Bayesian test reports, given the observed empirical results, three probabilities: named $p(left)$, $p(rope)$ and $p(right)$. In this thesis, $p(left)$ is the probability that the baseline is better than the proposal, $p(rope)$ is the probability that the baseline and the proposal are equivalent, and, $p(right)$ is the probability that the proposal is better than the baseline. These are the posterior probabilities on a next unseen dataset. The statistical assessment is done using these probabilities.

This statistical test was designed for experiments that were performed using cross-validation, which means that the training data overlaps among folds. In the *user cross-validation for biometric data streams* methodology adopted in this thesis, the users are cross-validated and, consequently both training and test data overlaps among user groupings. It is different than the scenario considered in (CORANI *et al.*, 2016; BENAVOLI *et al.*, 2016). Even though, the probabilities obtained by the statistical test throughout the thesis provide a good summary of the experimental results and they are reasonable considering the performance obtained by the biometric systems in the experiments.

3.7 Chapter remarks

A number of aspects regarding the evaluation of adaptive biometric systems were discussed in this chapter. Evaluation methodologies that have been adopted in previous work are not entirely suitable for biometrics in a data stream context due to several issues, such as considering session division to perform adaptation (information that may not be available in a practical scenario), use of true labels for adaptation, presence of impostor attacks from both registered

and unregistered users, and employed evaluation metrics (e.g. EER).

An evaluation methodology for biometrics in a data stream context proposed by the research of this thesis was introduced in this chapter. Furthermore, the adopted experimental setup was presented, including biometric modalities and baseline systems. This is the evaluation methodology adopted by all experiments reported in the next chapters.

USAGE CONTROL FOR SELF-DETECTOR

Recent work applying immune algorithms to keystroke dynamics has reported that *Self-Detector* can obtain good recognition performance (PISANI, 2012; PISANI; LORENA, 2015). However, the algorithms were evaluated in a scenario that did not consider the adaptation of the biometric reference. Even though, the results obtained there suggested that the typing rhythm changes over time. This raises the following question: is it possible to turn *Self-Detector* into an adaptive algorithm?

The *Self-Detector* algorithm (STIBOR; TIMMIS, 2005) keeps a set of detectors that are used to match the query samples. This immune algorithm, from the positive selection class of immune algorithms, can be considered to belong to the instance-based category, which makes it easier to adapt its model over time (MCEWAN; HART, 2009; MENA-TORRES; AGUILAR-RUIZ, 2014). Instance-based algorithms store training examples in memory and perform classification of new examples using nearby examples in the memory (MITCHELL, 1997). The detector set can be understood as this memory for *Self-Detector*. One possibility to adapt the detector set would be to control its usage for matching, in order to keep only the most used detectors, leading to the following hypothesis:

- **H1:** Positive selection can become an adaptive class of immune algorithms by controlling the usage of the detectors, enabling its use for biometrics in a data stream context.

The research carried out for this thesis proposed some alternatives in this line, they are named *Usage Control*. Four versions of *Usage Control* were proposed, as discussed in the next sections. They were also investigated in several papers: (PISANI; LORENA; CARVALHO, 2013), (PISANI; LORENA; CARVALHO, 2015b), (PISANI; LORENA; CARVALHO, 2014), (PISANI; LORENA; CARVALHO, 2017) and (PISANI; LORENA; CARVALHO, 2015a). The next sections are organized as follows: Section 4.1 presents an overview of the adaptive positive selection and four adaptation strategies for *Self-Detector*; Section 4.2 shows and discusses the obtained experimental results; and, Section 4.3 presents the main remarks of the chapter.

4.1 Positive selection and adaptation strategies

As previously stated, this chapter deals with adaptation strategies for *Self-Detector* (STIBOR; TIMMIS, 2005). This classification algorithm is from the positive selection class of artificial immune systems. Artificial immune systems are defined as computational systems inspired by the natural immune system and applied to solve various problems (CASTRO; TIMMIS, 2002). These systems have been employed in several applications related to pattern recognition, such as anomaly detection and optimization.

In the context of biometrics, the standard version of *Self-Detector* shown in Algorithm 7 receives enrollment samples $\mathcal{E}_j$ from the genuine user j and stores them as detectors (a radius is assigned to each detector). Afterwards, in the test/recognition, when a query is presented to the classification algorithm, it will be tested against all detectors. If at least one detector matches the query, it is classified as genuine (self) and, otherwise, as impostor (non-self). In the implementation used here, matching occurs if the distance between the detector and the query sample is smaller than the detector radius (also known as self-radius). The current study defines the radius using the methodology described in Section 3.2.2. Distances are computed using the cosine distance as in (PISANI; LORENA, 2015), defined in Equation 3.4. Due to its characteristic of storing enrollment samples in memory, *Self-Detector* can be considered to belong to the instance-based category. This makes it easier to adapt its model over time (MCEWAN; HART, 2009; MENA-TORRES; AGUILAR-RUIZ, 2014).

Figure 13 – *Self-Detector* and its adaptive model. Detectors may change when a query is classified as genuine. The figure was adapted from (PISANI; LORENA; CARVALHO, 2015b).

The standard non-adaptive *Self-Detector* is summarized in Figure 13 (a). In order to make this algorithm adaptive, the model presented in Figure 13 (b) was adopted. The main modification was the incorporation of a mechanism able to decide whether or not to change the detector set when a new query is matched as genuine. It is important to highlight that adaptation only occurs when a query is classified as genuine. The *Sliding* and *Growing* adaptation

strategies applied to *Self-Detector* are shown in Algorithm 9. The current implementation for *Self-Detector* is based on the ideas of (KANG; HWANG; CHO, 2007), as described in (PISANI; LORENA; CARVALHO, 2015b).

Since *Self-Detector* is an instance-based classification algorithm, the user model and the gallery are the same, so $ref_j.T = ref_j.GL$ (the user model is the set of detectors in *Self-Detector*). Therefore, adaptation can be done directly in $ref_j.T$. In Algorithm 9, *Sliding* removes the oldest detector from the set of detectors $ref_j.T$ and adds the query $\mathbf{q}$ as a new detector. This adaptation strategy keeps the amount of detectors constant over time. *Growing*, instead, does not remove the oldest detector. As a consequence, the set of detectors keeps growing over time.

Algorithm 9: *Self-Detector (Sliding and Growing)*. The algorithm shows how *Sliding* works in *Self-Detector*. *Growing* works in the same way, but without the lines 10 and 11, which remove the oldest detector.

Input: $ref_{j(t)}$, $\mathbf{q}$, $\theta_j^{verify} = \{selfRadius\}$, $\theta_j^{adapt} = \{\}$

Output: $(ref_{j(t+1)}, label^p)$

```

1  verified = FALSE
2  foreach  $d_j$  in  $ref_{j(t)}.T$  do
3      if  $dist(d_j, \mathbf{q}) < selfRadius$  then
4          verified = TRUE
5          break
6      end
7  end
8   $ref_{j(t+1)} = ref_{j(t)}$ 
9  if verified then
10      $d_r \leftarrow oldestDetector(ref_{j(t+1)}, .T)$ 
11      $ref_{j(t+1)}.T = ref_{j(t+1)}.T - \{d_r\}$ 
12      $ref_{j(t+1)}.T = ref_{j(t+1)}.T \cup \{detector(\mathbf{q})\}$ 
13     return  $(ref_{j(t+1)}, \mathbf{genuine})$ 
14 else
15     return  $(ref_{j(t+1)}, \mathbf{impostor})$ 
16 end

```

Self-Detector (Growing and Sliding) are considered baseline adaptive biometric systems for the experiments carried out in this thesis, as stated in Section 3.5. Next section present the *Usage Control* versions proposed by the research of this thesis.

4.1.1 Usage Control: first version

As its name suggests, *Usage Control* adapts the biometric reference ref_j by controlling the usage of the detectors (PISANI; LORENA; CARVALHO, 2013; PISANI; LORENA; CARVALHO, 2015b). The adaptation strategy is shown in Algorithm 10. For each detector d_i in the set of detectors $ref_j.T$, two new attributes are assigned in *Usage Control*:

- $d_i.usageCount$: increases every time the detector matches a query.
- $d_i.recentUsage$: decreases when another detector matches a query. If a detector matches a query, it returns to a maximum value MAX_RU (in the current experiments, $MAX_RU = 10$, the same value adopted in (PISANI; LORENA; CARVALHO, 2015b)). When the detector is firstly generated, it also assumes the maximum value for this attribute.

When a query matches a detector, the two additional attributes are updated. The first detector to match the query is considered to be the “used” detector, while the others are considered as “unused”. All detectors with $recentUsage \leq 0$ are ordered by $usageCount$. Among them, the detector with lowest $usageCount$ is removed from $ref_j.T$ and a new detector is added to the set using the matched query q . The effect of these additional attributes in *Usage Control* is the removal of detectors with low usage without removing new detectors instantly (as their $usageCount$ is zero when they are created). If there is no detector that $recentUsage \leq 0$, no adaptation occurs and the recognized query is discarded.

The storage of samples in memory and their replacement according to the usage was also discussed in the context of biometrics in a technical report (SCHEIDAT; MAKRUSHIN; VIELHAUER, 2007), although their approach is different from *Usage Control*. In that work, three strategies that explicit deal with the usage of samples are presented: LFU (least frequently used), LRU (least recently used) and *Extended replacement algorithm*. These strategies are inspired by cache management and are described in Section 2.2. However, that work did not execute tests to evaluate the performance of the presented approaches. Later, another work evaluated LFU along with other replacement strategies (FRENI; MARCIALIS; ROLI, 2008).

Although the replacement of examples based on usage of (SCHEIDAT; MAKRUSHIN; VIELHAUER, 2007) may seem similar to *Usage Control*, there are some key differences between them. In all three approaches from (SCHEIDAT; MAKRUSHIN; VIELHAUER, 2007), authentication is performed using the nearest neighbour. A common implementation needs to scan all samples to determine the nearest neighbour, while *Self-Detector (Usage Control)* does not necessarily scan all detectors to classify a query. One detector matching the query is enough for *Self-Detector*. This can result in faster recognition performance for *Usage Control*.

This difference also affects how the usage attributes are updated. As previously described, just the first detector to match the query is considered as “used”. This means that a detector less similar to the query than another one may have its counters updated as “used”, while the other detector does not. As a result, detectors that are more similar can eventually be eliminated, what may be seem as a problem in a first analysis. However, it is important to remember that when the detector set is updated, a new detector is included using the input sample. Consequently, this problem can be mitigated.

During the description of LFU (SCHEIDAT; MAKRUSHIN; VIELHAUER, 2007), it is claimed that this strategy is not totally appropriate to replace samples in memory. If a sample

Algorithm 10: *Usage Control* and *Usage Control R*. The first version checks detectors from the oldest to the newest one, while *Usage Control R* adopts the opposite order (line 3).

Input: $ref_{j(t)}$, $\mathbf{q}$, $\theta_j^{verify} = \{selfRadius\}$, $\theta_j^{adapt} = \{MAX_RU\}$
Output: $(ref_{j(t+1)}, label^p)$

```

1  verified = FALSE
2   $ref_{j(t+1)} = ref_{j(t)}$ 
3  foreach  $d_i$  in  $ref_{j(t+1)}.T$  do
4      if  $dist(d_i, \mathbf{q}) < selfRadius$  then
5           $d_j.usageCount = d_j.usageCount + 1$ 
6           $d_j.recentUsage = MAX\_RU$ 
7           $D_{unused} = ref_{j(t+1)}.T - d_i$ 
8          foreach  $d_u$  in  $D_{unused}$  do
9               $d_u.recentUsage = d_u.recentUsage - 1$ 
10         end
11         verified = TRUE
12         break
13     end
14 end
15 if verified then
16      $R = \{d_i \in ref_{j(t+1)}.T \mid d_i.recentUsage \leq 0\}$ 
17     if  $R \neq \{\}$  then
18          $d_r \leftarrow whichDetectorHasLowestUsageCount(R)$ 
19          $ref_{j(t+1)}.T = ref_{j(t+1)}.T - \{d_r\}$ 
20          $ref_{j(t+1)}.T = ref_{j(t+1)}.T \cup \{detector(\mathbf{q})\}$ 
21     end
22     return  $(ref_{j(t+1)}, \text{genuine})$ 
23 else
24     return  $(ref_{j(t+1)}, \text{impostor})$ 
25 end

```

is used too many times for a period, it will not be easily replaced, even if it does not represent the current behaviour of the user. This is due to its high value of usage frequency. The authors mentioned that this problem could be reduced if the replacement is done periodically, but the authors did not detail how to perform this procedure. Conversely, *Usage Control* mitigates this problem with the inclusion of the *recentUsage*. If a detector has not been used recently, it can be eliminated even with a *usageCount* higher than that of other detectors. Another related problem is that new detectors may be removed right after their inclusion in LFU, since their usage count is very low when inserted. In *Usage Control*, the *recentUsage* attribute also contributes to avoid this problem. Every time a new detector is created in *Usage Control*, it is flagged as being recently used, $recentUsage = MAX_RU$. As a result, new detectors are not instantly removed in *Usage Control*.

Another strategy discussed in the report is LRU. However, implementation details are not presented there. The authors claim that it may involve high computational cost because it needs to store what was used and when. This can be a problem in cache management, which needs very high efficiency, but may not be a critical issue in biometrics. In order to approximate the effect of LRU, the *clock algorithm* is presented. This algorithm keeps a circular list and controls recent usage of the samples by moving a pointer over this list.

Finally, the third strategy presented in that technical report was *Extended replacement algorithm* (SCHEIDAT; MAKRUSHIN; VIELHAUER, 2007). This strategy associates a measure of relevance to each sample in memory, which is then used for choosing which one is replaced. By the presented definition, this measure combines both LFU and LRU. However, the combination of everything into a single measure leads to some limitations. For instance, it inherits a problem from LFU: the difficulty to remove samples used with high frequency of usage even if it does not represent the current user behaviour. Although *Extended replacement algorithm* reduces this problem, it is still present. As discussed earlier, this problem is mitigated by *Usage Control*, which has two attributes: one for frequency of usage and another for recent usage. Thus, even detectors with high frequency of usage (and higher than other detectors) may be eliminated first if $recentUsage \leq 0$.

Apart from the several differences between *Usage Control* and the strategies in (SCHEIDAT; MAKRUSHIN; VIELHAUER, 2007), it is important to highlight that if all detectors in *Usage Control* have been recently used, the detector set is not modified. Hence, if the detector set is still representative, no adaptation is performed. The strategies in (SCHEIDAT; MAKRUSHIN; VIELHAUER, 2007), on the other hand, always perform replacement.

After the first version of *Usage Control* presented in last section, additional versions of this adaptation strategy were proposed by the research of this thesis. They are described in the next sections.

4.1.2 Usage Control R

In the *Usage Control* presented in last section, when a query is presented, detectors are checked from the oldest to the newest one. However, this can increase the likelihood that older detectors are considered as “used”. In order to avoid this issue, a new version of the adaptation strategy, named *Usage Control R* was proposed (PISANI; LORENA; CARVALHO, 2015b). *Usage Control R* checks the detectors from the newest to the oldest one. The rest of the algorithm is the same as the first version of *Usage Control*, as shown in Algorithm 10.

4.1.3 Usage Control S

For the adaptation strategies *Usage Control* and *Sliding/Growing* presented so far, the query being classified as genuine is usually enough to adapt the biometric reference and to add the query sample to the set of detectors $ref_j.T$. As a consequence, a misclassified query may be wrongly included in the detector set, increasing false acceptance rates. In view of this problem,

a more stringent approach is proposed by *Usage Control S* (PISANI; LORENA; CARVALHO, 2014). In *Usage Control S*, adaptation only occurs if more than one detector matches the query. This additional rule assumes that a query matched by two or more detectors has a higher level of confidence that it is a true genuine. Conversely, a query matched by only one detector has low level of confidence and, therefore, it is not added as a new detector.

Another modification in *Usage Control S* is that all detectors that match the query are now considered as “used”. This means that their *usageCount* and *recentUsage* attributes are update accordingly. This modification avoids the removal of detectors that could still represent the current behaviour of the user, even though they are not the first to match the query. *Usage Control S* is described in Algorithm 11.

Algorithm 11: *Usage Control S*. Adaptation only occurs if more than one detector matches the query.

Input: $ref_{j(t)}$, $\mathbf{q}$, $\theta_j^{verify} = \{selfRadius\}$, $\theta_j^{adapt} = \{MAX_RU\}$
Output: $(ref_{j(t+1)}, label^p)$

```

1  verified = FALSE
2   $ref_{j(t+1)} = ref_{j(t)}$ 
3  nMatch = 0
4  foreach  $d_i$  in  $ref_{j(t+1)}.T$  do
5      if  $dist(d_i, \mathbf{q}) < selfRadius$  then
6           $d_j.usageCount = d_j.usageCount + 1$ 
7           $d_j.recentUsage = MAX\_RU$ 
8           $nMatch = nMatch + 1$ 
9          verified = TRUE
10     end
11     else
12          $d_i.recentUsage = d_i.recentUsage - 1$ 
13     end
14 end
15 if verified and (nMatch > 1) then
16      $R = \{d_i \in ref_{j(t+1)}.T \mid d_i.recentUsage \leq 0\}$ 
17     if  $R \neq \{\}$  then
18          $d_r \leftarrow whichDetectorHasLowestUsageCount(R)$ 
19          $ref_{j(t+1)}.T = ref_{j(t+1)}.T - \{d_r\}$ 
20          $ref_{j(t+1)}.T = ref_{j(t+1)}.T \cup \{detector(\mathbf{q})\}$ 
21     end
22     return  $(ref_{j(t+1)}, \text{genuine})$ 
23 else
24     return  $(ref_{j(t+1)}, \text{impostor})$ 
25 end

```

4.1.4 Usage Control 2

The previous versions of *Usage Control* do not perform any adaptation when there is no detector that $recentUsage \leq 0$. Although this strategy can prevent the inclusion of impostor samples in the biometric reference, the algorithm may lose key information for the adaptation. A new version of *Usage Control*, named *Usage Control 2* (PISANI; LORENA; CARVALHO, 2015a), deals with this issue by always including a matched query as a new detector, regardless of the $recentUsage$ values of all detectors.

Nevertheless, when new detectors are included and there is no detector that $recentUsage \leq 0$, this modification can result in an endless increase in the set of detectors. This is related to the behaviour of the *Growing* strategy. To avoid such a problem, whenever a new query is recognized as genuine, instead of just removing a single detector (the one with the lowest $usageCount$), *Usage Control 2* removes all detectors that $recentUsage \leq 0$. Therefore, *Usage Control 2* strategy does not use the $usageCount$ attribute. It also checks detectors from the newest to the oldest one, the same way *Usage Control R* does. *Usage Control 2* is described in Algorithm 12.

Algorithm 12: *Usage Control 2*. Note that the set of detectors has variable size as the set R may: be empty, contain just one element or more than one element.

Input: $ref_{j(t)}$, $\mathbf{q}$, $\theta_j^{verify} = \{selfRadius\}$, $\theta_j^{adapt} = \{MAX_RU\}$
Output: $(ref_{j(t+1)}, label^p)$

```

1  verified = FALSE
2   $ref_{j(t+1)} = ref_{j(t)}$ 
3  foreach  $d_i$  in  $ref_{j(t+1)}.T$  do
4      if  $dist(d_i, \mathbf{q}) < selfRadius$  then
5           $d_j.recentUsage = MAX\_RU$ 
6           $D_{unused} = ref_{j(t+1)}.T - d_i$ 
7          foreach  $d_u$  in  $D_{unused}$  do
8               $d_u.recentUsage = d_u.recentUsage - 1$ 
9          end
10         verified = TRUE
11         break
12     end
13 end
14 if verified then
15      $R = \{d_i \in ref_{j(t+1)}.T \mid d_i.recentUsage \leq 0\}$ 
16      $ref_{j(t+1)}.T = ref_{j(t+1)}.T - R$ 
17      $ref_{j(t+1)}.T = ref_{j(t+1)}.T \cup \{detector(\mathbf{q})\}$ 
18     return  $(ref_{j(t+1)}, \mathbf{genuine})$ 
19 else
20     return  $(ref_{j(t+1)}, \mathbf{impostor})$ 
21 end

```

As a result of these changes, the set of detectors can increase (when there is no detector that $recentUsage \leq 0$) or decrease (when there is more than one detector that $recentUsage \leq 0$). Consequently, the number of detectors is not constant in *Usage Control 2*, as in the first versions of *Usage Control*. This variation is illustrated in Figure 14, which compares the number of detectors over time of *Usage Control* and *Usage Control 2* for one user. According to this graph, the number of detectors increases for a short time for *Usage Control 2*, then it sharply decreases and keeps varying within a lower range.

Figure 14 – *Usage Control 2* - number of detectors over time (example).

4.2 Experimental results

This section presents the results obtained from the experiments using *Usage Control*. First, the global results are reported and discussed. Then, the performance over time is studied.

4.2.1 Global results

The global results for all datasets are shown in Table 5 (best results for each group are highlighted in bold and standard deviation among runs is shown between parenthesis). Based on the reported balanced accuracy, almost all adaptive biometric systems attained higher performance than non-adaptive systems. This is mainly caused by the reduction in FNMR brought by the adaptation of the biometric reference over time. It suggests that the behaviour of the users change over time and that the adaptation of the reference has a key impact on the recognition performance. In fact, a fundamental aspect of an adaptive biometric system is to reduce the FNMR compared to a non-adaptive system, since the biometric reference should be adapted to the data from the genuine user. Ideally, the reduction in FNMR obtained by adaptive biometric systems should not result in an increase of FMR.

The greatest benefits of using adaptation were observed in the datasets that contain the largest average number of samples per user, on both biometric modalities: CMU and McGill. These datasets resulted in longer biometric data streams, meaning that there was more room for changes in the user behaviour in these datasets, making the difference between non-adaptive and

Table 5 – Results for *Usage Control* (all datasets). The best results for each group are highlighted in bold (standard deviation among runs is shown between parenthesis). Adaptive biometric systems are indicated by the adaptation strategy between parenthesis like, for example, *Self-Detector (Sliding)*. Conversely, non-adaptive biometric systems do not use an adaptation strategy, hence, they do not contain a parenthesis in their names like, for example, *Self-Detector*.

Biometric system	GREYC			CMU		
	FMR	FNMR	Acc (balanc.)	FMR	FNMR	Acc (balanc.)
<i>Self-Detector</i>	0.090 (0.010)	0.165 (0.005)	0.872 (0.006)	0.287 (0.023)	0.410 (0.016)	0.651 (0.009)
<i>Self-Detector (Sliding)</i>	0.092 (0.011)	0.129 (0.004)	0.890 (0.006)	0.291 (0.031)	0.211 (0.013)	0.749 (0.016)
<i>Self-Detector (Growing)</i>	0.105 (0.011)	0.119 (0.005)	0.888 (0.006)	0.562 (0.039)	0.118 (0.009)	0.660 (0.018)
<i>M2005</i>	0.221 (0.019)	0.130 (0.003)	0.824 (0.009)	0.273 (0.028)	0.451 (0.019)	0.638 (0.013)
<i>M2005 (DB)</i>	0.220 (0.019)	0.086 (0.003)	0.847 (0.009)	0.129 (0.014)	0.373 (0.014)	0.749 (0.010)
<i>M2005 (IDB)</i>	0.210 (0.018)	0.092 (0.004)	0.849 (0.008)	0.122 (0.011)	0.306 (0.008)	0.786 (0.006)
<i>M2005 (Adapted Thresholds - Growing)</i>	0.239 (0.019)	0.092 (0.003)	0.835 (0.010)	0.462 (0.019)	0.113 (0.007)	0.712 (0.009)
<i>M2005 (Adapted Thresholds - Sliding)</i>	0.184 (0.017)	0.115 (0.004)	0.851 (0.008)	0.160 (0.011)	0.295 (0.010)	0.773 (0.007)
<i>Self-Detector (Usage Control)</i>	0.091 (0.010)	0.140 (0.005)	0.884 (0.006)	0.351 (0.033)	0.211 (0.013)	0.719 (0.016)
<i>Self-Detector (Usage Control 2)</i>	0.069 (0.009)	0.168 (0.006)	0.882 (0.006)	0.143 (0.012)	0.323 (0.014)	0.767 (0.009)
<i>Self-Detector (Usage Control R)</i>	0.092 (0.010)	0.140 (0.005)	0.884 (0.006)	0.311 (0.030)	0.220 (0.013)	0.735 (0.015)
<i>Self-Detector (Usage Control S)</i>	0.089 (0.010)	0.149 (0.005)	0.881 (0.006)	0.213 (0.014)	0.275 (0.012)	0.756 (0.008)
Biometric system	GREYC-Web (Logins)			GREYC-Web (Passwords)		
	FMR	FNMR	Acc (balanc.)	FMR	FNMR	Acc (balanc.)
<i>Self-Detector</i>	0.066 (0.008)	0.141 (0.005)	0.896 (0.005)	0.388 (0.014)	0.180 (0.002)	0.716 (0.007)
<i>Self-Detector (Sliding)</i>	0.074 (0.011)	0.085 (0.004)	0.920 (0.007)	0.330 (0.021)	0.205 (0.008)	0.733 (0.012)
<i>Self-Detector (Growing)</i>	0.124 (0.015)	0.060 (0.003)	0.908 (0.008)	0.468 (0.017)	0.123 (0.002)	0.704 (0.008)
<i>M2005</i>	0.096 (0.013)	0.245 (0.016)	0.829 (0.008)	0.329 (0.035)	0.251 (0.015)	0.710 (0.024)
<i>M2005 (DB)</i>	0.083 (0.012)	0.179 (0.012)	0.869 (0.008)	0.251 (0.026)	0.240 (0.014)	0.754 (0.018)
<i>M2005 (IDB)</i>	0.095 (0.015)	0.131 (0.011)	0.887 (0.008)	0.247 (0.023)	0.190 (0.006)	0.781 (0.013)
<i>M2005 (Adapted Thresholds - Growing)</i>	0.132 (0.016)	0.303 (0.023)	0.782 (0.015)	0.273 (0.030)	0.348 (0.023)	0.690 (0.022)
<i>M2005 (Adapted Thresholds - Sliding)</i>	0.074 (0.011)	0.394 (0.020)	0.766 (0.010)	0.146 (0.022)	0.465 (0.022)	0.694 (0.018)
<i>Self-Detector (Usage Control)</i>	0.078 (0.011)	0.084 (0.003)	0.919 (0.006)	0.383 (0.021)	0.171 (0.006)	0.723 (0.011)
<i>Self-Detector (Usage Control 2)</i>	0.035 (0.007)	0.148 (0.010)	0.908 (0.007)	0.186 (0.013)	0.396 (0.012)	0.709 (0.010)
<i>Self-Detector (Usage Control R)</i>	0.069 (0.009)	0.086 (0.004)	0.922 (0.006)	0.360 (0.022)	0.188 (0.006)	0.726 (0.012)
<i>Self-Detector (Usage Control S)</i>	0.053 (0.007)	0.123 (0.005)	0.912 (0.005)	0.255 (0.017)	0.296 (0.009)	0.725 (0.009)
Biometric system	McGill			WISDM 1.1		
	FMR	FNMR	Acc (balanc.)	FMR	FNMR	Acc (balanc.)
<i>Self-Detector</i>	0.104 (0.018)	0.545 (0.023)	0.675 (0.019)	0.183 (0.013)	0.242 (0.013)	0.788 (0.009)
<i>Self-Detector (Sliding)</i>	0.166 (0.018)	0.282 (0.036)	0.776 (0.023)	0.227 (0.028)	0.186 (0.021)	0.794 (0.016)
<i>Self-Detector (Growing)</i>	0.490 (0.036)	0.116 (0.029)	0.697 (0.019)	0.349 (0.024)	0.123 (0.014)	0.764 (0.016)
<i>OCSVM</i>	0.323 (0.031)	0.486 (0.025)	0.596 (0.019)	0.183 (0.017)	0.426 (0.024)	0.695 (0.013)
<i>OCSVM (Growing)</i>	0.323 (0.031)	0.486 (0.025)	0.596 (0.019)	0.183 (0.017)	0.426 (0.024)	0.696 (0.013)
<i>OCSVM (Sliding)</i>	0.037 (0.006)	0.894 (0.006)	0.534 (0.005)	0.062 (0.013)	0.633 (0.014)	0.653 (0.010)
<i>Self-Detector (Usage Control)</i>	0.294 (0.035)	0.230 (0.037)	0.738 (0.027)	0.261 (0.030)	0.163 (0.024)	0.788 (0.017)
<i>Self-Detector (Usage Control 2)</i>	0.077 (0.007)	0.356 (0.038)	0.784 (0.021)	0.123 (0.015)	0.213 (0.021)	0.832 (0.011)
<i>Self-Detector (Usage Control R)</i>	0.233 (0.028)	0.250 (0.036)	0.758 (0.024)	0.227 (0.027)	0.174 (0.024)	0.799 (0.017)
<i>Self-Detector (Usage Control S)</i>	0.115 (0.016)	0.422 (0.035)	0.732 (0.023)	0.152 (0.014)	0.237 (0.015)	0.805 (0.010)
Biometric system	WISDM 2.0					
	FMR	FNMR	Acc (balanc.)			
<i>Self-Detector</i>	0.168 (0.006)	0.220 (0.007)	0.806 (0.005)			
<i>Self-Detector (Sliding)</i>	0.203 (0.015)	0.150 (0.008)	0.824 (0.008)			
<i>Self-Detector (Growing)</i>	0.302 (0.018)	0.116 (0.009)	0.791 (0.008)			
<i>OCSVM</i>	0.090 (0.006)	0.381 (0.007)	0.765 (0.006)			
<i>OCSVM (Growing)</i>	0.090 (0.006)	0.381 (0.007)	0.765 (0.006)			
<i>OCSVM (Sliding)</i>	0.028 (0.003)	0.572 (0.008)	0.700 (0.005)			
<i>Self-Detector (Usage Control)</i>	0.227 (0.016)	0.136 (0.008)	0.819 (0.008)			
<i>Self-Detector (Usage Control 2)</i>	0.123 (0.007)	0.176 (0.009)	0.851 (0.006)			
<i>Self-Detector (Usage Control R)</i>	0.198 (0.014)	0.144 (0.009)	0.829 (0.007)			
<i>Self-Detector (Usage Control S)</i>	0.156 (0.007)	0.185 (0.010)	0.829 (0.006)			

adaptive biometric systems more evident. The best biometric system is different for each dataset (in terms of balanced accuracy), although some tendencies can be observed. *Self-Detector (Sliding)* is among the best systems, indicating that the use of newer samples from the genuine user in the adaptation process is a good strategy

Self-Detector (Growing), however, did not perform well. On some datasets, such as GREYC-Web (Passwords) and both WISDMs, *Self-Detector (Growing)* obtained balanced accuracy worse than the non-adaptive *Self-Detector*. This adaptation strategy only adds new samples and never removes old samples from the detector set. Consequently, old patterns, which may not represent the current behaviour of the user are not removed. More importantly, wrongly classified patterns are also never removed from the detector set in *Growing*. Such strategy resulted in high FMR. In CMU, its FMR was above 50%, which means that more than half of the impostor attempts succeed. *Growing* works the same way as *Self-update* discussed in Section 2.2. For behavioural modalities, in which older unrepresentative patterns must be removed, *Self-update* may not be the best choice. For physical modalities, some experiments mainly adapted the biometric reference to different conditions (POH; KITTLER; RATTANI, 2014), like different acquisition conditions for face recognition. In this case, *Self-update* can be considered a suitable strategy.

Usage Control also performed well, showing that keeping the most used detectors and switching the least used by new samples is also a suitable strategy for adaptation. Among the *Usage Control* versions, *Usage Control R* obtained higher balanced accuracy than the first version. This suggests that the detector scanning order can impact the performance. According to the reported results, starting by the newer detectors is a better strategy. Both *Usage Control S* and 2 attained lower FMR than the other *Usage Control* versions. This illustrates that the more stringent rule to add new detectors of *Usage Control S* worked as expected. However, *Usage Control 2* obtained even lower FMR, although it simply adds any sample classified as genuine to the set of detectors. This can be explained by the reduced set of detectors in *Usage Control 2* shown in Figure 14. In this algorithm, the range for the amount of detectors is a function of the maximum value that *recent usage* assumes when a detector is used. As stated earlier, the value adopted here is 10, the same for all *Usage Control* versions. A reduced set of detectors, also means a reduced set of acceptable patterns and, consequently, the biometric system becomes more stringent to accept a sample as genuine.

The non-adaptive *M2005* obtained lower balanced accuracy than the non-adaptive *Self-Detector* in all datasets. However, when the adaptation strategies were applied, *M2005*-based systems obtained higher performance than the *Self-Detector*-based systems in two datasets: CMU and GREYC-Web (Passwords). This illustrates the key impact of adaptation for *M2005* classification algorithm. The adaptation strategy *Adapted Thresholds* obtained good results in some cases, though it was not consistent among the datasets. However, it cannot be concluded that it is not a good adaptation strategy. These results may be due to its design for offline adaptation. The evaluation methodology adopted in this thesis deals with online adaptation in a continuous biometric data stream and, therefore, perhaps *Adapted Thresholds* is not suitable for this application scenario. Regarding *Double Parallel*, the modified version using median attained higher balanced accuracy in all datasets, apart from the more efficient memory usage.

In the accelerometer-based gait biometrics datasets, *OCSVM*-based systems reached the lowest balanced accuracy. Even the adaptive versions did not succeed to improve the performance. This can be explained by the high FNMR of the non-adaptive *OCSVM*. A high FNMR results in fewer samples used for adaptation since only samples classified as genuine are employed by *Sliding* and *Growing*. Due to this low performance, *OCSVM* is not considered in the discussion of the next chapters in this thesis.

As described in Section 3.6, a Bayesian statistical test (CORANI *et al.*, 2016; BENAVALI *et al.*, 2016) was applied to check the results of the *Usage Control* versions against all baselines. The results of the statistical test are shown in Table 6. The Bayesian statistical test adopted here outputs the probabilities that each *Usage Control* version is worse, equivalent or better than the baselines in terms of balanced accuracy. These probabilities are $p(left)$, $p(rope)$ and $p(right)$, respectively. Hence, the higher the $p(right)$, the better is the performance of the *Usage Control* version compared to that of the baselines.

The results of the statistical test shows that the performance of the proposed *Usage Control* versions are better than that of most baselines, such as all non-adaptive and some adaptive biometric systems. This can be observed in the cases where the $p(right)$ values are above 90%. Nonetheless, some baselines, particularly *Self-Detector (Sliding)* and *M2005 (IDB)* performed better than some *Usage Control* versions. This is shown where the probabilities on the left are higher than the ones in the right. All in all, narrowing down the analysis to the *Self-Detector*-based systems, *Usage Control 2* obtained the best results in the statistical test among the *Usage Control* proposals.

Table 6 – Bayesian statistical test (balanced accuracy): *Usage Control* versions vs baseline systems. For each baseline biometric system, three probabilities are respectively reported: $p(left)$, $p(rope)$ and $p(right)$. The higher the probability on the right, the better is the performance of the *Usage Control* version compared to that of the baselines. The table is divided into three sections. The first section presents the results for *Self-Detector*, that was applied to all datasets. The second section presents the results for *M2005*, which was only employed in the keystroke dynamics datasets. The third section presents the results for *OCSVM*, which was only employed in the accelerometer-based gait biometrics datasets.

	Self-Detector (Usage Control)	Self-Detector (Usage Control 2)	Self-Detector (Usage Control R)	Self-Detector (Usage Control S)
<i>Self-Detector</i>	3% 0% 97%	3% 0% 97%	2% 0% 98%	3% 0% 97%
<i>Self-Detector (Sliding)</i>	79% 19% 2%	22% 12% 65%	12% 86% 2%	47% 41% 12%
<i>Self-Detector (Growing)</i>	1% 0% 99%	3% 0% 97%	2% 0% 98%	3% 0% 97%
<i>M2005</i>	4% 0% 96%	8% 0% 92%	4% 0% 96%	6% 0% 94%
<i>M2005 (DB)</i>	40% 0% 60%	30% 0% 69%	31% 0% 68%	26% 1% 73%
<i>M2005 (IDB)</i>	63% 0% 37%	60% 0% 40%	59% 0% 40%	59% 0% 41%
<i>M2005 (Adapted Thresholds - Growing)</i>	11% 0% 89%	7% 0% 93%	9% 0% 91%	6% 0% 94%
<i>M2005 (Adapted Thresholds - Sliding)</i>	25% 0% 75%	19% 0% 81%	22% 0% 78%	18% 0% 82%
<i>OCSVM</i>	9% 0% 91%	6% 0% 94%	8% 0% 92%	6% 0% 94%
<i>OCSVM (Growing)</i>	9% 0% 91%	7% 0% 93%	9% 0% 91%	6% 0% 94%
<i>OCSVM (Sliding)</i>	5% 0% 95%	4% 0% 96%	5% 0% 95%	3% 0% 97%

4.2.2 Performance over time

The performance over time is shown in figures 15, 16, 17 and 18. These figures report FNMR and FMR over time as described in Section 3.3. The curve is the average performance at the indicated window index, while the shaded area represents a confidence interval based on standard error (it shows how the performance varies among the users). Hence, a large shaded area means that the performance presented a high variation among the users.

First, regarding FNMR, non-adaptive biometric systems tend to increase this metric over time, as expected. At the same time, adaptive biometric systems tend to obtain better FNMR than their non-adaptive counterparts. The performance variability among the users is higher for the accelerometer-based gait biometrics datasets than that of the keystroke dynamics datasets. For the accelerometer datasets, FNMR increased over time even for the adaptive biometric systems. However, the adaptation strategies managed to decrease the rate of the FNMR increase.

In the keystroke dynamics datasets, the non-adaptive systems attained higher values of FNMR over time. This illustrates that adaptation was effective to adapt the biometric reference to the current user data. An exception is on the GREYC-Web (Passwords) dataset, where FNMR increased for while and then decreased. This may be because the graph plots the average among all users in the window. As the number of samples is not the same for all users in this dataset, the decrease can be a result of an average on a reduced set of users (which have low FNMR).

Regarding the *Usage Control* versions, they all performed similarly in terms of FNMR. The main difference was observed for *Usage Control S* and 2, which obtained a slightly higher FNMR than the other *Usage Control* versions. The increase in FNMR can be due to the decrease in FMR caused by the more stringent rules to add new samples to the detector set. Consequently, some genuine samples with low classification confidence may have been discarded.

All *M2005*-based biometric systems increased FNMR in the CMU dataset over time, though the adaptive systems managed to decrease the rate of increase compared to the non-adaptive system. In all keystroke dynamics datasets, between the two adaptive versions for *M2005* (*DB* and *IDB*), the FNMR of *IDB* was lower than that of *DB*. The same tendency seen over time here was also observed in the global results reported by Table 5.

Concerning *OCSVM*, as previously discussed, it presented high FNMR. As a consequence, few samples were used for the adaptation process. This can explain the fact that the adaptation strategies could not improve the performance of *OCSVM*-based systems. *Sliding*, for instance, obtained FNMR even worse than the non-adaptive *OCSVM*.

As observed in figures 15 and 16, overall, adaptive biometric systems managed to decrease FNMR over time. This is the expected result from these systems, since the biometric reference is adapted to data from the genuine users. However, at the same time, these adaptive systems should avoid increasing FMR. As shown in figures 17 and 18, most adaptive biometric systems do manage to keep FMR similar to that of the non-adaptive system.

(a) CMU.

(b) GREYC-Web (Logins).

(c) GREYC-Web (Passwords).

Figure 15 – FNMR over time of *Usage Control* and baselines (keystroke dynamics) - Average from all users.

(a) McGill.

(b) WISDM 1.1.

(c) WISDM 2.0.

Figure 16 – FNMR over time of *Usage Control* and baselines (accelerometer) - Average from all users.

(a) CMU.

(b) GREYC-Web (Logins).

(c) GREYC-Web (Passwords).

Figure 17 – FMR over time of *Usage Control* and baselines (keystroke dynamics) - Average from all users.

(a) McGill.

(b) WISDM 1.1.

(c) WISDM 2.0.

Figure 18 – FMR over time of *Usage Control* and baselines (accelerometer) - Average from all users.

Self-Detector (Growing), however, increased FMR over time in almost all datasets. This effect is clearer in the CMU and McGill datasets. As previously discussed, FMR can increase for this biometric system because *Growing* keeps adding samples to the detector set and does not remove any. In this case, if a sample wrongly classified as genuine is included in the detector set, it will never be removed.

Regarding *Usage Control*, the first version obtained the highest FMR values. Conversely, *Usage Control 2* attained the best result in terms of FMR. This is a good result, since this adaptation strategy reduced FNMR of the non-adaptive system on several datasets too.

Figure 19 – Correlation over time - keystroke dynamics (genuine samples only).

Apart from the FNMR and FMR over time, another graph was plot to check the performance of the *Self-Detector*-based systems in figures 19 and 20. These figures show the maximum correlation between detectors and the genuine sample in the biometric data stream (PISANI; LORENA; CARVALHO, 2013; PISANI; LORENA; CARVALHO, 2015b). This maximum correlation can indicate if the biometric reference was successfully adapted to the genuine samples over time. The plot is drawn per user, so three users for each biometric modality were selected to show different behaviours. The graphs in the top show the correlations for the baselines and the graphs in the bottom show the correlations for the *Usage Control* versions.

The non-adaptive *Self-Detector* can be used to show how the user behaviour change over time. In Figure 19, for instance, User A tends to decrease the correlation to the initial data over time, showing a clear change in the typing rhythm. User B decreases for a while and then keeps the correlation stable. User C is a user that does not seem to change the behaviour over

Figure 20 – Correlation over time - accelerometer biometrics (genuine samples only).

time. For all of these different users, the adaptation strategies maintained the correlation more stable and at a higher value. The highest the correlation, the closest the biometric reference is to the current genuine user data. In Figure 20, which shows three users for accelerometer-based gait biometrics, three different behaviours were also selected. As shown in these graphs, the variation over time in accelerometer datasets is higher than that of the keystroke dynamics datasets. This can explain the higher difficult to adapt the biometric reference in this modality.

4.3 Chapter remarks

Previous studies applying immune algorithms to keystroke dynamics have reported that *Self-Detector* can obtain good recognition performance in a scenario without adaptation (PISANI, 2012; PISANI; LORENA, 2015). Even though, it was suggested that the typing rhythm changes over time. *Self-Detector*, from the positive selection class of immune algorithms, can be considered to belong to the instance-based category, which makes it easier to adapt its model over time (MCEWAN; HART, 2009; MENA-TORRES; AGUILAR-RUIZ, 2014). All these factors motivated the investigation of adaptation strategies for the *Self-Detector* algorithm.

In this chapter, adaptation strategies based on the idea of controlling the usage of detectors were presented, named *Usage Control*. The general concept is to keep those detectors that are more frequently (and recently) used for the recognition process. According to the experimental results, this is a suitable strategy to adapt the biometric reference over time.

Throughout the chapter, several aspects regarding adaptation were discussed, including how FNMR and FMR evolve over time. An adaptive biometric system should mainly decrease FNMR, since it is adapting the biometric reference to the genuine user data. However, at the same time, the biometric system should avoid to increase FMR. The reported results confirms that this is the case for several adaptive biometric systems, including versions of the *Usage Control*.

All adaptive biometric systems studied in this chapter only used samples classified as genuine to perform adaptation. However, all samples could be used in order to improve FMR. This is investigated in Chapter 5.

Moreover, a plot showing the correlation of the detectors in the biometric reference illustrated different behaviour change patterns among users. Even for these different behaviours, adaptation overall results in performance improvement. However, this finding also raises the question of whether different adaptation strategies should be chosen per user. This is further investigated in Chapter 7.

COMBINING GENUINE AND IMPOSTOR MODELS

As presented in Chapter 2, the majority of the adaptation strategies only takes into account the queries classified as genuine to adapt the biometric reference. Thus, they only manage a genuine gallery $ref_j.GL^G$, used to induce the genuine model $ref_j.T^G$. However, such adaptation strategies may discard important information, since the queries classified as impostors are not considered for adaptation. To the best of the author's knowledge, there is no prior work that adapts a genuine and an impostor model for each user. This brings the following question: would the use of an additional impostor model generated from samples classified as impostor be able to enhance the recognition performance? This question leads to the following hypothesis:

- **H2:** The use of a negative/impostor model, in addition to the positive/genuine model, can improve the recognition performance of adaptive biometric systems.

The research carried out for this thesis then proposed a new framework, named *Enhanced Template Update (ETU)*, which considers all queries for the adaptation process. Several adaptation strategies were proposed within this new framework. In order to implement such strategies, apart from the genuine gallery $ref_j.GL^G$ and the genuine model $ref_j.T^G$, there is also an impostor gallery $ref_j.GL^I$ and an impostor model $ref_j.T^I$. The usage of an additional impostor model in the adaptation process has two main motivations:

- *Reduce false match:* As some impostor samples would be available, they may help to avoid the inclusion of impostor samples in the genuine gallery. The proposal of the *Positive Gallery Protection (PGP)* takes advantage of this reasoning to attempt to decrease the wrong inclusion of impostor samples in the genuine gallery, also referred to as positive gallery in the context of *ETU*.
- *Reduce false non-match:* This can be achieved by changing the classification decision. Sometimes the genuine model alone (induced from the genuine samples only) may reject a given genuine query, but with the help of an impostor model (induced from a gallery

of impostor samples), it may be possible to verify whether the query sample is closer to the genuine model than the impostor model. As a result, FNMR can be reduced. Four versions are presented here based on this reasoning: *ETU 0* to *3*.

The proposed *Enhanced Template Update* framework was studied in the context of keystroke dynamics in (PISANI *et al.*, 2016). This chapter, however, also includes the results for the accelerometer-based gait datasets. The next sections are organized as follows: Section 5.1 describes the *Enhanced Template Update*; Section 5.2 shows and discusses the obtained results; Section 5.3 presents the main remarks of the chapter.

5.1 Enhanced Template Update

The *Enhanced Template Update* framework (PISANI *et al.*, 2016) is shown in Figure 21. In the enrollment phase, all genuine enrollment samples are stored in the genuine gallery $ref_j.GL^G$, which is then used to induce the genuine model/classifier. Afterwards, in the test phase, biometric query samples classified as genuine are added to the genuine gallery (the genuine model is also updated). Conversely, queries classified as impostor are added to the impostor gallery $ref_j.GL^I$.

When the $ref_j.GL^I$ reaches a minimum size (MIN_SAMPLES_NEG_ACTIVATION), the negative procedure is enabled. At this point, the impostor classifier/model $ref_j.T^I$ is induced and updated every time the impostor gallery is adapted. The galleries are managed by the *Sliding* adaptation strategy, described in Section 2.2. The negative procedure changes the way the classification decision occurs by combining the results obtained from the genuine and the impostor data, instead of using just the genuine model.

The choice of the negative procedure in the *ETU* framework defines each version: *ETU 0* to *3*. Some of the proposed versions were designed to deal with particular issues of some classification algorithms, however, they can be applied to other classification algorithms as long as they output a similarity score (the single exception is *ETU 2*, which is based on an instance-based algorithm).

It must be observed that *Enhanced Template Update* is different from the work presented in (LEE; CHO, 2007), which only updated an impostor database to retrain novelty detectors for keystroke dynamics. *ETU* makes use of two models/galleries to support classification and model adaptation.

5.1.1 Positive Gallery Protection Check

As shown in the top part of the Figure 21, there is an optional *Positive Gallery Protection (PGP)* check. When this check is activated, it attempts to avoid the inclusion of impostor samples in the genuine gallery. Note that the impostor gallery must reach the minimum size (MIN_SAMPLES_NEG_ACTIVATION) to use *PGP*, since the impostor data supports its decision.

Figure 21 – *Enhanced Template Update* (ETU) Framework. ETU has a genuine and an impostor gallery. The additional impostor gallery and impostor model can be used in different ways, defined by the *negative procedure*. The choice of a *negative procedure* within the framework defines each version of ETU. Four of them are discussed in this thesis: ETU 0 to 3. The figure was adapted from (PISANI *et al.*, 2016).

When a new query is classified as genuine, *PGP* returns whether or not it can be added to the genuine gallery. *PGP* first clusters all samples from impostor gallery using *K-Means++* (ARTHUR; VASSILVITSKII, 2007), implementation from *Apache Commons* (COMMONS..., 2014). *K-Means++* is a *K-Means* (BISHOP, 2006) variant which is less prone to poor random initialization of the centers. The k parameter is tuned using the *Ordered multiple runs of k-means* (OMRk) (NALDI *et al.*, 2011). This method executes the clustering algorithm for several values of k and selects the value using a clustering validity index. In the experiments, k ranged from 3 to $(impostor_gallery_size)/2$. The validity index used here is based on the maximum standard deviation among all obtained clusters. If this maximum standard deviation is smaller than a *target* value, the current k value is returned. The *target* value is the standard deviation

of the genuine gallery, which is considered as a single cluster in this case.

After the clustering step, *PGP* computes the center of all obtained impostor clusters (*impostor centers*). The whole genuine gallery is considered as a single genuine cluster, where a *genuine center* is also obtained. Then, *PGP* looks for the cluster center which is closest to the biometric query sample classified as genuine. If the closest cluster is the genuine cluster, the query sample can be added to the genuine gallery, and, therefore, adaptation is allowed. Otherwise, the genuine gallery is not adapted. Some hypothetical situations are shown in Figure 22 to illustrate the *PGP* check. Query 1 is easy to identify as it is outside of the genuine cluster. However, queries 2 and 3 are inside the genuine cluster, so they are likely to be true genuine samples. As there is an overlapping impostor cluster closer to query 3, this sample is not used to update the genuine gallery.

Figure 22 – *ETU* - Positive Gallery Protection (*PGP*). Each positive circle (+) represents a sample in the genuine gallery and each negative circle (–) represents a sample in the impostor gallery. There are three queries q_i ($i \in [1; 3]$) in the figure to represent three situations. Note that adaptation is only allowed if the closest center is the one from the genuine gallery. This figure was adapted from (PISANI *et al.*, 2016).

5.1.2 *ETU 0*: Simple comparison of scores

This is the simplest *ETU* version, so it is called *ETU 0*. It adopts a simple rule: classify the query as genuine if the genuine score is higher than the impostor score as shown in Algorithm 13.

As discussed later in Section 5.2, *ETU 0* has some drawbacks. The impostor gallery can contain samples from several different users (different impostors). Thus, the standard deviation on this gallery is probably high and, in the case of the classification algorithm *M2005*, it can potentially result in a misleading high impostor score for several queries. As described in Section 3.5, *M2005* uses the standard deviation of the enrollment samples to verify queries. The higher the standard deviation, the higher is the acceptable range of the algorithm. Consequently, *ETU*

0 may result in high FNMR due to misleading high impostor scores. A solution for this issue in the case of *M2005* is presented in Section 5.1.5, which discusses *ETU 3*.

Algorithm 13: *Enhanced Template Update 0 (ETU 0)*. The procedure to output the classification decision is highlighted in grey.

Data: $ref_{j(t)}$, $\mathbf{q}$, $\theta_j^{verify} = \{\}$, $\theta_j^{adapt} = \{\}$
Output: $(ref_{j(t+1)}, label^p)$

```

1  $score^G \leftarrow classificationAlgorithm.getScore(ref_{j(t)}.T^G, \mathbf{q})$ 
2  $score^I \leftarrow classificationAlgorithm.getScore(ref_{j(t)}.T^I, \mathbf{q})$ 
3  $verified = (score^G > score^I)$ 
4  $ref_{j(t+1)} = ref_{j(t)}$ 
5 if  $verified$  then
6    $ref_{j(t+1)}.GL^G = ref_{j(t+1)}.GL^G \cup \{\mathbf{q}\}$ 
7    $ref_{j(t+1)}.T^G \leftarrow classificationAlgorithm.train(ref_{j(t+1)}.GL^G)$ 
8   return  $(ref_{j(t+1)}, \text{genuine})$ 
9 else
10   $ref_{j(t+1)}.GL^I = ref_{j(t+1)}.GL^I \cup \{\mathbf{q}\}$ 
11   $ref_{j(t+1)}.T^I \leftarrow classificationAlgorithm.train(ref_{j(t+1)}.GL^I)$ 
12  return  $(ref_{j(t+1)}, \text{impostor})$ 
13 end
```

5.1.3 ETU 1: Comparison of scores

ETU 1 is an incremental modification over *ETU 0*, which adds a new rule to classify a query as genuine. As shown in Algorithm 14, a query is classified as genuine if the genuine score is higher than the impostor score and the difference between them is higher than twice ($\alpha = 2$) the self-radius (it was designed for the *Self-Detector*). As a result, the classification becomes more rigorous, which may contribute to decrease false match when *Self-Detector* is employed.

In some situations, an impostor query can be very far from both the genuine and the impostor model, although it may be closer to the genuine model. In this case, the query should be rejected, but, instead, it would be accepted in *ETU 0* as the genuine score would be higher than the impostor one (even though both have low similarity values in this hypothetical example). To avoid this problem, *ETU* proposes to check if the difference between the scores is large enough. This method was designed to deal with a possible problem of *ETU 0* when applied to *Self-Detector*.

Algorithm 14: *Enhanced Template Update 1 (ETU 1)*. The procedure to output the classification decision is highlighted in grey.

Input: $ref_{j(t)}$, $\mathbf{q}$, $\theta_j^{verify} = \{selfRadius, \alpha\}$, $\theta_j^{adapt} = \{\}$

Output: $(ref_{j(t+1)}, label^p)$

```

1  $score^G \leftarrow classificationAlgorithm.getScore(ref_{j(t)}.T^G, \mathbf{q})$ 
2  $score^I \leftarrow classificationAlgorithm.getScore(ref_{j(t)}.T^I, \mathbf{q})$ 
3  $verified = (score^G - score^I) > (\alpha \times selfRadius)$ 
4  $ref_{j(t+1)} = ref_{j(t)}$ 
5 if  $verified$  then
6    $ref_{j(t+1)}.GL^G = ref_{j(t+1)}.GL^G \cup \{\mathbf{q}\}$ 
7    $ref_{j(t+1)}.T^G \leftarrow classificationAlgorithm.train(ref_{j(t+1)}.GL^G)$ 
8   return  $(ref_{j(t+1)}, \text{genuine})$ 
9 else
10   $ref_{j(t+1)}.GL^I = ref_{j(t+1)}.GL^I \cup \{\mathbf{q}\}$ 
11   $ref_{j(t+1)}.T^I \leftarrow classificationAlgorithm.train(ref_{j(t+1)}.GL^I)$ 
12  return  $(ref_{j(t+1)}, \text{impostor})$ 
13 end
```

5.1.4 ETU 2: k -NN like

Since there are two galleries of samples (genuine and impostor), they could be merged into a single set in order to apply the k -Nearest Neighbour algorithm (FLACH, 2012). This is implemented as shown in Algorithm 15. If most of the k closest samples are genuine, then the query sample is classified as genuine and, otherwise, as impostor. Note that if k is even, a draw can happen and, in this case, the sample is classified as an impostor. In the experiments, $k = 2$ as in (PISANI *et al.*, 2016). ETU 2 uses the cosine similarity to measure distances, the same way that *Self-Detector* does.

5.1.5 ETU 3: Clustering impostor samples

As mentioned in Section 5.1.2, the impostor gallery may have high standard deviation among its samples as it contains samples from different impostors. This can result in misleading high impostor scores when *M2005* is used. ETU 3 deals with it by applying a clustering algorithm to the impostor gallery, as described in Algorithm 16. For each cluster, an impostor model is obtained using *M2005*, which will be used to obtain a set of impostor scores by comparing each model to the query. The final impostor score is the average of them. Afterwards, the classification is performed by checking if the difference between the genuine and impostor scores is large enough ($\beta = 2$), as shown in line 4 of Algorithm 16. The use of clustering reduces the standard deviation of the sample sets, avoiding the misleading high impostor scores in *M2005*. The *KMeans++* (ARTHUR; VASSILVITSKII, 2007) algorithm was used to cluster the samples. The k parameter was tuned using *OMRk*, as described in Section 5.1.1.

Algorithm 15: *Enhanced Template Update 2 (ETU 2)*. The procedure to output the classification decision is highlighted in grey.

Input: $ref_{j(t)}$, $\mathbf{q}$, $\theta_j^{verify} = \{k\}$, $\theta_j^{adapt} = \{\}$
Output: $(ref_{j(t+1)}, label^p)$

- 1 $C \leftarrow getClosestSamples(ref_{j(t)}.GL^G \cup ref_{j(t)}.GL^I, k)$
- 2 $ref_{j(t+1)} = ref_{j(t)}$
- 3 $verified \leftarrow mostSamplesAreGenuine(C)$
- 4 **if** $verified$ **then**
 - 5 $ref_{j(t+1)}.GL^G = ref_{j(t+1)}.GL^G \cup \{\mathbf{q}\}$
 - 6 $ref_{j(t+1)}.T^G \leftarrow classificationAlgorithm.train(ref_{j(t+1)}.GL^G)$
 - 7 **return** $(ref_{j(t+1)}, \text{genuine})$
- 8 **else**
 - 9 $ref_{j(t+1)}.GL^I = ref_{j(t+1)}.GL^I \cup \{\mathbf{q}\}$
 - 10 $ref_{j(t+1)}.T^I \leftarrow classificationAlgorithm.train(ref_{j(t+1)}.GL^I)$
 - 11 **return** $(ref_{j(t+1)}, \text{impostor})$
- 12 **end**

Algorithm 16: *Enhanced Template Update 3 (ETU 3)*. The procedure to output the classification decision is highlighted in grey.

Input: $ref_{j(t)}$, $\mathbf{q}$, $\theta_j^{verify} = \{minScore, \beta\}$, $\theta_j^{adapt} = \{\}$
Output: $(ref_{j(t+1)}, label^p)$

- 1 $score^G \leftarrow classificationAlgorithm.getScore(ref_{j(t)}.T^G, \mathbf{q})$
- 2 $CL \leftarrow clusterSamples(ref_{j(t)}.GL^I)$
- 3 $score^I = \frac{\sum_{cl_i \in CL} classificationAlgorithm.getScore(classificationAlgorithm.train(cl_i), \mathbf{q})}{|CL|}$
- 4 $verified = (score^G - score^I) > (minScore/\beta)$
- 5 $ref_{j(t+1)} = ref_{j(t)}$
- 6 **if** $verified$ **then**
 - 7 $ref_{j(t+1)}.GL^G = ref_{j(t+1)}.GL^G \cup \{\mathbf{q}\}$
 - 8 $ref_{j(t+1)}.T^G \leftarrow classificationAlgorithm.train(ref_{j(t+1)}.GL^G)$
 - 9 **return** $(ref_{j(t+1)}, \text{genuine})$
- 10 **else**
 - 11 $ref_{j(t+1)}.GL^I = ref_{j(t+1)}.GL^I \cup \{\mathbf{q}\}$
 - 12 $ref_{j(t+1)}.T^I \leftarrow classificationAlgorithm.train(ref_{j(t+1)}.GL^I)$
 - 13 **return** $(ref_{j(t+1)}, \text{impostor})$
- 14 **end**

5.2 Experimental results

This section presents the results obtained from the experiments using *Enhanced Template Update*. First, the global results are reported and discussed. Afterwards, the performance over time is evaluated.

5.2.1 Global results

The global results for all datasets are shown in tables 7 and 8 (best results for each group are highlighted in bold and standard deviation among runs is shown between parenthesis). *PGP* was applied to *ETU 0* to 3 and to *Sliding*. Since the *ETU* framework uses *Sliding* to manage its galleries, the standard *Sliding* adaptation strategy can take advantage of *PGP* too. First, concerning *ETU 0*, it is clear that the obtained balanced accuracy was lower than the baselines in almost all cases. In several of them, the performance was even lower than the non-adaptive biometric system. However, this low performance was due to different factors, depending on the classification algorithm. *M2005 (ETU 0)*, for example, obtained high FNMR, while *Self-Detector (ETU 0)* obtained high FMR. *ETU 0* was the simplest rule, which simply compared whether the genuine or the impostor model obtains the highest score for the query.

In the case of *M2005*, the impostor model returned high scores for too many queries (including the genuine ones) and, consequently, *ETU 0* classified them as impostors. In the case of *M2005*, it can be explained by the high variance among samples in the impostor gallery. This gallery contains samples from several different impostor users, so it is expected that there is a high variance among the samples. As described in Section 3.5.2, the *M2005* classification algorithm increases the acceptable range as the standard deviation among the enrollment samples increases. As a result, *M2005* applied to a gallery with high variance is likely to result in high similarity scores for the queries. In the case of *ETU 0*, this can result in higher impostor scores than genuine scores in several cases. Consequently, it increases the rejection of queries, contributing to the high FNMR observed for *M2005 (ETU 0)*.

To deal with this problem, *ETU 3* was proposed. *ETU 3* first clusters the samples in the impostor gallery and then induces several impostor models, one for each cluster. The impostor score is then the average of all impostor models. By clustering the samples, *ETU 3* obtains a lower variance among the samples within the clusters, which mitigates the problem of obtaining unrealistic high impostor scores, as observed in *ETU 0*. The results in Table 7 illustrate that, in fact, this modification improved the balanced accuracy. In CMU, for instance, *M2005 (ETU 3)* obtained the highest overall performance. *ETU 3* obtained good performance among the *M2005*-based systems, overcoming the performance of *M2005 (DB)* in almost all datasets, except in *GREYC-Web (Logins)*.

However, *M2005 (ETU 3)* attained higher balanced accuracy than *M2005 (IDB)* only in the CMU dataset. This may be due to the use of a modified version of *M2005* by *IDB*, which uses mean instead of median in the computations, as described in Section 3.5. CMU is the dataset

Table 7 – *Enhanced Template Update* results - keystroke dynamics datasets. The best results for each group are highlighted in bold (standard deviation among runs is shown between parenthesis). Adaptive biometric systems are indicated by the adaptation strategy between parenthesis like, for example, *Self-Detector (Sliding)*. Conversely, non-adaptive biometric systems do not use an adaptation strategy, hence, they do not contain a parenthesis in their names like, for example, *Self-Detector*.

Biometric system	GREYC			CMU		
	FMR	FNMR	Acc (balanc.)	FMR	FNMR	Acc (balanc.)
<i>Self-Detector</i>	0.090 (0.010)	0.165 (0.005)	0.872 (0.006)	0.287 (0.023)	0.410 (0.016)	0.651 (0.009)
<i>Self-Detector (Sliding)</i>	0.092 (0.011)	0.129 (0.004)	0.890 (0.006)	0.291 (0.031)	0.211 (0.013)	0.749 (0.016)
<i>Self-Detector (Growing)</i>	0.105 (0.011)	0.119 (0.005)	0.888 (0.006)	0.562 (0.039)	0.118 (0.009)	0.660 (0.018)
<i>M2005</i>	0.221 (0.019)	0.130 (0.003)	0.824 (0.009)	0.273 (0.028)	0.451 (0.019)	0.638 (0.013)
<i>M2005 (DB)</i>	0.220 (0.019)	0.086 (0.003)	0.847 (0.009)	0.129 (0.014)	0.373 (0.014)	0.749 (0.010)
<i>M2005 (IDB)</i>	0.210 (0.018)	0.092 (0.004)	0.849 (0.008)	0.122 (0.011)	0.306 (0.008)	0.786 (0.006)
<i>M2005 (Adapted Thresholds - Growing)</i>	0.239 (0.019)	0.092 (0.003)	0.835 (0.010)	0.462 (0.019)	0.113 (0.007)	0.712 (0.009)
<i>M2005 (Adapted Thresholds - Sliding)</i>	0.184 (0.017)	0.115 (0.004)	0.851 (0.008)	0.160 (0.011)	0.295 (0.010)	0.773 (0.007)
<i>Self-Detector (Usage Control)</i>	0.091 (0.010)	0.140 (0.005)	0.884 (0.006)	0.351 (0.033)	0.211 (0.013)	0.719 (0.016)
<i>Self-Detector (Usage Control 2)</i>	0.069 (0.009)	0.168 (0.006)	0.882 (0.006)	0.143 (0.012)	0.323 (0.014)	0.767 (0.009)
<i>Self-Detector (Usage Control R)</i>	0.092 (0.010)	0.140 (0.005)	0.884 (0.006)	0.311 (0.030)	0.220 (0.013)	0.735 (0.015)
<i>Self-Detector (Usage Control S)</i>	0.089 (0.010)	0.149 (0.005)	0.881 (0.006)	0.213 (0.014)	0.275 (0.012)	0.756 (0.008)
<i>Self-Detector (ETU 0)</i>	0.104 (0.010)	0.121 (0.005)	0.887 (0.006)	0.538 (0.016)	0.102 (0.007)	0.680 (0.009)
<i>M2005 (ETU 0)</i>	0.189 (0.017)	0.115 (0.004)	0.848 (0.008)	0.064 (0.008)	0.623 (0.019)	0.656 (0.009)
<i>Self-Detector (ETU 1)</i>	0.089 (0.011)	0.144 (0.006)	0.884 (0.006)	0.251 (0.015)	0.203 (0.017)	0.773 (0.013)
<i>Self-Detector (ETU 2)</i>	0.096 (0.010)	0.126 (0.005)	0.889 (0.006)	0.285 (0.019)	0.207 (0.013)	0.754 (0.013)
<i>M2005 (ETU 3)</i>	0.192 (0.017)	0.100 (0.004)	0.854 (0.008)	0.244 (0.016)	0.143 (0.006)	0.807 (0.009)
<i>Self-Detector (ETU 0) PGP</i>	0.104 (0.010)	0.121 (0.005)	0.887 (0.006)	0.573 (0.018)	0.088 (0.009)	0.670 (0.009)
<i>M2005 (ETU 0) PGP</i>	0.189 (0.017)	0.116 (0.004)	0.848 (0.008)	0.064 (0.009)	0.669 (0.015)	0.633 (0.008)
<i>Self-Detector (ETU 1) PGP</i>	0.089 (0.011)	0.144 (0.006)	0.884 (0.006)	0.271 (0.012)	0.201 (0.015)	0.764 (0.010)
<i>Self-Detector (ETU 2) PGP</i>	0.096 (0.010)	0.126 (0.005)	0.889 (0.006)	0.268 (0.016)	0.224 (0.014)	0.754 (0.011)
<i>M2005 (ETU 3) PGP</i>	0.192 (0.017)	0.101 (0.004)	0.854 (0.008)	0.175 (0.017)	0.243 (0.011)	0.791 (0.010)
<i>Self-Detector (ETU - Sliding) PGP</i>	0.092 (0.011)	0.129 (0.004)	0.889 (0.006)	0.250 (0.021)	0.251 (0.015)	0.750 (0.011)

Biometric system	GREYC-Web (Logins)			GREYC-Web (Passwords)		
	FMR	FNMR	Acc (balanc.)	FMR	FNMR	Acc (balanc.)
<i>Self-Detector</i>	0.066 (0.008)	0.141 (0.005)	0.896 (0.005)	0.388 (0.014)	0.180 (0.002)	0.716 (0.007)
<i>Self-Detector (Sliding)</i>	0.074 (0.011)	0.085 (0.004)	0.920 (0.007)	0.330 (0.021)	0.205 (0.008)	0.733 (0.012)
<i>Self-Detector (Growing)</i>	0.124 (0.015)	0.060 (0.003)	0.908 (0.008)	0.468 (0.017)	0.123 (0.002)	0.704 (0.008)
<i>M2005</i>	0.096 (0.013)	0.245 (0.016)	0.829 (0.008)	0.329 (0.035)	0.251 (0.015)	0.710 (0.024)
<i>M2005 (DB)</i>	0.083 (0.012)	0.179 (0.012)	0.869 (0.008)	0.251 (0.026)	0.240 (0.014)	0.754 (0.018)
<i>M2005 (IDB)</i>	0.095 (0.015)	0.131 (0.011)	0.887 (0.008)	0.247 (0.023)	0.190 (0.006)	0.781 (0.013)
<i>M2005 (Adapted Thresholds - Growing)</i>	0.132 (0.016)	0.303 (0.023)	0.782 (0.015)	0.273 (0.030)	0.348 (0.023)	0.690 (0.022)
<i>M2005 (Adapted Thresholds - Sliding)</i>	0.074 (0.011)	0.394 (0.020)	0.766 (0.010)	0.146 (0.022)	0.465 (0.022)	0.694 (0.018)
<i>Self-Detector (Usage Control)</i>	0.078 (0.011)	0.084 (0.003)	0.919 (0.006)	0.383 (0.021)	0.171 (0.006)	0.723 (0.011)
<i>Self-Detector (Usage Control 2)</i>	0.035 (0.007)	0.148 (0.010)	0.908 (0.007)	0.186 (0.013)	0.396 (0.012)	0.709 (0.010)
<i>Self-Detector (Usage Control R)</i>	0.069 (0.009)	0.086 (0.004)	0.922 (0.006)	0.360 (0.022)	0.188 (0.006)	0.726 (0.012)
<i>Self-Detector (Usage Control S)</i>	0.053 (0.007)	0.123 (0.005)	0.912 (0.005)	0.255 (0.017)	0.296 (0.009)	0.725 (0.009)
<i>Self-Detector (ETU 0)</i>	0.353 (0.030)	0.034 (0.011)	0.807 (0.017)	0.483 (0.018)	0.115 (0.012)	0.701 (0.010)
<i>M2005 (ETU 0)</i>	0.073 (0.011)	0.427 (0.046)	0.750 (0.021)	0.138 (0.020)	0.583 (0.026)	0.640 (0.013)
<i>Self-Detector (ETU 1)</i>	0.065 (0.013)	0.146 (0.020)	0.894 (0.011)	0.441 (0.019)	0.140 (0.014)	0.710 (0.011)
<i>Self-Detector (ETU 2)</i>	0.103 (0.014)	0.071 (0.009)	0.913 (0.010)	0.364 (0.020)	0.171 (0.009)	0.733 (0.011)
<i>M2005 (ETU 3)</i>	0.163 (0.017)	0.118 (0.018)	0.860 (0.013)	0.294 (0.016)	0.177 (0.021)	0.765 (0.013)
<i>Self-Detector (ETU 0) PGP</i>	0.355 (0.032)	0.042 (0.010)	0.802 (0.020)	0.494 (0.018)	0.118 (0.012)	0.694 (0.010)
<i>M2005 (ETU 0) PGP</i>	0.067 (0.010)	0.505 (0.037)	0.714 (0.018)	0.134 (0.018)	0.614 (0.025)	0.626 (0.011)
<i>Self-Detector (ETU 1) PGP</i>	0.064 (0.013)	0.151 (0.019)	0.893 (0.010)	0.448 (0.019)	0.141 (0.014)	0.705 (0.011)
<i>Self-Detector (ETU 2) PGP</i>	0.099 (0.014)	0.078 (0.011)	0.911 (0.011)	0.331 (0.021)	0.213 (0.011)	0.728 (0.012)
<i>M2005 (ETU 3) PGP</i>	0.120 (0.016)	0.151 (0.018)	0.865 (0.013)	0.247 (0.019)	0.268 (0.017)	0.742 (0.014)
<i>Self-Detector (ETU - Sliding) PGP</i>	0.067 (0.010)	0.106 (0.008)	0.913 (0.008)	0.292 (0.023)	0.254 (0.011)	0.727 (0.012)

with the highest number of samples per user for keystroke dynamics and, therefore, obtaining the highest *balanced accuracy* on this dataset shows that *ETU 3* has good performance over time.

As previously discussed, *ETU 0* obtained low performance for *Self-Detector* too, though due to high FMR instead of the high FNMR observed for *M2005*. In some cases, an impostor query may result in low similarity score for both the genuine and the impostor models, though

Table 8 – *Enhanced Template Update* results - accelerometer datasets. The best results for each group are highlighted in bold (standard deviation among runs is shown between parenthesis).

Biometric system	McGill			WISDM 1.1		
	FMR	FNMR	Acc (balanc.)	FMR	FNMR	Acc (balanc.)
<i>Self-Detector</i>	0.104 (0.018)	0.545 (0.023)	0.675 (0.019)	0.183 (0.013)	0.242 (0.013)	0.788 (0.009)
<i>Self-Detector (Sliding)</i>	0.166 (0.018)	0.282 (0.036)	0.776 (0.023)	0.227 (0.028)	0.186 (0.021)	0.794 (0.016)
<i>Self-Detector (Growing)</i>	0.490 (0.036)	0.116 (0.029)	0.697 (0.019)	0.349 (0.024)	0.123 (0.014)	0.764 (0.016)
<i>Self-Detector (Usage Control)</i>	0.294 (0.035)	0.230 (0.037)	0.738 (0.027)	0.261 (0.030)	0.163 (0.024)	0.788 (0.017)
<i>Self-Detector (Usage Control 2)</i>	0.077 (0.007)	0.356 (0.038)	0.784 (0.021)	0.123 (0.015)	0.213 (0.021)	0.832 (0.011)
<i>Self-Detector (Usage Control R)</i>	0.233 (0.028)	0.250 (0.036)	0.758 (0.024)	0.227 (0.027)	0.174 (0.024)	0.799 (0.017)
<i>Self-Detector (Usage Control S)</i>	0.115 (0.016)	0.422 (0.035)	0.732 (0.023)	0.152 (0.014)	0.237 (0.015)	0.805 (0.010)
<i>Self-Detector (ETU 0)</i>	0.569 (0.023)	0.141 (0.034)	0.645 (0.017)	0.416 (0.022)	0.143 (0.024)	0.720 (0.012)
<i>Self-Detector (ETU 1)</i>	0.523 (0.030)	0.155 (0.039)	0.661 (0.015)	0.389 (0.030)	0.155 (0.027)	0.728 (0.015)
<i>Self-Detector (ETU 2)</i>	0.209 (0.026)	0.315 (0.044)	0.738 (0.033)	0.217 (0.028)	0.191 (0.021)	0.796 (0.017)
<i>Self-Detector (ETU 0) PGP</i>	0.585 (0.031)	0.168 (0.037)	0.623 (0.014)	0.441 (0.025)	0.155 (0.024)	0.702 (0.013)
<i>Self-Detector (ETU 1) PGP</i>	0.539 (0.036)	0.192 (0.040)	0.635 (0.014)	0.407 (0.033)	0.163 (0.025)	0.715 (0.015)
<i>Self-Detector (ETU 2) PGP</i>	0.186 (0.019)	0.361 (0.049)	0.727 (0.031)	0.210 (0.026)	0.195 (0.021)	0.798 (0.016)
<i>Self-Detector (ETU - Sliding) PGP</i>	0.134 (0.013)	0.435 (0.045)	0.715 (0.025)	0.202 (0.025)	0.195 (0.021)	0.801 (0.015)

WISDM 2.0			
Biometric system	FMR	FNMR	Acc (balanc.)
<i>Self-Detector</i>	0.168 (0.006)	0.220 (0.007)	0.806 (0.005)
<i>Self-Detector (Sliding)</i>	0.203 (0.015)	0.150 (0.008)	0.824 (0.008)
<i>Self-Detector (Growing)</i>	0.302 (0.018)	0.116 (0.009)	0.791 (0.008)
<i>Self-Detector (Usage Control)</i>	0.227 (0.016)	0.136 (0.008)	0.819 (0.008)
<i>Self-Detector (Usage Control 2)</i>	0.123 (0.007)	0.176 (0.009)	0.851 (0.006)
<i>Self-Detector (Usage Control R)</i>	0.198 (0.014)	0.144 (0.009)	0.829 (0.007)
<i>Self-Detector (Usage Control S)</i>	0.156 (0.007)	0.185 (0.010)	0.829 (0.006)
<i>Self-Detector (ETU 0)</i>	0.437 (0.009)	0.096 (0.006)	0.734 (0.005)
<i>Self-Detector (ETU 1)</i>	0.422 (0.011)	0.103 (0.007)	0.738 (0.006)
<i>Self-Detector (ETU 2)</i>	0.206 (0.010)	0.154 (0.007)	0.820 (0.006)
<i>Self-Detector (ETU 0) PGP</i>	0.462 (0.011)	0.093 (0.006)	0.723 (0.006)
<i>Self-Detector (ETU 1) PGP</i>	0.443 (0.012)	0.100 (0.007)	0.728 (0.006)
<i>Self-Detector (ETU 2) PGP</i>	0.197 (0.010)	0.167 (0.009)	0.818 (0.006)
<i>Self-Detector (ETU - Sliding) PGP</i>	0.179 (0.011)	0.178 (0.009)	0.822 (0.006)

the genuine score could be higher than the impostor score. *ETU 0* would wrongly classify the impostor query as a genuine in this hypothetical example since the genuine score is higher. Even worse, when this occurs, the impostor query is added to the genuine gallery, contributing to increase FMR over time. In order to avoid this problem, *ETU 1* only classifies a query as genuine if the genuine score is higher than the impostor score and if such difference is large enough. Such modification improved the balanced accuracy in almost all datasets.

Another proposed *ETU* version is *ETU 2*, which is based on a standard k -NN. It merges both the genuine and the impostor galleries into a single set and, then, check whether the k closest samples to the query are either genuine or impostor. This *ETU* version obtained even higher performance than *ETU 1* in almost all datasets.

The Bayesian statistical test (CORANI *et al.*, 2016; BENAVALI *et al.*, 2016) described in Section 3.6 was applied to check the results of the *ETU* versions against all baselines in terms of balanced accuracy. *Usage Control* versions presented in the last chapter are also considered as baselines here. The results of the statistical test are shown in tables 9 and 10. *Self-Detector*-based systems were applied to all datasets, while *M2005*-based systems were only applied to keystroke dynamics datasets since *M2005* was designed for this biometric modality. Due to this fact, the statistical comparison was divided into two tables.

According to the statistical test, the performance of *ETU 0* was clearly worse than that of all baselines for both *Self-Detector* and *M2005*. This is expected considering the results discussed in this section. *M2005 (ETU 3)*, the improved version of *ETU 0* for *M2005*, obtained higher performance than the non-adaptive system and the adaptive *M2005 (DB)*. However, the probability that *M2005 (IDB)* is better than *M2005 (ETU 3)* is above 50%. This is in agreement with the results discussed in this section. A possible explanation for this performance improvement of *M2005 (IDB)* is the use of the mean instead of the median in the *M2005* computations. Regarding *Self-Detector (ETU 1)*, it obtained higher performance than *Self-Detector (ETU 0)*, but the statistical test shows that the baselines have higher probability of performing better than *Self-Detector (ETU 1)*. Conversely, *Self-Detector (ETU 2)* performed better in the tests. *Self-Detector (ETU 2)* is equivalent or better than most baselines. These results show that the *ETU* versions can obtain competitive performance, but they still need improvements to be able to obtain higher performance than all baselines.

Table 9 – Bayesian statistical test (balanced accuracy): *ETU - Self-Detector*. For each baseline biometric system, three probabilities are respectively reported: $p(left)$, $p(rope)$ and $p(right)$. The higher the probability on the right, the better is the performance of the *ETU* system compared to that of the baselines.

	Self-Detector (ETU 0)	Self-Detector (ETU 1)	Self-Detector (ETU 2)
<i>Self-Detector</i>	93% 0% 7%	59% 0% 41%	3% 0% 97%
<i>Self-Detector (Sliding)</i>	100% 0% 0%	95% 0% 5%	9% 89% 2%
<i>Self-Detector (Growing)</i>	95% 0% 5%	64% 0% 36%	2% 0% 98%
<i>Self-Detector (Usage Control)</i>	99% 0% 1%	89% 0% 11%	4% 56% 40%
<i>Self-Detector (Usage Control 2)</i>	99% 0% 1%	95% 0% 5%	83% 3% 14%
<i>Self-Detector (Usage Control R)</i>	99% 0% 1%	94% 0% 6%	15% 74% 10%
<i>Self-Detector (Usage Control S)</i>	99% 0% 1%	95% 0% 5%	3% 93% 4%

Table 10 – Bayesian statistical test (balanced accuracy): *ETU - M2005*. For each baseline biometric system, three probabilities are respectively reported: $p(left)$, $p(rope)$ and $p(right)$. The higher the probability on the right, the better is the performance of *ETU* system compared to that of the baselines.

	M2005 (ETU 0)	M2005 (ETU 3)
<i>M2005</i>	75% 0% 25%	10% 0% 90%
<i>M2005 (DB)</i>	94% 0% 6%	22% 2% 76%
<i>M2005 (IDB)</i>	95% 0% 5%	53% 13% 34%
<i>M2005 (Adapted Thresholds - Growing)</i>	87% 0% 13%	4% 0% 96%
<i>M2005 (Adapted Thresholds - Sliding)</i>	87% 0% 13%	8% 0% 92%

5.2.2 Performance over time

The performance over time is shown in figures 23, 24, 25 and 26. These figures report FNMR and FMR over time as described in Section 3.3. Part of these plots can also be found in (PISANI *et al.*, 2016). The curve in the plots is the average performance at the indicated window index, while the shaded area represents a confidence interval based on standard error (it shows how the performance varies among the users). Hence, a large shaded area means that the performance presented a high variance among the users.

(a) CMU.

(b) GREYC-Web (Logins).

(c) GREYC-Web (Passwords).

Figure 23 – FNMR over time of *ETU* - Average from all users (keystroke dynamics).

(a) McGill.

(b) WISDM 1.1.

(c) WISDM 2.0.

Figure 24 – FNMR over time of *ETU* - Average from all users (accelerometer).

All *ETU* versions have the same behaviour in the very first moments since the negative procedure is disabled. As soon as the impostor gallery reaches a minimum amount of samples, the negative procedure is enabled. For the datasets that contain a higher average number of samples per user, this minimum amount of impostor samples is equal to the number of samples used for enrollment. By adopting the same value, the genuine and impostor galleries are balanced. This is the case of CMU and McGill datasets. However, for the remaining datasets, which contain a lower average number of samples per user, adopting such minimum amount for the impostor gallery would prevent the negative procedure to be enabled in several cases. Since the biometric data streams are shorter, there could not be enough impostor samples to enable the negative procedure over time. In order to avoid this problem, the minimum amount for the shorter datasets is half the amount of enrollment samples.

As a consequence of the different values to enable the negative procedure, this part of *ETU* is enabled later in CMU and McGill. Hence, the similar behaviour in the very beginning lasts longer in these two datasets than in the other ones. It can be easily observed in the CMU dataset. Nevertheless, it may not be easy to identify this similar behaviour in the McGill plots due to the large number of windows (McGill data streams are longer than those in CMU).

The issue of the low FNMR for *Self-Detector (ETU 0)* observed in the last section, can be seen in greater detail in figures 23 and 24, particularly for CMU and McGill. In these datasets, FNMR decreases over time, showing that the low FNMR keeps decreasing. The same applies for *M2005 (ETU 0)*, though the FNMR increases over time for this biometric system. Note that such findings cannot be easily observed for the datasets with shorter data streams because the length of the data streams also changes among users. Consequently, the average FNMR plotted in the first windows are computed over all users, while the last windows are computed over the users which contain the largest number of samples.

With regard to *M2005*, FNMR decreased over time when *ETU 3* was applied. This is clearer in CMU, but can also be observed in the first windows of the other datasets. This is a good result since the adaptation strategies for *M2005* presented in last Chapter (*DB* and *IDB*) tend to increase FNMR over time, though in a lower rate than the non-adaptive *M2005*. In Figure 23, *M2005 (ETU 3)* initially increased the FNMR for the CMU dataset, until the *negative procedure* was enabled, then FNMR decreased over time. Still regarding *ETU 3* in CMU, the plot shows that the deviation of FNMR performance among all users was lower than for the other *M2005*-based algorithms, as illustrated by the smaller shaded area.

Figures 25 and 26 show FMR over time just for the biometric systems which support *ETU PGP*. The objective is to discuss the FMR performance with *PGP* (bottom plots) and without *PGP* (top plots), as the main purpose of *PGP* is to reduce FMR. Firstly, the FMR behaviour is the opposite of the FNMR for the systems using *ETU 0*. This is expected as an increase in FNMR usually results in a decrease in FMR, and vice versa. In CMU, for instance, *Self-Detector (ETU 0)* attained a FMR above 50% most of the time. Still in terms of FMR,

(a) CMU.

(b) GREYC-Web (Logins).

(c) GREYC-Web (Passwords).

Figure 25 – FMR over time of *ETU* - Average from all users (keystroke dynamics).

ETU 2 obtained performance similar to *Sliding* in most cases. In the accelerometer-based gait biometrics datasets, the variance among users is smaller for *ETU 2* and *Sliding* than for *ETU 0* and *1*.

Regarding *PGP*, it presented a lower FMR for *Self-Detector (Sliding)*, *Self-Detector (ETU 2)* and *M2005 (ETU 3)*, as expected. However, it was not the case when *ETU 0* and *ETU 1*

(a) McGill.

(b) WISDM 1.1.

(c) WISDM 2.0.

Figure 26 – FMR over time of *ETU* - Average from all users (accelerometer).

were applied. The high error rates for these two *ETU* versions resulted in galleries containing many wrongly classified samples and, therefore, *PGP* could not work properly. This illustrates that *PGP* does not work well if the recognition performance is low.

A Bayesian statistical test (CORANI *et al.*, 2016; BENAOLI *et al.*, 2016) was applied to check the results of *PGP* as shown in Table 11. Since the focus now is on FMR, the test was applied on this rate instead of the balanced accuracy considered in previous statistical tests in this chapter. Since FMR is an error rate, the lower the better. Thus, now the higher the $p(left)$, the better is the performance of *PGP* against the baseline without *PGP*. The statistical test confirms the findings discussed in this section showing that *PGP* successfully decreased FMR for *ETU 2*, *ETU 3* and *Sliding*, but it failed for *ETU 0* and *1*. It must be noted, however, that reducing FMR also resulted in an increase of FNMR. As a consequence, biometric systems with *PGP* overall did not perform better than the systems without *PGP* in terms of balanced accuracy. However, biometric systems with *PGP* are suitable for application scenarios where decreasing FMR is more critical than reducing FNMR.

Table 11 – Bayesian statistical test (FMR): *ETU - PGP*. For each baseline biometric system, three probabilities are respectively reported: $p(left)$, $p(rope)$ and $p(right)$. Note that in this case the test was applied to an error rate, so the interpretation is different from the accuracy. The higher the probability on the left, the better is the performance of the *ETU* system compared to that of the baselines in terms of FMR.

	PGP
<i>Self-Detector (ETU 0)</i>	1% 9% 90%
<i>Self-Detector (ETU 1)</i>	1% 35% 64%
<i>Self-Detector (ETU 2)</i>	75% 24% 1%
<i>Self-Detector (Sliding)</i>	99% 0% 1%
<i>M2005 (ETU 0)</i>	3% 95% 2%
<i>M2005 (ETU 3)</i>	93% 0% 7%

5.3 Chapter remarks

Most adaptation strategies only consider the queries classified as genuine to adapt the biometric reference. Moreover, to the best of the author’s knowledge, all current adaptive biometric systems only keep a model for the genuine user in the biometric reference. This also means that samples classified as impostor are discarded. However, they can contain important information to support the adaptation process. This brings the following question: would the use of an additional impostor model generated from samples classified as impostor be able to enhance the recognition performance?

In order to answer this question, the research for this thesis proposed a new framework, named *Enhanced Template Update (ETU)*, which manages a genuine and an impostor gallery. Two user models are obtained from the galleries, genuine and impostor, which are employed to support both classification of queries and adaptation of the galleries/models. Four versions were proposed within the *ETU* framework. Each one change the way the query is classified using both galleries. In addition, a method to avoid the inclusion of impostor samples in the genuine gallery was proposed within the same framework, named *Positive Gallery Protection (PGP)*.

According to the obtained results, the *ETU 2* and *3* versions have competitive performance. In the CMU dataset, for example, *ETU 3* obtained the best overall balanced accuracy. Moreover, it managed to reduce FNMR over time for *M2005*, while the previous adaptation strategies for this classification algorithm had a tendency to increase this measure over time. Nevertheless, *ETU* systems did not attain the best performance in several datasets. This is also illustrated by the statistical test, which showed that *Usage Control 2* and *IDB* can be better than *ETU*. These results indicate the need for additional research on the combination of genuine and impostor galleries.

Furthermore, the fact that *ETU* can obtain the best performance on one dataset, but not in another suggests that different adaptation strategies should be chosen per dataset. This adds to the findings of the last chapter which also indicate the need to choose adaptation strategies per dataset and per user. This topic is further investigated in Chapter 7.

ADAPTATION USING SCORE NORMALIZATION

A key aspect for the performance of adaptive biometric systems is the choice of which query samples will be used to adapt the biometric reference. Most adaptation strategies are based on just one genuine gallery and, therefore, the inclusion of impostor samples in this gallery should be avoided. In order to deal with this issue, a number of solutions have been employed, such as using an adaptation threshold higher than the decision threshold (RATTANI *et al.*, 2009; RATTANI; MARCIALIS; ROLI, 2011b), only adapting the biometric reference if the query is recognized by more than one detector (PISANI; LORENA; CARVALHO, 2014), ETU - PGP (PISANI *et al.*, 2016). Basically, these solutions only include queries classified with higher confidence in the adaptation process. However, this may also imply in discarding genuine query samples with low similarity, which could be used to improve the user model.

Another way to improve this situation is by the improvement of the classification decision process to make the output more reliable. A more reliable output from the classification algorithm can result in more genuine samples used for adaptation, while avoiding the wrong usage of impostor samples during the adaptation process of the genuine user model. In biometrics, score normalization has been used as a way of refining the classification decision (POH; MERATI; KITTLER, 2009) and can be an alternative in this direction.

A preliminary study on the use of score normalization for supervised adaptation to cope with different acquisition conditions has been conducted in (POH *et al.*, 2010). Nonetheless, to the best of the author's knowledge, score normalization has not been used for biometrics in a data stream context. Hence, could score normalization be used to increase adaptive biometric systems performance by a better refinement of the final decision? It means that by using score normalization, a more effective threshold could also be chosen, which can in turn increase the amount of correctly accepted genuine samples for adaptation, while reducing the amount of impostor samples wrongly accepted for adaptation. This leads to the following hypothesis:

- **H3:** Score normalization can improve the recognition performance of adaptive biometric systems.

The application of score normalization in adaptive biometric systems is not straightforward. Aspects concerning how to obtain the normalization terms in such a dynamic context are discussed. The primary objective is to investigate the possibility of using score normalization (POH; MERATI; KITTLER, 2009), such as T-Norm, F-Norm and Z-Norm to improve the recognition performance of an adaptive biometric system.

The application of score normalization to biometrics in a data stream context discussed in the next sections, along with the obtained results were submitted to a journal (PISANI *et al.*, 2016). This chapter is organized as follows: Section 6.1 introduces the topic of score normalization; Section 6.2 describes how the score normalization terms are obtained within the *user cross-validation for biometric data streams* methodology; Section 6.3 reports the obtained results and discusses several questions regarding the application of score normalization in adaptive biometric systems; Section 6.4 presents the main remarks of this chapter.

6.1 Score normalization

Score normalization can be applied to biometric systems which perform classification on the basis of the score between the query and the biometric reference. This is formalized by the function *classificationAlgorithm.getSimilarityScore* (Section 2.1), which receives the user model $ref_j.T$ and the query $\mathbf{q}$ as input. The output is the similarity score. If the returned score is above a threshold value, the query is classified as genuine, otherwise, as impostor. Score normalization is applied to the output score of the classification algorithm. The normalized score is then compared to the threshold to perform the classification of the query. The normalization aims to reduce any undesirable variation in the output score, and this often increases the class separability between genuine and impostor query samples. This section describes four score normalization procedures (POH; MERATI; KITTLER, 2009): T-Norm, Z-Norm, F-Norm and Adaptive F-Norm.

Test Normalization, also known as T-Norm (AUCKENTHALER; CAREY; LLOYD-THOMAS, 2000), is a widely used score normalization. This normalization uses data from other known users who work as cohort users. T-Norm is applied over the output *score*, which is subject to z-scoring, and returns the normalized score $score^T$, as shown in Equation 6.1 (POH; MERATI; KITTLER, 2009), where μ_j^c is the average *cohort* score and σ_j^c is the standard deviation of the *cohort* scores, for the user j . *Cohort* scores are those obtained by comparing the query to the user models of a different set of users at runtime (known as *cohort* models). Depending on the number of cohort models, the computation of T-Norm can imply in high processing cost. This thesis considers that the set of cohort models can be different for each user.

$$score^T = \frac{score - \mu_j^c}{\sigma_j^c} \quad (6.1)$$

Another related method is Z-Norm (REYNOLDS, 1997), which also applies z-scoring to the score. However, unlike T-Norm, Z-Norm does not need any additional significant computation at runtime. Z-Norm makes use of a development dataset to estimate the normalization parameters offline. For each user $j \in \mathcal{J}$, the development set has a set of additional genuine samples. These additional genuine samples are used to compute the average impostor score $\mu_{I,j}^d$ and the standard deviation $\sigma_{I,j}^d$, for the user j . Z-Norm is defined as shown in Equation 6.2.

$$score^Z = \frac{score - \mu_{I,j}^d}{\sigma_{I,j}^d} \quad (6.2)$$

According to (POH; BENGIO, 2005), both T-Norm and Z-Norm are considered impostor centric because they only use data from impostor users to normalize the scores. Conversely, F-Norm is considered a client-impostor centric score normalization, as it also takes genuine data into account. F-Norm is defined in Equation 6.3, where $\mu_{I,j}^d$ is the average impostor score also used in Z-Norm and $\tilde{\mu}(\gamma, j)$ is the term that considers the genuine data. This term is defined in Equation 6.4, where $\mu_{G,j}^d$ is the average of the genuine scores for the reference user j and μ_G^d is the average of $\mu_{G,j}^d$ for all users. In the same equation, γ adjusts the weight of the expected value of the user-specific mean $\mu_{G,j}^d$ and the user-independent mean μ_G^d .

$$score^F = \frac{score - \mu_{I,j}^d}{\tilde{\mu}(\gamma, j) - \mu_{I,j}^d} \quad (6.3)$$

$$\tilde{\mu}(\gamma, j) = \gamma \mu_{G,j}^d + (1 - \gamma) \times \mu_G^d \quad (6.4)$$

There is also an adaptive version of the F-Norm which uses cohort scores instead of $\mu_{I,j}^d$ (POH; MERATI; KITTLER, 2009). It is named Adaptive F-Norm and work as shown in Equation 6.5. The advantage of using Adaptive F-Norm over the standard F-Norm is that the mean is dependent on each query sample and, therefore, it is adapted to the context of the operational environment.

$$score^{AF} = \frac{score - \mu_j^c}{\tilde{\mu}(\gamma, j) - \mu_j^c} \quad (6.5)$$

6.2 Computation of score normalization in the user cross-validation

This section describes how the score normalization terms are obtained from the *user cross-validation for biometric data streams methodology* adopted in this work. An overview is shown on Figure 27. Note that some score normalization procedures require an offline set of scores, such as Z-Norm and F-Norm. This section describes how to obtain these data within the evaluation methodology.

Figure 27 – Sets to obtain the score normalization terms. Adapted from (PISANI *et al.*, 2016).

First, three sets of scores are obtained: $S_{j,query}^c$, S_j^G and S_j^I . The first set contains the cohort scores for the user j considering a *query*. All registered users $i \neq j$ are considered here as cohort users. $S_{j,query}^c$ contains the scores obtained by comparing the query to all user models other than the claimed model with index j , as described in Equation 6.6.

$$S_{j,query}^c = \{classificationAlgorithm.getSimilarityScore(ref_i.T, \mathbf{q}) \mid i \in \mathcal{J} \wedge i \neq j\} \quad (6.6)$$

The second set contains the genuine scores for user j , which is used to calculate the F-Norm terms, as described in Equation 6.7. This set of scores is obtained only from the enrollment samples of the user j , which are the first samples from the user j in the user cross-validation for biometric data streams evaluation methodology. These scores are obtained by applying the *leave one out method* over the enrollment samples. In this method, the classification algorithm is trained using all enrollment samples from user j except one ($\mathcal{E}_j - e_n$), resulting in the user model $T^{(\mathcal{E}_j - e_n)}$. The remaining enrollment sample (e_n) is used to obtain a genuine

score: *classificationAlgorithm.getSimilarityScore*($T^{(\mathcal{E}_j - e_n)}, e_n$). This process is repeated until all enrollment samples are used once for matching, as described in Equation 6.7.

$$\mathcal{S}_j^G = \{\text{classificationAlgorithm.getSimilarityScore}(T^{(\mathcal{E}_j - e_n)}, e_n) \mid e_n \in \mathcal{E}_j\} \quad \text{Where:} \quad (6.7)$$

$$T^{(\mathcal{E}_j - e_n)} = \text{classificationAlgorithm.train}(\mathcal{E}_j - e_n)$$

The third set contains impostor scores for user j . They are needed to compute F-Norm and Z-Norm. This set is obtained as described in Equation 6.8, using only the enrollment samples from all other users (except those belonging to the reference user j). For such, the model of the reference user $ref_j.T$ is compared to all enrollment samples from the remaining registered users, defined as the set $\mathcal{H}_j^I$.

$$\mathcal{S}_j^I = \{\text{classificationAlgorithm.getSimilarityScore}(ref_j.T, s_n) \mid s_n \in \mathcal{H}_j^I\} \quad (6.8)$$

From these sets, the terms needed for computing T-Norm, Adaptive F-Norm, Z-Norm and F-Norm can be obtained as described in Equation 6.9 to 6.14:

$$\mu_j^c = \mu_{j,query}^c = \text{mean}(\mathcal{S}_{j,query}^c) \quad (6.9)$$

$$\sigma_j^c = \sigma_{j,query}^c = \text{std}(\mathcal{S}_{j,query}^c) \quad (6.10)$$

$$\mu_{G,j}^d = \text{mean}(\mathcal{S}_j^G) \quad (6.11)$$

$$\mu_G^d = \text{mean}(\{\mu_{G,j}^d \mid j \in \mathcal{J}\}) \quad (6.12)$$

$$\mu_{I,j}^d = \text{mean}(\mathcal{S}_j^I) \quad (6.13)$$

$$\sigma_{I,j}^d = \text{std}(\mathcal{S}_j^I) \quad (6.14)$$

6.3 Experimental results

This section presents the results obtained from the experiments using score normalization. The discussion of the results is guided by a sequence of questions. These questions start from checking the hypothesis (whether score normalization can improve the recognition performance of adaptive biometric systems) to questions concerning specific aspects, such as the performance of score normalization versus adaptation and differences among users.

Enhanced Template Update are not employed in the experiments of score normalization because it uses an additional impostor model and the studied score normalization procedures are designed to normalize genuine output scores. The baseline adaptation strategy *Adapted Thresholds* is also not considered in the experiments, since the main contribution of this strategy is to adapt the threshold over time, by making it more stringent in later moments. The usage of score normalization specifically changes the score output range and, therefore, it may affect the way that the threshold should be adapted in the adaptation strategy. Hence, to avoid misleading conclusions, it is not part of the experiments.

6.3.1 Adaptive biometric systems improved by score normalization

The first question asked in this chapter is: could score normalization be used to increase adaptive biometric systems performance by a better refinement of the final decision? To answer this question, the score normalization procedures previously described in this chapter were applied to several adaptive biometric systems. However, there are many combinations of score normalization procedures, biometric systems and datasets to be analyzed. This could make the analysis difficult, so a plot showing the relative performance of the biometric systems with and without score normalization is used in this section. The complete table of results can be found in Appendix B.

The relative performance plots shown in the Figure 28 are based on the balanced accuracy. It was based on the plots of (POH; TISTARELLI, 2012). The relative performance is computed as shown in Equation 6.15, $BAcc_a$ is the performance of the biometric system with score normalization and $BAcc_b$ is the performance of the baseline (without normalization). Each data point used to build the box plots was obtained from the Equation 6.15. One data point was computed for each run of the experiments, which means that there were 150 points per system per dataset ($30 \text{ executions} \times 5 \text{ user groupings}$). Each box plot is also composed by all datasets from the given biometric modality. As a result, each box plot for keystroke dynamics has 600 data points ($150 \times 4 \text{ datasets}$) and each box plot for accelerometer-based gait biometrics has 450 data points ($150 \times 3 \text{ datasets}$). By merging data points from all datasets into a single box plot, there is a reduced amount of plots to be analyzed, making it easier to draw conclusions. In case there is too much variance among the datasets, it would be shown in the box plot by a larger graph.

(a) Keystroke dynamics.

(b) Accelerometer-based gait biometrics.

Figure 28 – Biometric systems using score normalization - relative performance (PISANI *et al.*, 2016). There are 600 data points per keystroke dynamics box plot and 450 data points per accelerometer-based gait biometrics box plot. As these plots are based on balanced accuracy, the higher the values in the horizontal axis, the better.

$$RP_a = \frac{BAcc_a - BAcc_b}{BAcc_b} \quad (6.15)$$

According to the box plots, the performance was improved by the application of score normalization in most cases. Nevertheless, the results change depending on the biometric modality. While T-Norm and Adaptive F-Norm generally increase the recognition performance of biometric systems in the keystroke dynamics datasets, F-Norm and Adaptive F-Norm do so only for some biometric systems in the accelerometer-based gait datasets. Note that, for accelerometer-based gait biometrics, T-Norm, which was the best score normalization for keystroke dynamics, seems to obtain performance lower than the baseline without normalization. Moreover, the performance change obtained by the score normalization also depends on

the biometric system. For example, in the accelerometer datasets, F-Norm showed little improvement for *Self-Detector (Usage Control 2)*, while the performance increase is clear for *Self-Detector (Usage Control)* using the same score normalization procedure.

The results reported in Figure 28 suggest that score normalization can increase the recognition performance of adaptive biometric systems for keystroke dynamics, but it is not clear for accelerometer-based gait biometrics due to the higher variance.

By checking all the results, it can be seen that T-Norm was the best score normalization for keystroke dynamics and Adaptive F-Norm for accelerometer-based gait biometrics in most cases. An additional set of box plots for these two cases is shown in Figure 29. These graphs consider all adaptive biometric systems for each dataset in terms of FMR and FNMR. They are based on the absolute difference as shown in Equation 6.16, where Metric can be either FNMR or FMR. The purpose is to compare which performance metric (FMR or FNMR) was more affected by the use of score normalization in our experiments with adaptive biometric systems. For such, the absolute difference was considered, instead of the relative performance. Each data point in the box plots was obtained using the Equation 6.16, where $Metric_a$ is the value of the metric for the system with score normalization and $Metric_b$ is the value of the same metric for the baseline system without score normalization (one data point was computed for each run). There are 150 data points per system per dataset, similarly to the previous box plots in Figure 28. There are eight adaptive biometric systems for keystroke dynamics, which yields 1200 (8×150) data points for each box plot for this biometric modality. For accelerometer-based gait biometrics, there are six adaptive biometric systems, which results in 900 (6×150) data points for each box plot.

$$PD_a = Metric_a - Metric_b \quad (6.16)$$

By comparing the two plots (FMR and FNMR), it is clear that the adaptive biometric systems in the keystroke dynamics datasets benefited most from the decrease in FNMR (T-Norm boxplots are to the left of the baseline without normalization), though at the cost of an increase in FMR. However, the increase in FMR is not as high as the decrease in FNMR. This suggests that the adaptive biometric system was able to accept a higher amount of genuine samples, that would have been rejected without any score normalization. Nevertheless, the opposite effect was observed for the accelerometer-based gait biometrics datasets. For these datasets, FMR decreased in most datasets, while FNMR increased. However, again, the improvement in FMR is higher than the increase in FNMR, except in McGill. This result suggests that the better refinement of the final decision brought by the score normalization reduces the amount of impostor samples wrongly accepted for adaptation.

Finally, as described in Section 3.6, a Bayesian statistical test (CORANI *et al.*, 2016; BENAOLI *et al.*, 2016) was applied to check the results of the score normalization procedures over the biometric systems in terms of balanced accuracy. The results of the statistical test are

Figure 29 – FMR and FNMR absolute performance difference per dataset. As these plots are based on error rates (FMR and FNMR), the lower the values in the horizontal axis, the better. There are 1200 data points per keystroke dynamics box plot and 900 data points per accelerometer-based gait biometrics box plot. The plots for keystroke dynamics were submitted to (PISANI *et al.*, 2016).

shown in Table 12. The Bayesian statistical test adopted here outputs the probabilities that the performance of the biometric systems using each score normalization is worse, equivalent or better than the performance of the biometric systems without any score normalization.

According to the results reported in Table 12, the best overall score normalization was Adaptive F-Norm. The $p(right)$ values are around 90% for most baseline systems. As shown in Figure 28, this normalization procedure improved the balanced accuracy on both keystroke dynamics and accelerometer-based gait biometrics. However, as illustrated by the same figure, the best score normalization for keystroke dynamics was T-Norm. T-Norm obtained $p(right)$ values around 90% for several cases too. However, for example, *Self-Detector (Sliding)* resulted in $p(right)$ around 50%. The same applies to F-Norm, which also performed well just in some cases. Z-Norm, on the other hand, attained low performance in the experiments (the $p(left)$ values are around 90% in most cases).

Both T-Norm and Adaptive F-Norm adapt their normalization terms over time, since they make use of cohort models. The cohort models here were implemented as the current user models from the registered users other than the claimed user. F-Norm and Z-Norm, conversely,

Table 12 – Bayesian statistical test (balanced accuracy): score normalization. For each baseline biometric system, three probabilities are respectively reported: $p(left)$, $p(rope)$ and $p(right)$. The higher the probability on the right, the better is the performance of the system using the given score normalization compared to that of the baselines, which do not use any score normalization.

	T-Norm	Z-Norm	F-Norm	Adaptive F-Norm
<i>Self-Detector</i>	41% 0% 59%	90% 3% 6%	2% 95% 3%	24% 5% 72%
<i>Self-Detector (Sliding)</i>	50% 0% 50%	97% 0% 3%	6% 15% 79%	10% 0% 90%
<i>Self-Detector (Growing)</i>	10% 0% 90%	95% 0% 5%	10% 0% 90%	20% 0% 80%
<i>Self-Detector (Usage Control)</i>	15% 1% 84%	96% 0% 4%	6% 3% 92%	5% 0% 95%
<i>Self-Detector (Usage Control R)</i>	37% 0% 63%	96% 0% 4%	4% 27% 69%	9% 0% 91%
<i>Self-Detector (Usage Control S)</i>	43% 0% 57%	96% 0% 4%	1% 86% 13%	9% 0% 91%
<i>Self-Detector (Usage Control 2)</i>	69% 0% 31%	95% 0% 5%	10% 81% 9%	12% 1% 87%
<i>M2005</i>	2% 0% 98%	1% 1% 97%	3% 2% 95%	4% 5% 92%
<i>M2005 (DB)</i>	1% 0% 99%	62% 18% 20%	10% 2% 87%	3% 1% 97%
<i>M2005 (IDB)</i>	1% 4% 95%	71% 6% 23%	14% 3% 82%	10% 18% 72%

do not update their normalization terms, which are only obtained at the enrollment phase. As the experiments were conducted considering a data stream context, where the biometric features can change over time, this can be a reason for the superior performance of T-Norm and Adaptive F-Norm over the other normalization procedures.

6.3.2 Impact of score normalization versus adaptation

As observed in the last section, the recognition performance increases when score normalization is applied. However, the improvement obtained by the application of score normalization can be different depending on the biometric system. This brings another important question in the context of adaptive biometric system: can a non-adaptive biometric system with score normalization attain higher recognition performance than an adaptive biometric system without any score normalization? By answering this question, it is possible to know whether adaptation or score normalization results in higher impact on the performance of a non-adaptive biometric system. In order to answer this question, three experimental scenarios were assessed, as follows:

1. Non-adaptive biometric system (without score normalization) vs. adaptive biometric system (without score normalization)
2. Non-adaptive biometric system (with the best score normalization) vs. adaptive biometric system (without score normalization)
3. Non-adaptive biometric system (with the best score normalization) vs. adaptive biometric system (with the best score normalization)

The first scenario is the standard one, which has been tested in the previous chapters of this thesis and, therefore, the result is already known: adaptive biometric systems attained higher performance than non-adaptive biometric systems in most cases. The second and third scenarios, however, have not been assessed so far in the thesis. In the second scenario, score

normalization is applied only to the non-adaptive biometric system. It allows to check whether score normalization or adaptation alone has a higher impact in the recognition performance. This assessment takes into account only the best score normalization per modality here. It was chosen from the last section, where T-Norm performed better for keystroke dynamics and Adaptive F-Norm for accelerometer-based gait biometrics. In the third scenario, both non-adaptive and adaptive biometric systems use the score normalization. This allows to check whether the benefits from score normalization and adaptation can be combined in order to obtain an even better biometric system than using each of them alone.

The three scenarios are shown using the relative performance box plots in Figure 30, where each row of graphs represents from the first to the third scenarios, respectively. Since there are two baseline classification algorithms for keystroke dynamics, the plots for this modality are divided into two sets of graphs: one for *M2005* and another for *Self-Detector*.

Figure 30 – Relative Performance: score normalization vs adaptation (PISANI *et al.*, 2016). There are two different base classification algorithms for keystroke dynamics, so there is one set of graphs for each one: *M2005* and *Self-Detector*.

Among the evaluated scenarios in Figure 30, perhaps the most interesting one is the second scenario, which compares the non-adaptive biometric system using score normalization to the adaptive biometric system without any score normalization. According to the results, the adaptive biometric systems attained higher performance than the non-adaptive ones enhanced by score normalization in most cases. The relative difference between the two systems was reduced, although adaptation still resulted in better performance. *Self-Detector*-based adaptive biometric systems, however, did not perform better than the non-adaptive baselines. It illustrates that the answer for the question asked in the beginning of this section depends on the biometric modality and the classification algorithm. In short, the improvement obtained from adaptation is higher than that of score normalization for most cases, but for *Self-Detector*-based systems, the conclusion is the opposite.

A Bayesian statistical test (CORANI *et al.*, 2016; BENAOLLI *et al.*, 2016) as described in Section 3.6 was applied to check the results of score normalization vs adaptation in terms of balanced accuracy (the second scenario in Figure 30: (d), (e) and (f)). The results of the statistical test are shown in Table 13. For this test, the baselines are the non-adaptive biometric systems using score normalization. The higher the $p(right)$, the better is the performance of the adaptive biometric system compared to the that of the non-adaptive system using score normalization.

The results of the statistical test show that almost all adaptive biometric systems obtain $p(right)$ above 80%, meaning that adaptation resulted in higher impact on the overall biometric system performance. The single exception was *Self-Detector (Growing)*. As discussed previously in this thesis, the Growing adaptation strategy has not obtained good performance in the experiments. A possible reason is the lack of a forgetting mechanism, which can be more critical when dealing with behavioural biometric modalities.

Table 13 – Bayesian statistical test (balanced accuracy): score normalization vs adaptation. Each line is for a baseline non-adaptive biometric system using score normalization, which are compared to the adaptive biometric systems without score normalization (columns). For each baseline, three probabilities are respectively reported: $p(left)$, $p(rope)$ and $p(right)$. The higher the probability on the right, the better is the performance of the adaptive biometric system compared to the non-adaptive baseline.

	Self-Detector (Growing)	Self-Detector (Sliding)	Self-Detector (Usage Control)
Self-Detector - T-Norm	60% 0% 40%	9% 0% 90%	18% 0% 81%
Self-Detector - Z-Norm	2% 4% 94%	1% 0% 99%	1% 0% 99%
Self-Detector - F-Norm	26% 39% 34%	4% 0% 96%	4% 1% 96%
Self-Detector - Adaptive F-Norm	70% 3% 27%	10% 1% 89%	15% 3% 82%
	Self-Detector (Usage Control R)	Self-Detector (Usage Control S)	Self-Detector (Usage Control 2)
Self-Detector - T-Norm	12% 0% 88%	10% 0% 90%	10% 0% 90%
Self-Detector - Z-Norm	1% 0% 99%	1% 0% 99%	2% 0% 98%
Self-Detector - F-Norm	3% 0% 97%	4% 0% 95%	5% 0% 95%
Self-Detector - Adaptive F-Norm	11% 1% 88%	10% 0% 90%	8% 0% 92%
	M2005 (DB)	M2005 (IDB)	
M2005 - T-Norm	21% 1% 78%	15% 0% 85%	
M2005 - Z-Norm	13% 0% 86%	9% 0% 91%	
M2005 - F-Norm	17% 0% 83%	13% 0% 87%	
M2005 - Adaptive F-Norm	17% 0% 83%	12% 0% 88%	

Finally, Figure 30 shows in the third row both non-adaptive and adaptive biometric sys-

tems using score normalization. The results show that the combined use of score normalization and adaptation further improves the recognition performance, rather than applying score normalization or adaptation individually.

6.3.3 The best score normalization can be different among users

In Chapter 4, it was suggested that each user may have a different *change patterns*, i.e., they change the biometric features in diverse ways over time. As a consequence, it raises the question of whether different adaptation strategies should be chosen per user. Moreover, it also raises the question of whether the score normalization procedure should be chosen per user too.

To answer this question, a set of heat maps of the balanced accuracy per user for each score normalization experiment were plot. Two of each are shown in figures 31 and 32, CMU and GREYC-Web (Logins) datasets, respectively, since they are enough for the discussion of this chapter. There is an additional set of plots at Appendix C, which can be used to reinforce the findings of this section. There are two types of graphs. The first one is a heat map of the performance per user for each score normalization. The stronger the green, the better is the performance (balanced accuracy). The second graph just highlights which score normalization performed best per user.

Figure 31 – Score normalization performance per user - CMU. The first column presents the balanced accuracy per user (green is better), while the second column highlights (black) the best score normalization for each user. The plots for M2005 (IDB) were submitted to (PISANI *et al.*, 2016).

Figure 32 – Score normalization performance per user - GREYC-Web (Logins). The first column presents the balanced accuracy per user (green is better), while the second column highlights (black) the best score normalization for each user.

From these graphs, it can be observed that the best score normalization can be different for each user in the same dataset. In some cases, like *Self-Detector*-based systems in the CMU dataset, the same best score normalization is shared among the users. Nonetheless, this is not the case for all datasets and algorithms. In the same CMU dataset, *M2005*-based systems do not share the best score normalization for all users. In this particular case, the best normalization for most users are distributed among T-Norm, F-Norm and Adaptive F-Norm. Sometimes, a user obtains the best performance without any score normalization. Moreover, the heat maps show that the performance difference among the score normalization procedures may be high.

All in all, these findings answer the question asked in the beginning of this section by suggesting that an adaptive biometric system able to choose the score normalization per user can obtain a better overall recognition performance than a system that chooses a single/common normalization for all users.

6.4 Chapter remarks

Most adaptation strategies rely on the output of the classification algorithm to choose which samples are used for adaptation. Some of them also include additional steps to avoid the wrong inclusion of impostor samples in the genuine gallery, by only including queries classified with higher confidence in the adaptation process. However, this may also imply in discarding

genuine query samples with low similarity, which could be used to improve the genuine user model. Another approach to deal with this problem is by the improvement of the classification decision process to make the output more reliable. This can be implemented by the use of score normalization, which has been applied to refine the classification decision in biometric systems.

Score normalization was used in adaptive biometric systems to deal with different acquisition conditions under a supervised setting (POH *et al.*, 2010). However, to the best of the author's knowledge, score normalization has not been used for biometrics in a data stream context. This raised the main question of this chapter: could score normalization be used to increase adaptive biometric systems performance by a better refinement of the final decision? According to the reported results, score normalization can enhance the recognition performance of adaptive biometric systems. Moreover, this chapter discussed additional questions concerning the application of score normalization for biometrics in a data stream context.

Overall, T-Norm and Adaptive F-Norm obtained the best performance among the assessed score normalization procedures. Both of them adapted their normalization terms over time since they used cohort models. The cohort models here were implemented as the current user models from the registered users other than the claimed one. The other score normalization procedures studied here did not adapt the normalization terms, which are only computed once at the enrollment phase. Considering biometrics in a data stream context, where the biometric features can change over time, this can be a reason for the superior performance of T-Norm and Adaptive F-Norm over the other normalization procedures.

Each biometric system benefits in different ways from the score normalization. In view of this result, this chapter studied whether score normalization or adaptation used alone has a higher impact on the recognition performance. According to the experiments, in most cases, the use of adaptation without score normalization results in higher performance than the use of score normalization without adaptation. Furthermore, the combined use of both adaptation and score normalization can result in even higher performance than any of them applied alone.

As observed in the experiments, the best score normalization can be different among biometric modalities, which means that it can differ from one dataset to another. Moreover, in the last section of the chapter, it was also shown that the best score normalization can change among users in the same dataset. This suggests that an adaptive biometric system able to choose the score normalization per user can obtain a better overall recognition performance than a system that applies a single/common normalization to all users.

MODULAR ADAPTIVE BIOMETRIC SYSTEM

A number of adaptation strategies have been proposed for biometric systems, as shown in Section 2.2. In addition, new adaptation strategies were proposed and evaluated by the research of this thesis, as presented in previous chapters. These adaptation strategies are based on different approaches, some of them are:

- Chronological order of the samples (ROLI; MARCIALIS, 2006; KANG; HWANG; CHO, 2007; RATTANI *et al.*, 2009): *Self-update/Growing, Sliding*;
- Usage of detectors (PISANI; LORENA; CARVALHO, 2015b; PISANI; LORENA; CARVALHO, 2014; PISANI; LORENA; CARVALHO, 2015a): *Usage Control* versions;
- Fusion of scores (GIOT; ROSENBERGER; DORIZZI, 2012b): *Double Parallel*;
- Clustering (ULUDAG; ROSS; JAIN, 2004; FRENI; MARCIALIS; ROLI, 2008): DEND, MDIST;
- Graph theory (RATTANI; MARCIALIS; ROLI, 2008; RATTANI; MARCIALIS; ROLI, 2013a): *Graph min-cut*.

As reported in the experiments carried out for this thesis, a given adaptation strategy can present the best recognition performance for one dataset, measured by the average of the performance for all users, but this superiority does not necessarily occurs for another dataset. This is related to the *no free lunch theorem* (LUXBURG; SCHÖLKOPF, 2011). Different *change patterns* among datasets, even for the same biometric modality, can be the reason for this result. A *change pattern* is defined here as the way that the biometric features change over time. Consequently, a biometric system able to choose the most suitable adaptation strategy for each dataset would be, in theory, able to perform better in all datasets.

Furthermore, a few studies in the literature suggest that characteristics, including the *change pattern*, may also be user dependent (PISANI; LORENA; CARVALHO, 2015b; POH

et al., 2015). This thesis illustrates that the *change pattern* can be different among users in Chapter 4. In (PISANI; LORENA; CARVALHO, 2015b), it was reported that the typing rhythm change can differ depending on the user. Moreover, in (POH *et al.*, 2015), the authors showed that the recognition performance changes in different fashions depending on the user (e.g. the performance decreases for some users over time, while it can also increase for others). Another study suggested that different groups of users may need modifications in the adaptation strategy and parameters (RATTANI; MARCIALIS; ROLI, 2009).

As a consequence, a biometric system able to choose the adaptation strategy per user would be able to perform even better. To the best of the author's knowledge, an adaptive biometric system with these capabilities has never been proposed. This motivates the following hypothesis:

- **H4:** An adaptive biometric system can be divided into modules, such that it can choose different module implementations for each user. Therefore, distinct adaptation rules can be assigned to each user.

This thesis hypothesizes that this new adaptive biometric system can be designed following a modular framework, in which each aspect of the adaptive biometric system is divided into modules. By selecting particular implementations of the modules, the adaptation rules of several current adaptive biometric systems could be reproduced. Within this framework, the choice of the adaptation rules per user is equivalent to choosing the implementations of each module per user.

Assuming this modular system is able to generalize all baseline biometric systems, it would have the potential to obtain a recognition performance higher than (or at least equal to) that obtained by any baseline. Since the best adaptation strategy may differ from one user to another, this modular system would also potentially be able to obtain the best recognition performance on all datasets.

This chapter is organized as follows: Section 7.1 presents the proposed modular adaptive biometric system; Section 7.2 describes how the combination of module implementations is selected; Section 7.3 reports the experimental results using the modular system and discusses several questions regarding this new framework; Section 7.4 points out some alternatives for future extension of the modular system; and, Section 7.5 presents the main remarks of the chapter. Contents of this chapter were submitted to a journal for publication.

7.1 ModBioS: modular adaptive biometric system

ModBioS is a modular adaptive biometric system that introduces a modular framework, in which each aspect of the adaptive biometric system is divided into modules. The behaviour of this modular biometric system can be changed by selecting different module implementations. The proposed system can generalize several current adaptive biometric systems from the literature (Section 2.2) as well as all *Usage Control* versions presented in Chapter 4.

In the new modular system, the choice of the module implementations, named here a *combination*, can be made per user. By choosing a different combination per user, distinct adaptation strategies can be adopted for each user. This is a key contribution of this new modular adaptive biometric system. The proposed framework can also choose a different classification algorithm and threshold for each user within the framework.

Formally, *ModBioS* adds a new field, $ref_j.M$, to store the selected modules in the biometric reference of each user. The next sections describe how this new field is used, including the general algorithm adopted by the modular framework. Afterwards, a method to choose the contents of $ref_j.M$ is presented.

7.1.1 Modules

A key contribution of *ModBioS* is its capability to select the modules per user, through the new field $ref_j.M$, which enables the choice of different adaptation strategies. In the proposed implementation, this field is composed by three sub-fields:

- $ref_j.M.U$: global modules for the user;
- $ref_j.M.G^1$: modules specific to model/gallery 1 of the user;
- $ref_j.M.G^2$: modules specific to model/gallery 2 of the user.

ModBioS works with two user models/galleries. This enables the modular framework to implement adaptation rules of single gallery strategies (e.g. Sliding) and dual gallery strategies (e.g. I. Double Parallel). A graphical representation of the biometric reference in *ModBioS* is shown in Figure 33, which illustrates the modules contained in each of the sub-fields in $ref_j.M$. *ModBioS* has two types of modules, user specific and gallery specific, which are described next:

- **User specific** (blue): these are modules that establish global user settings or are shared by the two galleries.
 - **FusionMode**: indicates whether the biometric reference will work on single or dual gallery (biometric score fusion).
 - **ClassificationScore**: defines the classification algorithm used to obtain the score.
 - **FusionDecisionScoreIndex**: for dual gallery, the threshold is defined by this parameter. Since *ModBioS* is able to use different classification algorithms, the range of the threshold values can be different depending on the adopted algorithm. To keep the same range within the framework, *ModBioS* stores the *decisionScoreIndex* in this module and translates it to the proper range at runtime, according to the classification algorithm.
- **Gallery specific** (yellow): as the name suggests, these are the modules that are specific to each gallery.
 - **DecisionScoreIndex**: threshold for the gallery (only used in single gallery mode);

- **AdaptationScoreDiffIndex**: depending on the strategy used to adapt the gallery, a higher threshold value should be adopted to allow adaptation. This parameter defines how much the decision score should be increased for that;
- **DetectorParamManager**: mechanism employed to update the parameters of the detectors in the gallery;
- **AdaptationControl**: mechanism used to return whether the gallery will be adapted using the current query;
- **RemoveDetectorManager**: mechanism employed to remove detectors from the gallery;
- **DetectorScanOrder**: order in which the biometric samples/detectors are scanned.

Figure 33 – User biometric reference in *ModBioS*.

7.1.2 Single gallery mode

As discussed throughout this thesis, most adaptation strategies deals with a single gallery to perform adaptation. When *ModBioS* works in this mode, it attempts to simulate the behaviour of several current single gallery adaptation strategies.

The test and adaptation algorithm for *ModBioS* in single gallery mode (*FusionMode*=0) is shown in Algorithm 17. In this mode, the *ClassificationScore* module outputs a similarity score by matching the query sample to the user model of the current gallery. The *ClassificationScore* module receives the gallery as input, so it induces the model from these samples before computing the similarity score. Afterwards, the minimum score (threshold) required for classifying the sample as genuine is obtained by translating the *DecisionScoreIndex* according

to the *ClassificationScore* module. If the obtained score is higher than the threshold, the query is classified as genuine, otherwise, as impostor.

If the query is classified as genuine, the adaptation procedure is launched. In order to proceed, the adaptation threshold is obtained and the detectors in the gallery are updated by the *DetectorParamManager* (according to the scanning order of *DetectorScanOrder*). The *AdaptationControl* will then return whether the query should be used to update the gallery. If it allows adaptation, the *RemoveDetectorManager* will remove detectors from the gallery (depending on the implementation, the *RemoveDetectorManager* may not remove any detector at all or remove more than one). Finally, the query sample is added as a new detector to the gallery.

Algorithm 17: ModBioS (single gallery).

```

Input:  $ref_{j(t)}$ ,  $\mathbf{q}$ 
Output:  $(ref_{j(t+1)}, label^p)$ 

verificationScore =  $ref_{j(t)}.M.$ 

|   |   |
|---|---|
| U | 1 |
|---|---|

ClassificationScore.getScore( $\mathbf{q}$ ,  $ref_{j(t)}.T^1$ );

decisionScore =  $ref_{j(t)}.M.$ 

|   |   |
|---|---|
| U | 1 |
|---|---|

ClassificationScore.translateDecisionScore( $ref_{j(t)}.M.$ 

|    |   |
|----|---|
| G' | 1 |
|----|---|

DecisionScoreIndex);

verified = (verificationScore > decisionScore);
 $ref_{j(t+1)} = ref_{j(t)}$ 
if (verified) {
    adaptScoreDiff =  $ref_{j(t+1)}.M.$ 

|   |   |
|---|---|
| U | 1 |
|---|---|

ClassificationScore.translateAdaptScoreDiff(
         $ref_{j(t+1)}.M.$ 

|    |   |
|----|---|
| G' | 2 |
|----|---|

AdaptationScoreDiffIndex);
    adaptThreshold = (decisionScore + adaptScoreDiff);
     $ref_{j(t+1)}.T^1 = ref_{j(t+1)}.M.$ 

|    |   |
|----|---|
| G' | 3 |
|----|---|

DetectorParamManager.UpdateDetectorParameters( $\mathbf{q}$ ,  $ref_{j(t+1)}.T^1$ ,
         $ref_{j(t+1)}.M.$ 

|    |   |
|----|---|
| G' | 6 |
|----|---|

DetectorScanOrder);
    if ( $ref_{j(t+1)}.M.$ 

|    |   |
|----|---|
| G' | 4 |
|----|---|

AdaptationControl.CheckAllowAdaptation( $\mathbf{q}$ ,  $ref_{j(t+1)}.T^1$ , verificationScore,
        adaptThreshold)) {
         $ref_{j(t+1)}.T^1 = ref_{j(t+1)}.M.$ 

|    |   |
|----|---|
| G' | 5 |
|----|---|

RemoveDetectorManager.RemoveDetectors( $\mathbf{q}$ ,  $ref_{j(t+1)}.T^1$ );
         $ref_{j(t+1)}.T^1 = ref_{j(t+1)}.T^1 \cup \{detector(\mathbf{q})\}$ 
    }
    return ( $ref_{j(t+1)}$ , genuine)
} else {
    return ( $ref_{j(t+1)}$ , impostor)
}

```

7.1.3 Dual gallery mode (biometric fusion)

Although most adaptation strategies deal with a single gallery, a few adaptation strategies manage two galleries. A key example is *Double Parallel* (GIOT; ROSENBERGER; DORIZZI, 2012b). The current implementation for dual gallery in *ModBioS* employs biometric

fusion of scores ($FusionMode=1$). The main algorithm for test and adaptation in this mode is shown in Algorithm 18.

Algorithm 18: *ModBioS* (dual gallery).

```

Input:  $ref_{j(t)}, \mathbf{q}$ 
Output:  $(ref_{j(t+1)}, label^p)$ 

verificationScore1 =  $ref_{j(t)}.M.$ 

|                     |   |
|---------------------|---|
| U                   | 1 |
| ClassificationScore |   |

 $.getScore(\mathbf{q}, ref_{j(t)}.T^1);$ 

verificationScore2 =  $ref_{j(t)}.M.$ 

|                     |   |
|---------------------|---|
| U                   | 1 |
| ClassificationScore |   |

 $.getScore(\mathbf{q}, ref_{j(t)}.T^2);$ 

fusionVScore = (verificationScore1 + verificationScore2) / 2.0;

fusionDecisionScore =  $ref_{j(t)}.M.$ 

|                     |   |
|---------------------|---|
| U                   | 1 |
| ClassificationScore |   |

 $.translateDecisionScore($ 
   $ref_{j(t)}.M.$ 

|                          |   |
|--------------------------|---|
| U                        | 2 |
| FusionDecisionScoreIndex |   |

 $);$ 

fusionVerified = (fusionVScore > fusionDecisionScore);

 $ref_{j(t+1)} = ref_{j(t)}$ 

if (fusionVerified) {

  adaptScoreDiff1 =  $ref_{j(t+1)}.M.$ 

|                     |   |
|---------------------|---|
| U                   | 1 |
| ClassificationScore |   |

 $.translateAdaptScoreDiff($ 
     $ref_{j(t+1)}.M.$ 

|                          |   |
|--------------------------|---|
| G <sup>1</sup>           | 2 |
| AdaptationScoreDiffIndex |   |

 $);$ 

  adaptThreshold1 = (fusionDecisionScore + adaptScoreDiff1);

   $ref_{j(t+1)}.T^1 = ref_{j(t+1)}.M.$ 

|                      |   |
|----------------------|---|
| G <sup>1</sup>       | 3 |
| DetectorParamManager |   |

 $.UpdateDetectorParameters(\mathbf{q}, ref_{j(t+1)}.T^1,$ 
     $ref_{j(t+1)}.M.$ 

|                   |   |
|-------------------|---|
| G <sup>1</sup>    | 6 |
| DetectorScanOrder |   |

 $);$ 

  if ( $ref_{j(t+1)}.M.$ 

|                   |   |
|-------------------|---|
| G <sup>1</sup>    | 4 |
| AdaptationControl |   |

 $.CheckAllowAdaptation(\mathbf{q}, ref_{j(t+1)}.T^1, fusionVScore,$ 
    adaptThreshold1)) {

     $ref_{j(t+1)}.T^1 = ref_{j(t+1)}.M.$ 

|                       |   |
|-----------------------|---|
| G <sup>1</sup>        | 5 |
| RemoveDetectorManager |   |

 $.RemoveDetectors(\mathbf{q}, ref_{j(t+1)}.T^1);$ 
     $ref_{j(t+1)}.T^1 = ref_{j(t+1)}.T^1 \cup \{detector(\mathbf{q})\}$ 
  }

  adaptScoreDiff2 =  $ref_{j(t+1)}.M.$ 

|                     |   |
|---------------------|---|
| U                   | 1 |
| ClassificationScore |   |

 $.translateAdaptScoreDiff($ 
     $ref_{j(t+1)}.M.$ 

|                          |   |
|--------------------------|---|
| G <sup>2</sup>           | 2 |
| AdaptationScoreDiffIndex |   |

 $);$ 

  adaptThreshold2 = (fusionDecisionScore + adaptScoreDiff2);

   $ref_{j(t+1)}.T^2 = ref_{j(t+1)}.M.$ 

|                      |   |
|----------------------|---|
| G <sup>2</sup>       | 3 |
| DetectorParamManager |   |

 $.UpdateDetectorParameters(\mathbf{q}, ref_{j(t+1)}.T^2,$ 
     $ref_{j(t+1)}.M.$ 

|                   |   |
|-------------------|---|
| G <sup>2</sup>    | 6 |
| DetectorScanOrder |   |

 $);$ 

  if ( $ref_{j(t+1)}.M.$ 

|                   |   |
|-------------------|---|
| G <sup>2</sup>    | 4 |
| AdaptationControl |   |

 $.CheckAllowAdaptation(\mathbf{q}, ref_{j(t+1)}.T^2, fusionVScore,$ 
    adaptThreshold2)) {

     $ref_{j(t+1)}.T^2 = ref_{j(t+1)}.M.$ 

|                       |   |
|-----------------------|---|
| G <sup>2</sup>        | 5 |
| RemoveDetectorManager |   |

 $.RemoveDetectors(\mathbf{q}, ref_{j(t+1)}.T^2);$ 
     $ref_{j(t+1)}.T^2 = ref_{j(t+1)}.T^2 \cup \{detector(\mathbf{q})\}$ 
  }

  return ( $ref_{j(t+1)}, \text{genuine}$ )
} else {

  return ( $ref_{j(t+1)}, \text{impostor}$ )
}

```

In dual gallery mode, two galleries (indexes 1 and 2) are used and each of them may have differently, depending on the choice of the modules for each gallery. First, the verification score is obtained from each gallery. The fusion score is the average between the two scores. Afterwards, the minimum score (threshold) required for classifying the sample as genuine is obtained by translating the *FusionDecisionScoreIndex* according to the *ClassificationScore* module. If the fusion score is higher than the threshold for the fusion (*fusionDecisionScore*), the query is classified as genuine and, otherwise, as impostor.

Similarly to the single gallery mode, if the query is classified as genuine, the adaptation procedure is launched. The same process used for adapting the single gallery is used here, but executed separately for each of the two galleries. The main difference is that the output score used is obtained from the fusion instead of using a score specific to the gallery.

7.2 Combiner

The *Combiner* is responsible for choosing the combination of modules for each user registered in the biometric system. In order to perform this task, the *Combiner* assumes that all users $j \in \mathcal{J}$ registered in the system can share information. It is a feasible assumption in several scenarios, like a company, in which the biometric system would have access to data from all employees. For *ModBioS*, it means that the biometric system is aware of the enrollment samples of all registered users.

Note that it does not mean that all users in the dataset are assumed to be known in the evaluation, since the user cross-validation for biometric data streams methodology has a separate set for unknown users. Thus, they are not used by the *Combiner*. First, the *Combiner* lists all possible combinations of module implementations in $\mathcal{M}_{all}$, which results in several hundred alternatives. This list depends on the *fusionMode*, the parameter which defines whether biometric score fusion is used or not by *ModBioS*. Then, it estimates the performance of each combination to select the best combination for the users. This performance estimation is done using the enrollment samples $\mathcal{E}$ from all registered users $j \in \mathcal{J}$.

The *ModBioS Combiner* can work in three settings: *Grouped*, *User* and *Hybrid*. In the *Grouped* setting (Algorithm 19), the *Combiner* considers the average performance for all registered users. As a consequence, all registered users assume the same combination of modules. It is named *Grouped* due to the evaluation methodology that assigns a different set of registered users depending on the *user grouping*. In the *User* setting (Algorithm 20), the performance is estimated per user. As a result, the combinations of modules can be different among all registered users. Additionally, a *Hybrid* setting was proposed, combining both *Grouped* and *User* characteristics. In the *Hybrid* setting (Algorithm 21), the w best combinations considering the average performance for all registered users are stored in a set. Then, the performance estimation per user is obtained from this filtered set, instead of using the complete set of combinations $\mathcal{M}_{all}$.

Algorithm 19: ModBioS Combiner (Grouped).

Data: $\mathcal{J}, \mathcal{E}, fusionMode$

- 1 $\mathcal{M}_{all} \leftarrow listAllCombinations(fusionMode)$
- 2 $\mathcal{M}_{all}^{perf} \leftarrow estimatePerformanceOnUserGrouping(\mathcal{J}, \mathcal{E}, \mathcal{M}_{all})$
- 3 $M_{best} \leftarrow getBestCombination(\mathcal{M}_{all}^{perf})$
- 4 **foreach** j **in** $\mathcal{J}$ **do**
- 5 $ref_j.M \leftarrow M_{best}$
- 6 **end**

Algorithm 20: ModBioS Combiner (User).

Data: $\mathcal{J}, \mathcal{E}, fusionMode$

- 1 $\mathcal{M}_{all} \leftarrow listAllCombinations(fusionMode)$
- 2 **foreach** j **in** $\mathcal{J}$ **do**
- 3 $\mathcal{M}_{all}^{perf} \leftarrow estimatePerformanceForUser(\mathcal{J}, \mathcal{E}, \mathcal{M}_{all})$
- 4 $M_{best} \leftarrow getBestCombination(\mathcal{M}_{all}^{perf})$
- 5 $ref_j.M \leftarrow M_{best}$
- 6 **end**

Algorithm 21: ModBioS Combiner (Hybrid).

Data: $\mathcal{J}, \mathcal{E}, fusionMode, w$

- 1 $\mathcal{M}_{all} \leftarrow listAllCombinations(fusionMode)$
- 2 $\mathcal{M}_{all}^{perf} \leftarrow estimatePerformanceOnUserGrouping(\mathcal{J}, \mathcal{E}, \mathcal{M}_{all})$
- 3 $\mathcal{M}_{wbest} \leftarrow getBestCombinations(\mathcal{M}_{all}^{perf}, w)$
- 4 **foreach** j **in** $\mathcal{J}$ **do**
- 5 $\mathcal{M}_{wbest}^{perf} \leftarrow estimatePerformanceForUser(\mathcal{J}, \mathcal{E}, \mathcal{M}_{wbest})$
- 6 $M_{best} \leftarrow getBestCombination(\mathcal{M}_{wbest}^{perf})$
- 7 $ref_j.M \leftarrow M_{best}$
- 8 **end**

As previously mentioned, the estimation of the recognition performance is done on the enrollment samples (there are N enrollment samples for each user). For such, the genuine samples of each user are split in order to use the first samples for enrollment and the remaining samples for test/validation in a short data stream. Enrollment samples from the other registered users are used as impostors in the data stream. Since this is a short validation data stream in a controlled scenario, the rate of genuine and impostor samples was set to 50%.

The genuine samples are split using a method based on the *serial waterfall* data split (RAYKAR; SAHA, 2015). First, N_E samples are used for enrollment and the subsequent N_{ds} samples for the validation data stream. Next, $N_E + 1$ samples are used for enrollment and the subsequent N_{ds} samples for the validation data stream. The number of enrollment validation samples keeps increasing at each iteration until the sum of the amount of enrollment and genuine data stream validation samples reach N (the total amount of enrollment samples of the user). The final performance value is the average performance obtained among all iterations. In the

current implementation, where $N = 40$ (Section 3.2), N_E was empirically set to the range $[11; 20]$ and $N_{ds} = 20$.

The use of *serial waterfall* keeps enrollment and validation sets always fresh as new data come, avoiding model fitting to a single set of data only. This is illustrated in Figure 34. Note that a standard cross-validation or leave-one-out would not be appropriate here since the order of the biometric samples must be maintained.

Figure 34 – *Serial waterfall* data split in *ModBioS*.

7.2.1 Current module implementations

The current implementations for each of the modules of *ModBioS* are shown in Table 14. Currently, there are several implementations for each module, as shown in the table. The modules G.3 to G.6 define the adaptation strategy, hence it is possible to obtain 80 combinations for single gallery using the current module implementations ($80 = 2 \times 5 \times 4 \times 2$).

Table 14 – Current module implementations.

Component	Cardinality	Domain
U 0 FusionMode	2	None (0), UpdateBothByFusion (1)
U 1 ClassificationScore	2	Self-Detector (1), M2005 (2)
U 2 FusionDecisionScoreIndex	5	[1;5]
G 1 DecisionScoreIndex	5	[1;5]
G 2 AdaptationScoreDiffIndex	1	[1;1]
G 3 DetectorParamManager	2	UpdateAllDetectors (1), UpdateOnlyFirstDetector (2)
G 4 AdaptationControl	5	AlwaysUpdateForPositive (0), CheckUsage (1), OnlyIfMoreThanOneMatch (2), CheckUsagePlusMoreThanOneMatch (3), AdditionalThresholdForAdaptation (4)
G 5 RemoveDetectorManager	4	NeverRemove (0), RemoveOldestOnly (1), RemoveSingleLessUsed (2), RemoveAllNotUsedRecently (3)
G 6 DetectorScanOrder	2	OlderToNewer (1), NewerToOlder (2)

In order to illustrate the search space for the *Combiner* of *ModBioS*, the number of combinations are described next (the values in bold represent the search space for each biometric modality). The amount for dual gallery also includes the combinations for single gallery. It must be observed that some combinations can result in the same behaviour/recognition performance. For example, when the combination simulates the behaviour of the *Sliding* adaptation strategy, the *DetectorScanOrder* module does not impact the performance.

- Adaptation strategies for single gallery: 80
- Adaptation strategies for single gallery compatible with *Self-Detector*: 80 (all)
- Adaptation strategies for single gallery compatible with *M2005*: 4 (just 4 adaptation strategies can be applied to *M2005*, since they are based on retraining: *Growing*, *Sliding*, *Growing* with additional adaptation threshold, *Sliding* with additional adaptation threshold; the remaining 76 are based on usage of detectors, so they can only be used with *Self-Detector*)
- Threshold variation: 5 (the range is composed by 5 values within the range [1; 5])
- Combinations for single gallery: $400 = 80 \times 5$ (adaptive strategy plus the threshold)
- Combinations for single gallery compatible with *Self-Detector*: 400 (all)
- Combinations for single gallery compatible with *M2005*: $20 = 4 \times 5$
- Combinations for single gallery (Keystroke dynamics): **420** (*Self-Det.* + *M2005*)
- Combinations for single gallery (Accelerometer-based gait biometrics): **400** (*Self-Det.*)
- Adaptation strategies for dual gallery: $3240 = \binom{80}{2} + 80$
- Adaptation strategies for dual gallery compatible with *Self-Detector*: 3240 (all)
- Adaptation strategies for dual gallery compatible with *M2005*: $10 = \binom{4}{2} + 4$
- Combinations for dual gallery (biometric fusion): $16200 = 3240 \times 5$
- Combinations for dual gallery compatible with *Self-Detector*: 16200 (all)
- Combinations for dual gallery compatible with *M2005*: $50 = 10 \times 5$
- Combinations for dual gallery (Keystroke dynamics): **16250** (*Self-Det.* + *M2005*)
- Combinations for dual gallery (Accelerometer-based gait biometrics): **16200** (*Self-Det.*)

7.2.2 Generalized adaptive biometric systems

As shown in the previous section, there are 3240 possible combinations of adaptation strategies modules that can be obtained from the current implementations of *ModBioS*. Some of them simulate the behaviour of baseline biometric systems in the literature as well as all *Usage Control* versions proposed by the research carried out for this thesis. These particular cases are illustrated in Table 15. Note that *ModBioS* can generate far more adaptation strategies than the

amount it generalizes from the literature, meaning that possible sound undiscovered adaptation strategies could be obtained from *ModBioS*.

Table 15 – Adaptive biometric systems generalized by *ModBioS*.

Adaptive Biometric System	FusionMode		ClassificationScore		FusionDecisionScoreIndex		DecisionScoreIndex		AdaptationScoreDiffIndex		DetectorParamManager		AdaptationControl		RemoveDetectorManager		DetectorScanOrder		DecisionScoreIndex		AdaptationScoreDiffIndex		DetectorParamManager		AdaptationControl		RemoveDetectorManager		DetectorScanOrder	
	U	0	U	1	U	2	G ¹	1	G ¹	2	G ¹	3	G ¹	4	G ¹	5	G ¹	6	G ²	1	G ²	2	G ²	3	G ²	4	G ²	5	G ²	6
<i>Self-Detector (Growing/Self-Update)</i>	0	1	1	1	-	-	*	-	-	-	0	0	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>Self-Detector (Growing/Self-Update with additional adaptation threshold)</i>	0	1	1	1	-	-	*	1	-	-	4	0	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>Self-Detector (Sliding/FIFO)</i>	0	1	1	1	-	-	*	-	-	-	0	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>Self-Detector (Sliding/FIFO with additional adaptation threshold)</i>	0	1	1	1	-	-	*	1	-	-	4	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>Self-Detector (Usage Control)</i>	0	1	1	1	-	-	*	-	2	1	2	1	2	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>Self-Detector (Usage Control R)</i>	0	1	1	1	-	-	*	-	2	1	2	2	2	2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>Self-Detector (Usage Control S)</i>	0	1	1	1	-	-	*	-	1	3	2	2	2	2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>Self-Detector (Usage Control 2)</i>	0	1	1	1	-	-	*	-	2	0	3	2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
<i>M2005 (I. Double Parallel)</i>	1	2	2	2	*	-	-	-	-	-	0	0	-	-	-	-	-	-	-	-	-	-	-	-	0	1	-	-	-	-

7.3 Experimental results

This section presents the results obtained from the experiments using *ModBioS*. All results are shown in Appendices A, D and E. The presentation of the results in this chapter is guided by a sequence of questions. These questions start from exploratory questions to check the hypothesis (checking that the best adaptation strategy changes among datasets and users) to questions regarding the practical usage of the modular adaptive biometric system.

7.3.1 Is the best adaptation strategy different for distinct datasets?

Firstly, is the best combination of adaptation strategy and classification algorithm different for distinct datasets? This question evaluates the case when all users assume the same adaptation strategy and classification algorithm. From the experimental results presented in the previous chapters, it was shown that some adaptation strategies are statistically better than others, though a single adaptation strategy is never *the best* on all datasets. Hence, the answer for this question is yes. However, here this question is further investigated, looking at the performance per user grouping, and using the proposed modular system. This thesis adopted the user cross-validation for biometric data streams methodology ($k = 5$), which divides the users in five sets to separate part of the users (one set) as unregistered impostors.

In order to answer this question, the performance for each adaptation strategy is plotted in the heat maps of Figure 35 (balanced accuracy). Each heat map was plotted using the test data. Since the evaluation of all combinations for all users in the test data requires too much computer resources, the results for this plot refer to the average performance after five executions. This

number of executions has little effect on this study, since the standard deviation among different executions is usually low.

Figure 35 – Performance of combinations per user grouping/dataset (adaptation strategy and classification algorithm). There are 84 combinations in the plots (single gallery). Each one is divided into five parts, corresponding to each user grouping of the evaluation methodology, as described in Section 3.2. The heat maps on the left present the balanced accuracy on the test data (the greener the better). The color scale is the same for all plots, meaning that the average performance in the GREYC-Web dataset is higher than in the CMU dataset. The plots on the right just highlight the best combination. When a draw occurs, more than one combination is highlighted.

As previously discussed, each *ModBioS* combination contains modules for the adaptation strategy, classification algorithm and other parameters (e.g. threshold). In line with this, if the plot considers the threshold variable, it is clear that different combinations would result in different performance. Hence, the analysis should remove this variable to properly evaluate if the best combination of adaptation strategy and classification algorithm differ among the users. The plots, therefore, show the performance of each combination of adaptation strategy and classification algorithm using the best threshold value. There are 84 possible combinations (80 for *Self-Detector* and 4 for *M2005*, as stated in Section 7.2.1). Just the combinations for single gallery are considered to answer this question, since they are enough for the analysis performed

in this section and in the next section too, which deals with a similar question.

In Figure 35, the plots on the left are heat maps in terms of balanced accuracy for each combination per user grouping (the greener the better), the scale of the colors is the same for all heat maps. From these plots, it is clear that the balanced accuracy obtained in GREYC-Web is higher than in the CMU dataset. This was also observed in the previous experiments reported by the research carried out for this thesis. Furthermore, the performance has higher variability among the combinations in the CMU dataset than in GREYC-Web. It illustrates that the performance improvement by a suitable combination in some cases (e.g. CMU) can be more significant than in others (e.g. GREYC-Web).

The plots on the right of Figure 35 highlight the best combination for each grouping (when there is a draw, more than one combination is highlighted). Figure 35 shows just two datasets for ease of reading, without affecting the obtained conclusions. The remaining plots can be found in Appendix D. From these two datasets, it is clear that the best combination varies among them and, sometimes, even for different user groupings in the same dataset.

Moreover, this analysis can be narrowed down to the adaptation strategy. From these plots, it can also be observed that different adaptation strategies for *Self-Detector* (the first 80 combinations in the graph) resulted in different performance depending on the dataset (the heatmap is mostly red in CMU and it is mostly green in GREYC-Web Logins, showing that the average performance was higher in GREYC-Web Logins than in CMU). This suggests that each adaptation strategy is better fitted for different application scenarios (datasets). Next section expands the current discussion to a per user perspective.

7.3.2 Is the best adaptation strategy different among users (same dataset)?

After verifying that the best combination of adaptation strategy and classification algorithm can differ among different datasets, another important question is whether the best adaptation strategy is different among users in the same dataset. Note that, in this new question, each user can assume a different combination of adaptation strategy and classification algorithm.

Previous studies suggest that the way the biometric features change over time, the *change pattern*, can be different among users in the same dataset (PISANI; LORENA; CARVALHO, 2015b; POH *et al.*, 2015). Another study suggested that different groups of users may need modifications in the adaptation strategy and parameters (RATTANI; MARCIALIS; ROLI, 2009). However, to the best of the author's knowledge, there is no previous study that experimentally illustrated the performance of different *adaptation strategies* to investigate whether it implies that different strategies should be chosen per user.

To answer this question, the performance for each adaptation strategy per user is plotted in the heat maps of figures 36 and 37 (balanced accuracy). Only the plots for two datasets are shown in the figures for ease of reading. They are enough for the conclusions drawn in

(a) CMU (balanced accuracy)

(b) CMU (best combination)

Figure 36 – Performance of combinations per user - CMU (adaptation strategy and classification algorithm). There are 84 combinations in the plots (single gallery). Each plot is divided into five parts, corresponding to each user grouping of the user cross-validation for biometric data streams methodology, as described in Section 3.2. The heat maps in the top present the balanced accuracy on the test data (the greener the better). The plots in the bottom just highlight the best combination. When a draw occurs, more than one combination is highlighted.

(a) GREYC-Web Logins (balanced accuracy)

(b) GREYC-Web Logins (best combination)

Figure 37 – Performance of combinations per user - GREYC-Web (Logins). There are 84 combinations (adaptation strategy and classification algorithm) in the plots (single gallery). Each plot is divided into five parts, corresponding to each user grouping of the user cross-validation for biometric data streams methodology, as described in Section 3.2. The heat maps in the top present the balanced accuracy on the test data (the greener the better). The plots in the bottom just highlight the best combination. When a draw occurs, more than one combination is highlighted.

this section (the remaining plots can be found in Appendix E). These plots follow the same methodology of the Figure 35, though now the balanced accuracy is computed per user.

From the plots in figures 36 and 37, it can be observed that the best combination varies among users in the same dataset. Moreover, the best combination varies for the same user depending on the user grouping too. This may be explained by the differences on the impostors that attack the user, which change in each user grouping. Furthermore, several combinations have the same best performance for some users, particularly user index User ID 32 in GREYC-Web, in which almost all combinations result in the same balanced accuracy.

Similarly to the the last section, the discussion can be narrowed down to the adaptation strategies for *Self-Detector* (the first 80 combinations in the plot). The plots suggest that the best adaptation strategy can be different depending on the user, indicating that each user may have different characteristics, including a distinct *change pattern*. It is in line with the results observed in previous work (PISANI; LORENA; CARVALHO, 2015b; POH *et al.*, 2015). Consequently, an adaptive biometric system like *ModBioS*, which is able to choose the adaptation strategy per user, would be capable of performing even better. To the best of the author's knowledge, *ModBioS* is the first adaptive biometric system proposed with these capabilities. In this proposal, selecting a different combination of modules can be understood as choosing a different adaptation strategy. The next section discusses the proposed methods to select the combination of modules within the *ModBioS* framework.

7.3.3 How to choose the combination of modules?

After the analysis of the previous questions, it is clear that the adaptive strategy should be chosen per user and per dataset to improve the recognition performance. However, these conclusions were obtained using the test data. Hence, the question now is: how to choose the combination of module implementations using only the available enrollment data? In *ModBioS*, this task is performed by the *Combiner*.

Two straightforward ways to select the combinations are to perform it by dataset/user grouping or by user. They represent the settings *Grouped* and *User* of the *ModBioS Combiner* described in Section 7.2, respectively. In the *Grouped* setting, all registered users assume the same combination (chosen for the *user grouping*), while, in the *User* setting, each registered user can assume a different combination (chosen for the user). The obtained results for both settings are shown in Table 16 (best results for each group are highlighted in bold and standard deviation among runs is shown between parenthesis). In this table, *ModBioS - Fusion* considers combinations with biometric fusion (two galleries), while *ModBioS* without the *Fusion* indication means that only the single gallery combinations are considered.

The purpose of this first analysis is to check whether it is better to choose the combinations in either *Grouped* or *User* settings. According to the results shown in Table 16, choosing combinations per user grouping resulted in better balanced accuracy in most cases, mainly for

Table 16 – Results for *ModBioS* using *Grouped* and *User* settings (all datasets). *ModBioS - Fusion* considers combinations with biometric fusion (two galleries), while *ModBioS* without the *Fusion* indication means that only the single gallery combinations are considered. The best results for each group are highlighted in bold (standard deviation among runs is shown between parenthesis).

Biometric system	GREYC			CMU		
	FMR	FNMR	Acc (balanc.)	FMR	FNMR	Acc (balanc.)
<i>ModBioS (User)</i>	0.153 (0.020)	0.095 (0.011)	0.876 (0.007)	0.215 (0.015)	0.222 (0.013)	0.781 (0.010)
<i>ModBioS (Grouped)</i>	0.077 (0.018)	0.151 (0.012)	0.886 (0.005)	0.212 (0.071)	0.186 (0.085)	0.801 (0.016)
<i>ModBioS - Fusion (User)</i>	0.158 (0.020)	0.086 (0.011)	0.878 (0.007)	0.213 (0.020)	0.196 (0.021)	0.796 (0.009)
<i>ModBioS - Fusion (Grouped)</i>	0.130 (0.107)	0.117 (0.042)	0.877 (0.033)	0.165 (0.066)	0.236 (0.067)	0.800 (0.004)
Biometric system	GREYC-Web (Logins)			GREYC-Web (Passwords)		
	FMR	FNMR	Acc (balanc.)	FMR	FNMR	Acc (balanc.)
<i>ModBioS (User)</i>	0.131 (0.021)	0.060 (0.011)	0.904 (0.008)	0.254 (0.025)	0.254 (0.020)	0.746 (0.012)
<i>ModBioS (Grouped)</i>	0.070 (0.019)	0.115 (0.030)	0.907 (0.009)	0.290 (0.016)	0.231 (0.011)	0.739 (0.010)
<i>ModBioS - Fusion (User)</i>	0.137 (0.020)	0.055 (0.006)	0.904 (0.009)	0.226 (0.018)	0.266 (0.020)	0.754 (0.010)
<i>ModBioS - Fusion (Grouped)</i>	0.090 (0.019)	0.084 (0.019)	0.913 (0.007)	0.292 (0.022)	0.219 (0.018)	0.745 (0.009)
Biometric system	McGill			WISDM 1.1		
	FMR	FNMR	Acc (balanc.)	FMR	FNMR	Acc (balanc.)
<i>ModBioS (User)</i>	0.350 (0.038)	0.211 (0.029)	0.719 (0.016)	0.129 (0.015)	0.244 (0.026)	0.814 (0.015)
<i>ModBioS (Grouped)</i>	0.320 (0.066)	0.168 (0.048)	0.756 (0.014)	0.166 (0.056)	0.201 (0.028)	0.817 (0.021)
<i>ModBioS - Fusion (User)</i>	0.370 (0.037)	0.214 (0.044)	0.708 (0.023)	0.130 (0.017)	0.228 (0.024)	0.821 (0.012)
<i>ModBioS - Fusion (Grouped)</i>	0.386 (0.145)	0.162 (0.045)	0.726 (0.054)	0.128 (0.025)	0.203 (0.023)	0.835 (0.011)
Biometric system	WISDM 2.0					
	FMR	FNMR	Acc (balanc.)			
<i>ModBioS (User)</i>	0.152 (0.009)	0.167 (0.009)	0.841 (0.006)			
<i>ModBioS (Grouped)</i>	0.130 (0.015)	0.172 (0.007)	0.849 (0.008)			
<i>ModBioS - Fusion (User)</i>	0.147 (0.011)	0.164 (0.012)	0.844 (0.007)			
<i>ModBioS - Fusion (Grouped)</i>	0.122 (0.017)	0.178 (0.029)	0.850 (0.008)			

the accelerometer-based gait biometrics datasets. Although the performance difference between *Grouped* and *User* is not large, it is an unexpected result. In theory, choosing the combinations per user should result in better overall performance than choosing a common combination for all registered users. It is certainly true if the test data could be used. However, using only the enrollment data, the performance of the *Grouped* setting is slightly better than *User*. This may be a result of *overfitting* to the user data in the *User* setting. Such a problem seems to be avoided when an average combination for all users is chosen, as the *Grouped* setting does. This suggests that new ways to choose the combinations considering information obtained over time could improve the choice of the modules.

Table 16 also reports the results of *ModBioS* using one and two galleries (biometric fusion). The balanced accuracy of *ModBioS - Fusion* is usually higher than that of *ModBioS* single gallery. This is expected since the two galleries setting can be understood as a small ensemble of two classifiers, increasing the robustness of the classification decision. Although the classification algorithm is the same, each classifier can be adapted using a different adaptation strategy in *ModBioS - Fusion*. Since the classification algorithm is the same, the diversity of this small ensemble is impaired and, therefore, it can explain the small performance improvement obtained by the biometric fusion over the single gallery system. A biometric system able to choose different classification algorithms for each gallery might result in even higher recognition performance for the *ModBioS - Fusion*.

7.3.3.1 Can both Grouped and User settings be combined?

Considering the results reported in the last section showing that the selection of modules per user grouping can be better than selecting it per user, a new question arises: can the recognition performance be improved by combining both settings: per user and per user grouping/dataset? To answer this question, a new setting for the *Combiner* was proposed, named *Hybrid*, as described in Section 7.2. This setting first applies the *Combiner* in the *Grouped* setting to filter the w best combinations that are then used as input for the *Combiner* in the *User* setting ($w = 10$ in the experiments).

The results for this new setting are shown in tables 17 and 18 (best results for each group are highlighted in bold and standard deviation among runs is shown between parenthesis). According to the reported results, the *Hybrid* setting obtained higher balanced accuracy in most cases, although by a small margin. A Bayesian statistical test between *Grouped* and *Hybrid* shows a higher probability that their balanced accuracy are equivalent. However, the author will consider the *ModBioS - Fusion (Hybrid)* as the reference *ModBioS* version for the next analysis. The *Hybrid* setting provides a compromise between the other settings, by combining the good performance of *Grouped* and the theoretical better concept of choosing a different adaptation strategy for each user. Such an approach can potentially avoid the overfitting to the data of a single user and, at the same time, choose a different combination per user.

Overall, the results shown in tables 17 and 18 suggest that *ModBioS - Fusion (Hybrid)* attains higher recognition performance in terms of balanced accuracy than most baselines. Although the choice of the combinations is done only on the enrollment data, the obtained results indicate that choosing the adaptation strategy for each user and dataset can result in higher recognition performance. Hence, a biometric system able to choose its modules for each case would be capable of reaching the best overall performance.

As described in Section 3.6, a Bayesian statistical test (CORANI *et al.*, 2016; BENAOLI *et al.*, 2016) was applied to check the results of *ModBioS - Fusion (Hybrid)* against all baselines in terms of balanced accuracy (some of the baselines here are biometric systems proposed in previous chapters of this thesis too, such as *Usage Control* and *ETU*). The results of the statistical test are shown in Table 19. The higher the probability on the right, the better is the performance of *ModBioS - Fusion (Hybrid)* compared to that of the baselines.

The results of the statistical test show that the proposed modular adaptive biometric system has a high probability of performing better than the baselines. In most of the cases, this probability is higher than 90%, showing that the current version of *ModBioS* has a good recognition performance. From the statistical test, it is also possible to observe $p(right)$ is lower than 80% in only three cases: *Self-Detector (Usage Control 2)*, *M2005 (IDB)* and *M2005 (ETU 3)*. It suggests that these baselines, overall, performed better than the other baselines. Nevertheless, even for these cases, the probability that *ModBioS - Fusion (Hybrid)* performs better is still above 70%.

Table 17 – Results for *ModBioS* using *Hybrid* setting (comparison to other baselines) - keystroke dynamics datasets. Only best ETU systems according to balanced accuracy are presented in this table. *ModBioS - Fusion* considers combinations with biometric fusion (two galleries), while *ModBioS* without the *Fusion* indication means that only the single gallery combinations are considered. Best results for each group are highlighted in bold (standard deviation among runs is shown between parenthesis). Adaptive biometric systems are indicated by the adaptation strategy between parenthesis like, for example, *Self-Detector (Sliding)*. Conversely, non-adaptive biometric systems do not use an adaptation strategy, hence, they do not contain a parenthesis in their names like, for example, *Self-Detector*.

Biometric system	GREYC			CMU		
	FMR	FNMR	Acc (balanc.)	FMR	FNMR	Acc (balanc.)
<i>Self-Detector</i>	0.090 (0.010)	0.165 (0.005)	0.872 (0.006)	0.287 (0.023)	0.410 (0.016)	0.651 (0.009)
<i>Self-Detector (Sliding)</i>	0.092 (0.011)	0.129 (0.004)	0.890 (0.006)	0.291 (0.031)	0.211 (0.013)	0.749 (0.016)
<i>Self-Detector (Growing)</i>	0.105 (0.011)	0.119 (0.005)	0.888 (0.006)	0.562 (0.039)	0.118 (0.009)	0.660 (0.018)
<i>M2005</i>	0.221 (0.019)	0.130 (0.003)	0.824 (0.009)	0.273 (0.028)	0.451 (0.019)	0.638 (0.013)
<i>M2005 (DB)</i>	0.220 (0.019)	0.086 (0.003)	0.847 (0.009)	0.129 (0.014)	0.373 (0.014)	0.749 (0.010)
<i>M2005 (IDB)</i>	0.210 (0.018)	0.092 (0.004)	0.849 (0.008)	0.122 (0.011)	0.306 (0.008)	0.786 (0.006)
<i>M2005 (Adapted Thresholds - Growing)</i>	0.239 (0.019)	0.092 (0.003)	0.835 (0.010)	0.462 (0.019)	0.113 (0.007)	0.712 (0.009)
<i>M2005 (Adapted Thresholds - Sliding)</i>	0.184 (0.017)	0.115 (0.004)	0.851 (0.008)	0.160 (0.011)	0.295 (0.010)	0.773 (0.007)
<i>Self-Detector (Usage Control)</i>	0.091 (0.010)	0.140 (0.005)	0.884 (0.006)	0.351 (0.033)	0.211 (0.013)	0.719 (0.016)
<i>Self-Detector (Usage Control 2)</i>	0.069 (0.009)	0.168 (0.006)	0.882 (0.006)	0.143 (0.012)	0.323 (0.014)	0.767 (0.009)
<i>Self-Detector (Usage Control R)</i>	0.092 (0.010)	0.140 (0.005)	0.884 (0.006)	0.311 (0.030)	0.220 (0.013)	0.735 (0.015)
<i>Self-Detector (Usage Control S)</i>	0.089 (0.010)	0.149 (0.005)	0.881 (0.006)	0.213 (0.014)	0.275 (0.012)	0.756 (0.008)
<i>Self-Detector (ETU 2)</i>	0.096 (0.010)	0.126 (0.005)	0.889 (0.006)	0.285 (0.019)	0.207 (0.013)	0.754 (0.013)
<i>M2005 (ETU 3)</i>	0.192 (0.017)	0.100 (0.004)	0.854 (0.008)	0.244 (0.016)	0.143 (0.006)	0.807 (0.009)
<i>Self-Detector (ETU - Sliding) PGP</i>	0.092 (0.011)	0.129 (0.004)	0.889 (0.006)	0.250 (0.021)	0.251 (0.015)	0.750 (0.011)
<i>ModBioS (User)</i>	0.153 (0.020)	0.095 (0.011)	0.876 (0.007)	0.215 (0.015)	0.222 (0.013)	0.781 (0.010)
<i>ModBioS (Grouped)</i>	0.077 (0.018)	0.151 (0.012)	0.886 (0.005)	0.212 (0.071)	0.186 (0.085)	0.801 (0.016)
<i>ModBioS - Fusion (User)</i>	0.158 (0.020)	0.086 (0.011)	0.878 (0.007)	0.213 (0.020)	0.196 (0.021)	0.796 (0.009)
<i>ModBioS - Fusion (Grouped)</i>	0.130 (0.107)	0.117 (0.042)	0.877 (0.033)	0.165 (0.066)	0.236 (0.067)	0.800 (0.004)
<i>ModBioS (Hybrid)</i>	0.102 (0.051)	0.132 (0.034)	0.883 (0.010)	0.214 (0.024)	0.207 (0.028)	0.790 (0.011)
<i>ModBioS - Fusion (Hybrid)</i>	0.102 (0.050)	0.123 (0.032)	0.888 (0.010)	0.216 (0.021)	0.159 (0.017)	0.812 (0.009)

Biometric system	GREYC-Web (Logins)			GREYC-Web (Passwords)		
	FMR	FNMR	Acc (balanc.)	FMR	FNMR	Acc (balanc.)
<i>Self-Detector</i>	0.066 (0.008)	0.141 (0.005)	0.896 (0.005)	0.388 (0.014)	0.180 (0.002)	0.716 (0.007)
<i>Self-Detector (Sliding)</i>	0.074 (0.011)	0.085 (0.004)	0.920 (0.007)	0.330 (0.021)	0.205 (0.008)	0.733 (0.012)
<i>Self-Detector (Growing)</i>	0.124 (0.015)	0.060 (0.003)	0.908 (0.008)	0.468 (0.017)	0.123 (0.002)	0.704 (0.008)
<i>M2005</i>	0.096 (0.013)	0.245 (0.016)	0.829 (0.008)	0.329 (0.035)	0.251 (0.015)	0.710 (0.024)
<i>M2005 (DB)</i>	0.083 (0.012)	0.179 (0.012)	0.869 (0.008)	0.251 (0.026)	0.240 (0.014)	0.754 (0.018)
<i>M2005 (IDB)</i>	0.095 (0.015)	0.131 (0.011)	0.887 (0.008)	0.247 (0.023)	0.190 (0.006)	0.781 (0.013)
<i>M2005 (Adapted Thresholds - Growing)</i>	0.132 (0.016)	0.303 (0.023)	0.782 (0.015)	0.273 (0.030)	0.348 (0.023)	0.690 (0.022)
<i>M2005 (Adapted Thresholds - Sliding)</i>	0.074 (0.011)	0.394 (0.020)	0.766 (0.010)	0.146 (0.022)	0.465 (0.022)	0.694 (0.018)
<i>Self-Detector (Usage Control)</i>	0.078 (0.011)	0.084 (0.003)	0.919 (0.006)	0.383 (0.021)	0.171 (0.006)	0.723 (0.011)
<i>Self-Detector (Usage Control 2)</i>	0.035 (0.007)	0.148 (0.010)	0.908 (0.007)	0.186 (0.013)	0.396 (0.012)	0.709 (0.010)
<i>Self-Detector (Usage Control R)</i>	0.069 (0.009)	0.086 (0.004)	0.922 (0.006)	0.360 (0.022)	0.188 (0.006)	0.726 (0.012)
<i>Self-Detector (Usage Control S)</i>	0.053 (0.007)	0.123 (0.005)	0.912 (0.005)	0.255 (0.017)	0.296 (0.009)	0.725 (0.009)
<i>Self-Detector (ETU 2)</i>	0.103 (0.014)	0.071 (0.009)	0.913 (0.010)	0.364 (0.020)	0.171 (0.009)	0.733 (0.011)
<i>M2005 (ETU 3)</i>	0.163 (0.017)	0.118 (0.018)	0.860 (0.013)	0.294 (0.016)	0.177 (0.021)	0.765 (0.013)
<i>Self-Detector (ETU - Sliding) PGP</i>	0.067 (0.010)	0.106 (0.008)	0.913 (0.008)	0.292 (0.023)	0.254 (0.011)	0.727 (0.012)
<i>ModBioS (User)</i>	0.131 (0.021)	0.060 (0.011)	0.904 (0.008)	0.254 (0.025)	0.254 (0.020)	0.746 (0.012)
<i>ModBioS (Grouped)</i>	0.070 (0.019)	0.115 (0.030)	0.907 (0.009)	0.290 (0.016)	0.231 (0.011)	0.739 (0.010)
<i>ModBioS - Fusion (User)</i>	0.137 (0.020)	0.055 (0.006)	0.904 (0.009)	0.226 (0.018)	0.266 (0.020)	0.754 (0.010)
<i>ModBioS - Fusion (Grouped)</i>	0.090 (0.019)	0.084 (0.019)	0.913 (0.007)	0.292 (0.022)	0.219 (0.018)	0.745 (0.009)
<i>ModBioS (Hybrid)</i>	0.084 (0.016)	0.099 (0.021)	0.908 (0.006)	0.263 (0.036)	0.230 (0.030)	0.754 (0.019)
<i>ModBioS - Fusion (Hybrid)</i>	0.082 (0.020)	0.092 (0.027)	0.913 (0.010)	0.295 (0.017)	0.214 (0.012)	0.745 (0.010)

Table 18 – Results for *ModBioS* using *Hybrid* setting (comparison to other baselines) - accelerometer datasets. Only best ETU systems according to balanced accuracy are presented in this table. *ModBioS - Fusion* considers combinations with biometric fusion (two galleries), while *ModBioS* without the *Fusion* indication means that only the single gallery combinations are considered. The best results for each group are highlighted in bold (standard deviation among runs is shown between parenthesis).

Biometric system	McGill			WISDM 1.1		
	FMR	FNMR	Acc (balanc.)	FMR	FNMR	Acc (balanc.)
<i>Self-Detector</i>	0.104 (0.018)	0.545 (0.023)	0.675 (0.019)	0.183 (0.013)	0.242 (0.013)	0.788 (0.009)
<i>Self-Detector (Sliding)</i>	0.166 (0.018)	0.282 (0.036)	0.776 (0.023)	0.227 (0.028)	0.186 (0.021)	0.794 (0.016)
<i>Self-Detector (Growing)</i>	0.490 (0.036)	0.116 (0.029)	0.697 (0.019)	0.349 (0.024)	0.123 (0.014)	0.764 (0.016)
<i>Self-Detector (Usage Control)</i>	0.294 (0.035)	0.230 (0.037)	0.738 (0.027)	0.261 (0.030)	0.163 (0.024)	0.788 (0.017)
<i>Self-Detector (Usage Control 2)</i>	0.077 (0.007)	0.356 (0.038)	0.784 (0.021)	0.123 (0.015)	0.213 (0.021)	0.832 (0.011)
<i>Self-Detector (Usage Control R)</i>	0.233 (0.028)	0.250 (0.036)	0.758 (0.024)	0.227 (0.027)	0.174 (0.024)	0.799 (0.017)
<i>Self-Detector (Usage Control S)</i>	0.115 (0.016)	0.422 (0.035)	0.732 (0.023)	0.152 (0.014)	0.237 (0.015)	0.805 (0.010)
<i>Self-Detector (ETU 2)</i>	0.209 (0.026)	0.315 (0.044)	0.738 (0.033)	0.217 (0.028)	0.191 (0.021)	0.796 (0.017)
<i>Self-Detector (ETU - Sliding) PGP</i>	0.134 (0.013)	0.435 (0.045)	0.715 (0.025)	0.202 (0.025)	0.195 (0.021)	0.801 (0.015)
<i>ModBioS (User)</i>	0.350 (0.038)	0.211 (0.029)	0.719 (0.016)	0.129 (0.015)	0.244 (0.026)	0.814 (0.015)
<i>ModBioS (Grouped)</i>	0.320 (0.066)	0.168 (0.048)	0.756 (0.014)	0.166 (0.056)	0.201 (0.028)	0.817 (0.021)
<i>ModBioS - Fusion (User)</i>	0.370 (0.037)	0.214 (0.044)	0.708 (0.023)	0.130 (0.017)	0.228 (0.024)	0.821 (0.012)
<i>ModBioS - Fusion (Grouped)</i>	0.386 (0.145)	0.162 (0.045)	0.726 (0.054)	0.128 (0.025)	0.203 (0.023)	0.835 (0.011)
<i>ModBioS (Hybrid)</i>	0.355 (0.090)	0.166 (0.045)	0.740 (0.031)	0.146 (0.025)	0.195 (0.023)	0.829 (0.009)
<i>ModBioS - Fusion (Hybrid)</i>	0.360 (0.104)	0.158 (0.035)	0.741 (0.037)	0.129 (0.023)	0.200 (0.023)	0.835 (0.010)

WISDM 2.0			
Biometric system	FMR	FNMR	Acc (balanc.)
<i>Self-Detector</i>	0.168 (0.006)	0.220 (0.007)	0.806 (0.005)
<i>Self-Detector (Sliding)</i>	0.203 (0.015)	0.150 (0.008)	0.824 (0.008)
<i>Self-Detector (Growing)</i>	0.302 (0.018)	0.116 (0.009)	0.791 (0.008)
<i>Self-Detector (Usage Control)</i>	0.227 (0.016)	0.136 (0.008)	0.819 (0.008)
<i>Self-Detector (Usage Control 2)</i>	0.123 (0.007)	0.176 (0.009)	0.851 (0.006)
<i>Self-Detector (Usage Control R)</i>	0.198 (0.014)	0.144 (0.009)	0.829 (0.007)
<i>Self-Detector (Usage Control S)</i>	0.156 (0.007)	0.185 (0.010)	0.829 (0.006)
<i>Self-Detector (ETU 2)</i>	0.206 (0.010)	0.154 (0.007)	0.820 (0.006)
<i>Self-Detector (ETU - Sliding) PGP</i>	0.179 (0.011)	0.178 (0.009)	0.822 (0.006)
<i>ModBioS (User)</i>	0.152 (0.009)	0.167 (0.009)	0.841 (0.006)
<i>ModBioS (Grouped)</i>	0.130 (0.015)	0.172 (0.007)	0.849 (0.008)
<i>ModBioS - Fusion (User)</i>	0.147 (0.011)	0.164 (0.012)	0.844 (0.007)
<i>ModBioS - Fusion (Grouped)</i>	0.122 (0.017)	0.178 (0.029)	0.850 (0.008)
<i>ModBioS (Hybrid)</i>	0.141 (0.007)	0.176 (0.013)	0.842 (0.008)
<i>ModBioS - Fusion (Hybrid)</i>	0.112 (0.012)	0.187 (0.025)	0.851 (0.008)

Table 19 – Bayesian statistical test (balanced accuracy): *ModBioS - Fusion (Hybrid)* vs baseline systems. For each baseline biometric system, three probabilities are respectively reported: $p(left)$, $p(rope)$ and $p(right)$. The higher the probability on the right, the better is the performance of *ModBioS - Fusion (Hybrid)* compared to that of the baselines. The table is divided into 3 sections. The first section presents the results for *Self-Detector*, that was applied to all datasets. The second section presents the results for *M2005*, which was only employed in the keystroke dynamics datasets. The third section presents the results for *OCSVM*, which was only employed in the accelerometer-based gait biometrics datasets.

	ModBioS - Fusion (Hybrid)
<i>Self-Detector</i>	2% 0% 98%
<i>Self-Detector (Sliding)</i>	12% 1% 87%
<i>Self-Detector (Growing)</i>	2% 0% 98%
<i>Self-Detector (Usage Control)</i>	6% 0% 94%
<i>Self-Detector (Usage Control 2)</i>	13% 10% 77%
<i>Self-Detector (Usage Control R)</i>	8% 0% 91%
<i>Self-Detector (Usage Control S)</i>	2% 1% 97%
<i>Self-Detector (ETU 2)</i>	4% 1% 95%
<i>Self-Detector (ETU - Sliding) PGP</i>	3% 0% 97%
<i>M2005</i>	6% 0% 94%
<i>M2005 (DB)</i>	9% 0% 91%
<i>M2005 (IDB)</i>	26% 0% 74%
<i>M2005 (Adapted Thresholds - Growing)</i>	3% 0% 97%
<i>M2005 (Adapted Thresholds - Sliding)</i>	8% 0% 92%
<i>M2005 (ETU 3)</i>	21% 1% 78%
<i>OCSVM</i>	5% 0% 95%
<i>OCSVM (Growing)</i>	5% 0% 95%
<i>OCSVM (Sliding)</i>	2% 0% 98%

7.3.3.2 Why the estimation on the enrollment data is not optimal?

ModBioS is able to generalize the behaviour of several adaptive biometric systems, as described in Section 7.2.2. Therefore, it should obtain equal or higher overall recognition performance than these systems in all datasets. However, sometimes a baseline algorithm performs better than *ModBioS*. For example, in GREYC-Web (Passwords), the balanced accuracy of *M2005 (IDB)* was 0.781 while *ModBioS - Fusion (Hybrid)* obtained 0.745. As *ModBioS - Fusion (Hybrid)* can generalize *M2005 (IDB)*, it means that the choice of the module implementations was not optimal for this dataset.

A key aspect regarding the *ModBioS Combiner* is that the choice of the combination is made using a limited amount of data (enrollment samples from the registered users). There is a chance that the user characteristics, including the behaviour *change pattern*, is not properly present in this data for all users, resulting in a non-optimal choice of modules. In order to check it, the correlations graph presented in Section 4.2.2 is plot here for all users in the CMU dataset (first user grouping) in Figure 38. The figure contains two sets of plots. The first set shows the correlations on the enrollment data only, using the first half to train and the second to check the correlations. The second set of plots shows the correlation in the actual test setting, using the test data. In summary, the goal is to show the behaviour observed in the data used for *ModBioS* (first set of plots) and the behaviour on the complete dataset (second set of plots).

(a) Enrollment samples only.

(b) All samples.

Figure 38 – Correlations over time: comparing enrollment samples to all samples. The top part shows the correlation to the closest detector considering only the enrollment samples, while the bottom part shows the correlation to the closest detector considering all genuine samples (first user grouping of CMU dataset).

By comparing both sets of plots, it is clear that the behaviour change pattern is different on both sets for several users (e.g. users 1 and 48), though the length of the data streams are different. Conversely, for other users (e.g. 26 and 32), the behaviour on the enrollment data seems to have some relation to the behaviour observed in the test data. This variation in the user characteristics between enrollment and test makes it difficult for the *Combiner* to choose the best modules for these users.

This finding highlights the fact that *ModBioS* requires more data to be able to choose optimal combinations for the users. Although the modular framework provides extensive capabilities, such as being able to generalize some baseline systems, the task of selecting the optimal modules is hard in view of the limited amount of data used. An idea for future work is to use data acquired over time to improve the selection of modules. By using these data, additional information regarding the user and its *change pattern* could be obtained to refine the choice of the module combinations.

7.3.3.3 Is the current choice of combinations any better than random choice?

After the finding of the previous section, another question could be asked: is the current choice of combinations any better than random choice? Since several users show a different behaviour between enrollment and test data, choosing combinations randomly might result in similar performance to that of the adopted *Combiner* settings. In order to answer this question, the performance of the *Hybrid Combiner* is compared to a random combination choice, shown in Table 20 (best results for each group are highlighted in bold and standard deviation among runs is shown between parenthesis).

As shown in Table 20, *Random Combiner* is clearly worse than using the *Hybrid Combiner* (balanced accuracy). Hence, the *Hybrid Combiner* is being able to identify patterns on the enrollment data that indicate the performance on the test data.

However, a further question could be asked: is the current *Combiner* just choosing good threshold values, but it is failing to choose a good adaptation strategy? In order to answer this additional question, another variation of *Random* is evaluated: *Random - Tuned*. In *Random - Tuned*, the modules that refer to the threshold are optimized by the *Combiner*, leaving the other modules random. As expected, the balanced accuracy of *Random - Tuned* is higher than *Random*. However, it is still lower than the performance of the *Hybrid Combiner*. A single exception is in the GREYC-Web datasets. It may be a result of a larger amount of adaptation strategies that work well in these datasets, meaning that even a random choice was able to select an appropriate strategy.

The results were also checked using a Bayesian statistical test (CORANI *et al.*, 2016; BENAVALI *et al.*, 2016). The baselines are the two random *Combiners*, while the proposal is the *Hybrid* version. The results of the statistical test are shown in Table 21. The Bayesian statistical test done here outputs the probabilities that *ModBioS - Fusion (Hybrid)* is worse,

equivalent or better than the random baselines in terms of balanced accuracy. As reported in Table 21, the values of $p(right)$ are above 90%, showing that the choice made by the *Hybrid Combiner* is better than a random choice of modules.

Table 20 – Performance comparing *ModBioS* using *Hybrid* and *Random Combiner*. There are two random Combiners: *Random* chooses all modules randomly, while *Random - Tuned* works similarly, but it also optimizes the modules that refer to the threshold. The best results for each group are highlighted in bold (standard deviation among runs is shown between parenthesis).

Biometric system	GREYC			CMU		
	FMR	FNMR	Acc (balanc.)	FMR	FNMR	Acc (balanc.)
<i>ModBioS (Hybrid)</i>	0.102 (0.051)	0.132 (0.034)	0.883 (0.010)	0.214 (0.024)	0.207 (0.028)	0.790 (0.011)
<i>ModBioS - Fusion (Hybrid)</i>	0.102 (0.050)	0.123 (0.032)	0.888 (0.010)	0.216 (0.021)	0.159 (0.017)	0.812 (0.009)
<i>ModBioS (Random)</i>	0.036 (0.009)	0.452 (0.035)	0.756 (0.016)	0.473 (0.055)	0.234 (0.040)	0.646 (0.017)
<i>ModBioS (Random - Tuned)</i>	0.089 (0.012)	0.156 (0.007)	0.877 (0.006)	0.344 (0.035)	0.223 (0.019)	0.716 (0.014)
<i>ModBioS - Fusion (Random)</i>	0.030 (0.008)	0.458 (0.034)	0.756 (0.016)	0.472 (0.052)	0.224 (0.036)	0.652 (0.017)
<i>ModBioS - Fusion (Random - Tuned)</i>	0.079 (0.010)	0.157 (0.005)	0.882 (0.005)	0.326 (0.036)	0.229 (0.019)	0.722 (0.014)
Biometric system	GREYC-Web (Logins)			GREYC-Web (Passwords)		
	FMR	FNMR	Acc (balanc.)	FMR	FNMR	Acc (balanc.)
<i>ModBioS (Hybrid)</i>	0.084 (0.016)	0.099 (0.021)	0.908 (0.006)	0.263 (0.036)	0.230 (0.030)	0.754 (0.019)
<i>ModBioS - Fusion (Hybrid)</i>	0.082 (0.020)	0.092 (0.027)	0.913 (0.010)	0.295 (0.017)	0.214 (0.012)	0.745 (0.010)
<i>ModBioS (Random)</i>	0.049 (0.016)	0.290 (0.052)	0.831 (0.023)	0.550 (0.054)	0.113 (0.026)	0.668 (0.021)
<i>ModBioS (Random - Tuned)</i>	0.089 (0.015)	0.080 (0.015)	0.916 (0.008)	0.447 (0.043)	0.147 (0.025)	0.703 (0.015)
<i>ModBioS - Fusion (Random)</i>	0.044 (0.014)	0.286 (0.050)	0.835 (0.023)	0.545 (0.044)	0.104 (0.023)	0.676 (0.018)
<i>ModBioS - Fusion (Random - Tuned)</i>	0.078 (0.015)	0.077 (0.010)	0.923 (0.008)	0.425 (0.038)	0.150 (0.023)	0.713 (0.012)
Biometric system	McGill			WISDM 1.1		
	FMR	FNMR	Acc (balanc.)	FMR	FNMR	Acc (balanc.)
<i>ModBioS (Hybrid)</i>	0.355 (0.090)	0.166 (0.045)	0.740 (0.031)	0.146 (0.025)	0.195 (0.023)	0.829 (0.009)
<i>ModBioS - Fusion (Hybrid)</i>	0.360 (0.104)	0.158 (0.035)	0.741 (0.037)	0.129 (0.023)	0.200 (0.023)	0.835 (0.010)
<i>ModBioS (Random)</i>	0.549 (0.067)	0.169 (0.050)	0.641 (0.026)	0.590 (0.049)	0.101 (0.029)	0.655 (0.022)
<i>ModBioS (Random - Tuned)</i>	0.432 (0.066)	0.197 (0.057)	0.686 (0.030)	0.249 (0.031)	0.186 (0.031)	0.783 (0.018)
<i>ModBioS - Fusion (Random)</i>	0.521 (0.068)	0.187 (0.048)	0.646 (0.026)	0.580 (0.051)	0.108 (0.027)	0.656 (0.023)
<i>ModBioS - Fusion (Random - Tuned)</i>	0.401 (0.061)	0.233 (0.050)	0.683 (0.030)	0.228 (0.028)	0.189 (0.027)	0.791 (0.017)
Biometric system	WISDM 2.0					
	FMR	FNMR	Acc (balanc.)			
<i>ModBioS (Hybrid)</i>	0.141 (0.007)	0.176 (0.013)	0.842 (0.008)			
<i>ModBioS - Fusion (Hybrid)</i>	0.112 (0.012)	0.187 (0.025)	0.851 (0.008)			
<i>ModBioS (Random)</i>	0.503 (0.022)	0.094 (0.010)	0.701 (0.010)			
<i>ModBioS (Random - Tuned)</i>	0.235 (0.015)	0.133 (0.010)	0.816 (0.008)			
<i>ModBioS - Fusion (Random)</i>	0.494 (0.021)	0.098 (0.011)	0.704 (0.010)			
<i>ModBioS - Fusion (Random - Tuned)</i>	0.220 (0.013)	0.138 (0.009)	0.821 (0.007)			

Table 21 – Bayesian statistical test (balanced accuracy): *Hybrid* Combiner vs random choice. For each random baseline, three probabilities are respectively reported: $p(left)$, $p(rope)$ and $p(right)$. The higher the probability on the right, the better is the performance of *Hybrid* compared to that of the baselines.

	ModBioS (User)	ModBioS (Grouped)	ModBioS (Hybrid)
ModBioS (Random)	0% 0% 100%	0% 0% 100%	0% 0% 100%
ModBioS (Random - Tuned)	3% 0% 97%	2% 0% 98%	2% 0% 98%
	ModBioS - Fusion (User)	ModBioS - Fusion (Grouped)	ModBioS - Fusion (Hybrid)
ModBioS - Fusion (Random)	0% 0% 100%	0% 0% 100%	0% 0% 100%
ModBioS - Fusion (Random - Tuned)	5% 0% 94%	4% 0% 96%	3% 0% 97%

7.3.3.4 Could the baselines benefit from the *ModBioS* parameter tuning?

According to the results presented in the previous section, the *Hybrid Combiner* is better than a random choice of combinations for *ModBioS*. However, what if the baseline systems used the same tuning system from *ModBioS* to tune the threshold? Would the choices made by the *Hybrid Combiner* be still better than the tuned baselines? Currently, the baselines adopt a

global parameter setting, i.e., all users assume the same parameter values, similarly to previous work (PISANI; LORENA; CARVALHO, 2015b; PISANI *et al.*, 2016). By using the *ModBioS Combiner*, the parameters (e.g. threshold) can be set per user. To answer this question, the balanced accuracy of the baselines using *ModBioS* tuning is compared to the *ModBioS (Hybrid)* in Table 22. The baselines use the *ModBioS* system just to tune the threshold, so all users assume the same adaptation strategy. Note that only the adaptive biometric systems that *ModBioS* can generalize are part of the comparison.

Table 22 – Balanced accuracy of baseline systems using the *ModBioS* tuning. The best results are highlighted in bold (standard deviation among runs is shown between parenthesis).

Biometric system	GREYC	CMU	GREYC-Web (Logins)	GREYC-Web (Passwords)
<i>ModBioS (Hybrid)</i>	0.883 (0.010)	0.790 (0.011)	0.908 (0.006)	0.754 (0.019)
<i>ModBioS - Fusion (Hybrid)</i>	0.888 (0.010)	0.812 (0.009)	0.913 (0.010)	0.745 (0.010)
<i>Self-Detector (Sliding - ModBioS Tuned)</i>	0.885 (0.005)	0.730 (0.017)	0.925 (0.006)	0.688 (0.012)
<i>Self-Detector (Growing - ModBioS Tuned)</i>	0.885 (0.005)	0.662 (0.016)	0.911 (0.008)	0.666 (0.011)
<i>Self-Detector (Usage Control - ModBioS Tuned)</i>	0.882 (0.005)	0.699 (0.016)	0.920 (0.008)	0.670 (0.017)
<i>Self-Detector (Usage Control R - ModBioS Tuned)</i>	0.881 (0.005)	0.714 (0.016)	0.925 (0.006)	0.685 (0.013)
<i>Self-Detector (Usage Control S - ModBioS Tuned)</i>	0.878 (0.005)	0.744 (0.009)	0.918 (0.004)	0.724 (0.011)
<i>Self-Detector (Usage Control 2 - ModBioS Tuned)</i>	0.876 (0.005)	0.763 (0.009)	0.920 (0.006)	0.722 (0.012)
<i>M2005 (I. Double Parallel - ModBioS Tuned)</i>	0.858 (0.010)	0.813 (0.012)	0.878 (0.019)	0.780 (0.011)
Biometric system	McGill	WISDM 1.1	WISDM 2.0	
<i>ModBioS (Hybrid)</i>	0.740 (0.031)	0.829 (0.009)	0.842 (0.008)	
<i>ModBioS - Fusion (Hybrid)</i>	0.741 (0.037)	0.835 (0.010)	0.851 (0.008)	
<i>Self-Detector (Sliding - ModBioS Tuned)</i>	0.704 (0.024)	0.779 (0.020)	0.819 (0.008)	
<i>Self-Detector (Growing - ModBioS Tuned)</i>	0.631 (0.011)	0.756 (0.017)	0.786 (0.008)	
<i>Self-Detector (Usage Control - ModBioS Tuned)</i>	0.661 (0.029)	0.768 (0.020)	0.806 (0.009)	
<i>Self-Detector (Usage Control R - ModBioS Tuned)</i>	0.685 (0.027)	0.780 (0.020)	0.819 (0.006)	
<i>Self-Detector (Usage Control S - ModBioS Tuned)</i>	0.719 (0.028)	0.789 (0.013)	0.826 (0.005)	
<i>Self-Detector (Usage Control 2 - ModBioS Tuned)</i>	0.767 (0.014)	0.824 (0.012)	0.852 (0.006)	

It is usually expected that adjusting the parameters per user would result in higher recognition performance. Nonetheless, it is not the case for all biometric systems. Some of them performed worse than before by using the *ModBioS* parameter tuning, while others managed to increase the balanced accuracy. These results can be observed by comparing the results from tables 17 and 18 to Table 22. For example, *Sliding* in the CMU dataset reduced the balanced accuracy with the per user tuning. Conversely, *Usage Control 2* improved the performance using the per user tuning in *GREYC-Web*. It suggests that the same overfitting for the data of a single user observed by the *User* setting of the *Combiner* can also occur when the parameter tuning per user is performed. In addition, each time, a different biometric system attains the best performance among the baselines. This issue is also observed when the tuning is done globally, i.e., when all users assume the same parameter values.

The same Bayesian statistical test was also applied to these results, as shown in Table 23. The test reports that the probability of the performance of *ModBioS - Fusion (Hybrid)* being better than that of the baselines is higher than 95% for most tuned baselines. There are two exceptions: *Self-Detector (Usage Control 2)* and *M2005 (I. Double Parallel)*, whose probabilities of *ModBioS* performing better are 76% and 63% respectively. According to these results, these two biometric systems benefited from the threshold tuning per user of *ModBioS*. The probabilities are still higher for *ModBioS* due to its ability to also choose the adaptation strategy and

classification algorithm per user. However, the obtained results highlight the fact the method to choose the combinations per user for *ModBioS* should be further investigated in future work.

Table 23 – Bayesian statistical test (balanced accuracy): *Hybrid* vs tuned baselines. For each baseline, three probabilities are respectively reported: $p(left)$, $p(rope)$ and $p(right)$. The higher the probability on the right, the better is the performance of the *Hybrid* compared to that of the baselines.

	ModBioS - Fusion (Hybrid)
<i>Self-Detector (Sliding - ModBioS Tuned)</i>	2% 0% 98%
<i>Self-Detector (Growing - ModBioS Tuned)</i>	1% 0% 99%
<i>Self-Detector (Usage Control - ModBioS Tuned)</i>	1% 0% 99%
<i>Self-Detector (Usage Control R - ModBioS Tuned)</i>	3% 0% 97%
<i>Self-Detector (Usage Control S - ModBioS Tuned)</i>	3% 0% 97%
<i>Self-Detector (Usage Control 2 - ModBioS Tuned)</i>	12% 12% 76%
<i>M2005 (I. Double Parallel - ModBioS Tuned)</i>	35% 2% 63%

7.3.4 Adaptive modular biometric system performance over time

Adaptive biometric systems adapt the biometric reference to the genuine data over time. In terms of metrics, it can be understood as to avoid higher FNMR values over time without increasing FMR. FNMR can increase if the biometric system is not able to properly recognize the genuine users over time. In order to study FMR and FNMR, the performance over time of these metrics are shown in figures 39 and 41. The GREYC dataset is not part of this graph since its data streams are the shortest among the seven datasets. According to the Figure 39, it is clear that the behaviour in the keystroke dynamics datasets is different from that of the accelerometer-based gait biometrics datasets. Hence, one dataset for each modality was chosen for the study of this section: CMU and McGill (figures 40 and 42). They contain the highest average number of samples per user in each biometric modality, as shown in tables 2 and 3.

For keystroke dynamics, *ModBioS Fusion (Hybrid)* manages to keep this metric at a stable value over time. In the case of CMU, shown in Figure 40, FNMR is above 0.25 for several adaptive biometric systems, while this metric for *ModBioS Fusion (Hybrid)* is below 0.25. This is a good result, since *ModBioS* obtained better FNMR over time for this dataset. Regarding the accelerometer-based gait biometrics, the FNMR has a tendency to increase over time for *ModBioS* as shown in Figure 39. However, the FNMR obtained by *ModBioS* is lower than most baseline systems in the McGill dataset, as shown in Figure 40. For instance, the average FNMR is below 0.50 most of the time for *ModBioS*, while the baselines reach values above 0.50 in several cases.

FMR is also evaluated in this section. As shown in Figure 41, FMR tends to decrease over time for keystroke dynamics and has a higher variation over time for accelerometer-based gait biometrics. Comparing this result to the FMR of the keystroke dynamics dataset CMU in Figure 42, it can be seen that the *ModBioS* performance is better than most baselines. The FMR obtained by *ModBioS* in this dataset is closer to the best baselines in terms of FMR: *M2005 (DB and IDB)* and *Self-Detector (Usage Control 2)*. For accelerometer-based gait biometrics in the McGill dataset, however, the baselines obtained better values for FMR.

Figure 39 – FNMR over time of *ModBioS Fusion (Hybrid)* - Average from all users. Figure 40 – FNMR over time of *Usage Control* and baselines - Average from all users.

Figure 41 – FMR over time of *ModBioS Fusion (Hybrid)* - Average from all users.

Figure 42 – FMR over time of *Usage Control* and baselines - Average from all users.

7.4 Extensions for the modular system

The current version of *ModBioS* has already obtained good recognition performance in the experiments previously reported. However, the *ModBioS* performance could be further improved by expanding the proposed modular adaptive biometric system.

One of the first and perhaps easiest next expansion would be to include additional implementations for the current modules. Another possible extension would be to extend the modular framework in order to take into account other aspects such as an impostor gallery. As a result, *ModBioS* would be able to generalize ETU systems too.

Nevertheless, perhaps the change that would result in higher benefits for the recognition performance of *ModBioS* is a module for score normalization. As reported in Chapter 6, score normalization can improve the recognition performance of adaptive biometric systems. As discussed in the last chapter, the best score normalization can be different depending on the user. Moreover, the heat maps in Appendix C show that the performance difference among the score normalization procedures can be high. In some cases, a user obtained higher balanced accuracy without any score normalization processing. This result shows that an adaptive biometric system able to choose the score normalization per user may obtain higher recognition performance than a system that adopts the same score normalization for all registered users. Hence, this can be a promising new module to be included in *ModBioS*. The implementation of this module is not a trivial task, since the score normalization would need to consider the translation between score indexes and thresholds. This translation exists to enable *ModBioS* to deal with different classification algorithms, which can output scores at a different range of values.

7.5 Chapter remarks

A number of adaptive biometric systems have been proposed in the literature. As discussed in this chapter, the best adaptation strategy can be different depending on the dataset. Moreover, different users may have different *change patterns* (PISANI; LORENA; CARVALHO, 2015b; POH *et al.*, 2015). Another study also suggested that different groups of users may need modifications in the adaptation strategy and parameters (RATTANI; MARCIALIS; ROLI, 2009). All in all, the different characteristics among users, including the *change pattern*, indicates that the adaptation strategy should be chosen per user. To the best of the author's knowledge, an adaptive biometric system with these capabilities has never been proposed. In line with this concept, a modular adaptive biometric system was proposed, named *ModBioS*. This new system can choose different adaptation strategies per user by selecting the module implementations for each registered user.

ModBioS can generalize several baseline adaptive biometric systems, including some systems proposed by the research carried out for this thesis, such as the *Usage Control* versions. Additionally, *ModBioS* could implement possible sound undiscovered adaptation strate-

gies through the combination of the current modules. Hence, in theory, *ModBioS* should attain an equal or higher recognition performance than the generalized systems. However, according to the reported results, *ModBioS* could not reach the best balanced accuracy in all datasets. This happened due to a sub-optimal choice of modules. As previously discussed, a key problem for such a sub-optimal choice is the use of just the enrollment samples for choosing the combinations. In future work, additional sources of information could be used to improve the choice of modules. The information obtained over time by the biometric system is a possible source of information to improve the combinations.

Additional module implementations can be developed in the future to enable *ModBioS* to generalize adaptive biometric systems other than those discussed here. Furthermore, the proposed modular framework may be expanded to deal with additional modules. For instance, the capability to deal with an impostor gallery is not considered in the current modular framework. Moreover, a module to deal with score normalization may further improve the recognition performance of *ModBioS*. As discussed in the last section, the best score normalization may be different depending on the user, justifying the development of this new module. The current *Combiner* performs an exhaustive search and this requires a considerable amount of computer resources. Thus, the application of other search methods, such as by the use of evolutionary computing (CASTRO, 2006), could improve this aspect of the modular system. The application of proposals from *Metalearning* is another possibility to enhance the choice of the modules per user (LEMKE; BUDKA; GABRYS, 2015; BRAZDIL *et al.*, 2009). Another aspect for further research is to study whether the adaptation strategies should be changed over time depending on the user. Altogether, this new modular adaptive biometric system opens a myriad of research lines for future work.

CONCLUSION

Biometrics has been presented as a suitable alternative to enhance the security of user authentication, in order to face the weaknesses of password-based or card-based systems. Biometric features should meet several requirements such as universality, distinctiveness, collectability and permanence (JAIN; ROSS; PRABHAKAR, 2004). Permanence establishes that the features will not change over time. However, recent studies have shown that it is not the case in various cases. In fact, features extracted from several biometric traits may change over time (ROLI; DIDACI; MARCIALIS, 2008). Consequently, the biometric reference may not reflect the current user data anymore, an issue known as *template ageing* (JAIN; NANDAKUMAR; ROSS, 2016). As a result, the recognition performance of the biometric system may be affected.

Such a problem raises the need to update the biometric reference. A simple way to solve this problem is to periodically perform enrollment sessions for all registered users. Nevertheless, this can be unfeasible due to operational costs and user annoyance. Another possibility is the use of an adaptive biometric system, which is able to automatically adapt the biometric reference over time (ROLI; DIDACI; MARCIALIS, 2008; POH; RATTANI; ROLI, 2012). Adaptive biometric systems are the focus of this thesis.

This thesis studied adaptive biometric systems considering *biometrics in a data stream context*. In such a context, the biometric reference is obtained using the first/oldest samples of the user for enrollment, while the test is done on the remaining/newest samples. The test assumes a data stream context, in which the query samples are presented one after another in chronological order. The biometric system then has to classify each query and, at the same time, adapt the biometric reference. The true label of the queries is never informed to the biometric system during the test and the queries are not divided into sessions either. Biometrics in a data stream context proposes to simulate a practical scenario, where constraints, such as respecting chronological order, no information about the label of queries and no division into sessions are likely to be present.

Among the biometric modalities, this thesis focused on the behavioural ones, particularly on keystroke dynamics and accelerometer-based gait biometrics. This choice is supported by several facts:

- First, there are suitable public datasets to study these modalities under a data stream context. Datasets for this kind of study need to contain several samples per user.
- Second, behavioural modalities tend to be subject to faster changes of the biometric features over time than physical modalities, like face, for example (GIOT; ROSENBERGER; DORIZZI, 2012c).
- Moreover, behavioural modalities usually imply in lower discriminative power than physical modalities (e.g. face, fingerprint, iris), introducing additional challenges to the biometric system.
- Finally, in the current studies dealing with adaptive biometric systems, there is a gap to be filled by further research on behavioural modalities.

Throughout the thesis, several aspects to enhance the design of adaptive biometric systems for behavioural modalities in a data stream context were discussed: *Usage Control* versions for the immune-based *Self-Detector*, combinations of genuine and impostor models in the *Enhanced Template Update* framework and application of score normalization for biometrics in a data stream context. Each of these chapters dealt with a diverse aspect of adaptive biometric systems. Then, finally, the modular adaptive biometric system, *ModBioS*, was proposed, which is capable of generalizing several baselines and proposals into a single modular framework, along with the possibility of assigning different adaptation strategies per user. The main contributions and results from the research carried out for this thesis are presented in the next section.

8.1 Contributions and results

The first contribution of this thesis is the evaluation methodology to study *biometrics in a data stream context*. Evaluation methodologies adopted in previous work can control when adaptation is performed in different ways, mainly by the session-oriented approach (RATTANI; MARCIALIS; ROLI, 2013b). Biometrics in a data stream context considers the test as being formed by a single biometric data stream, which is presented query by query to the biometric system. The adaptive biometric system has to classify each query and decide when adaptation should be performed, without the help of session information. Moreover, several previous methodologies use different sets of samples for test and adaptation processes (ULUDAG; ROSS; JAIN, 2004; RATTANI; MARCIALIS; ROLI, 2013b). However, biometrics in a data stream context uses the same set of samples, the biometric data stream, for both processes, which is closer to a practical application scenario. Such an approach has also the advantage of making a more efficient use of the dataset, since all test samples are used for test and adaptation, in contrast to methodologies which dedicate disjoint parts of the dataset for each process. Fur-

thermore, some previous methodologies provide the true label to the biometric system, though it may not be feasible in a practical scenario (ULUDAG; ROSS; JAIN, 2004). A similar approach is adopted by several studies in Data Stream Mining, which provide the true label some time after the classification decision (ŽLIOBAITĖ *et al.*, 2015). This additional information is then used to help the adaptation process.

A number of papers report the results in terms of Equal Error Rate (EER) (KANG; HWANG; CHO, 2007; FRENI; MARCIALIS; ROLI, 2008; MARCIALIS; RATTANI; ROLI, 2008; RATTANI; MARCIALIS; ROLI, 2011b), which is obtained by adjusting the parameters (e.g. decision threshold) using the test data. The threshold can vary among different sessions too (GIOT; DORIZZI; ROSENBERGER, 2011). Nevertheless, the EER parameter adjustment is not possible in a practical scenario due to the use of the test data. A more realistic approach is to tune the parameters using the enrollment data only and then using the tuned parameters in the test, similar to (BENGIO; MARIÉTHOZ; KELLER, 2005). Consequently, the performance is not reported in terms of EER, but instead, it is reported in terms of False Match Rate (FMR) and False Non-Match Rate (FNMR) for the obtained parameters values. The research carried out for this thesis also reports the performance over time in terms of FMR and FNMR for several biometric systems. It is usually expected that an adaptive biometric system reduces FNMR compared to a non-adaptive biometric system, while the adaptive system does not increase FMR compared to the baseline non-adaptive system. In addition, this thesis applied the Bayesian Hierarchical statistical test (CORANI *et al.*, 2016; BENAVALI *et al.*, 2016) over the reported results, which overcomes some drawbacks of the null hypothesis significance tests. The Bayesian test is capable of reporting the probability that the performance of a system is better than that of another given the observed experimental results.

After the discussion on the evaluation methodology, two main proposals regarding the adaptation were presented in this thesis: *Usage Control* and *Enhanced Template Update (ETU)*. *Self-Detector* is an immune-based classification algorithm which obtained good recognition performance for keystroke dynamics in a non-adaptive scenario (PISANI, 2012; PISANI; LORENA, 2015). The research carried out for this thesis then proposed four adaptation strategies for this classification algorithm. They are based on the concept of controlling the usage of the detectors for matching, by keeping only the most used ones. According to the experimental results, the *Usage Control* versions can be considered suitable adaptation strategies for *Self-Detector*. In the same chapter, the correlation over time plot was introduced. This plot showed the correlation between the closest detector and each genuine sample in the test biometric data stream. Adaptive biometric systems tend to keep this correlation more stable and at a higher level over time, while this correlation can decrease over time for non-adaptive biometric systems. From this plot, it was also reported that different users change this correlation in different ways over time. This indicates that each user has its particular *change pattern*, which suggests that different adaptation strategies should be chosen per user. This is further investigated using the modular adaptive biometric system proposed in this research.

Another proposal to improve current adaptation strategies is the *Enhanced Template Update*. Most adaptation strategies implement adaptation by only adapting a genuine model. To the best of the author's knowledge, there is no prior work which adapts a genuine and an impostor model for each user. It was implemented by *ETU*, which considers all queries for adaptation and manages a genuine and an impostor model. Four *ETU* versions were proposed: *ETU 0* to *3*. They combine both user models in different ways to improve the recognition performance. In addition, the *Positive Gallery Protection* (PGP) was proposed within the *ETU* framework. *PGP* aims to reduce the wrong inclusion of impostor samples in the genuine gallery by the use of the additional information provided by an impostor gallery. According to the reported results, *ETU 2* and *3* have competitive performance, though they did not obtain the best performance in several datasets. This illustrates the need for further investigation on the combination of genuine and impostor models.

Apart from the adaptation strategies *Usage Control* and *ETU*, the thesis also studied the application of score normalization for biometrics in a data stream context, another important aspect of an adaptive biometric system. The correct choice of the samples to be used to adapt the biometric reference is key to the final performance of the system. As previously discussed, most adaptation strategies manage a single genuine user model and, therefore, the mistaken inclusion of impostor samples in this set should be avoided. In order to deal with this issue, several adaptation strategies only use queries classified as genuine with higher confidence in the adaptation process. Another way to improve this situation is by the improvement of the classification decision process to make the output more reliable. A more reliable output from the classification algorithm can result in more genuine samples used for adaptation, while avoiding the wrong usage of impostor samples during the adaptation process of the genuine user model. In biometrics, score normalization has been used as a way of refining the classification decision (POH; MERATI; KITTLER, 2009). A preliminary study on the use of score normalization for supervised adaptation to cope with different acquisition conditions has been conducted in (POH *et al.*, 2010).

However, to the best of the author's knowledge, score normalization has not been used for biometrics in a data stream context. The research for this thesis investigated the application of four score normalization procedures in this context. Since it is not a straightforward task, the thesis discussed aspects on how to obtain the normalization terms in such a dynamic context. The results were discussed in a series of questions. First, according to the experimental results, score normalization can actually improve the recognition of adaptive biometric systems. Overall, T-Norm and Adaptive F-Norm attained the best performance among the assessed score normalization procedures. It may be due to the adaptation of the normalization terms over time for these two score normalization procedures, owing to how the cohort models were implemented here. Second, it was observed that each biometric system benefits in different degrees from the score normalization. In light of this result, the thesis compared which aspect implies in higher performance improvement for a biometric system: adaptation or score normalization.

According to the results, adaptation can have a higher impact on the performance than score normalization. Moreover, it was also found that the combined use of adaptation and score normalization can result in even higher recognition performance than any of them applied alone. Finally, the experiments on score normalization have reported that the best score normalization can be different depending on the user, even in the same dataset. It suggests that the score normalization should be chosen per user, rather than using a common score normalization for all users in the biometric system.

As discussed in this section, the research for this thesis has contributed for several aspects of adaptive biometric systems. However, for each of them, it was observed that the best solution for a dataset may not be the best solution for another dataset. This claim is supported by the fact that the best biometric system is not the same for all datasets. This is related to the *no free lunch theorem* (LUXBURG; SCHÖLKOPF, 2011). The same applies to the different score normalization procedures. Moreover, in the study of each aspect it was also suggested that the best solution for one user may not be the best solution for another one. This is supported by several findings throughout the thesis. The genuine correlation plot in Chapter 4 showed that different users can have different *change patterns*, as suggested by (PISANI; LORENA; CARVALHO, 2015b). Apart from that, the plot reporting the performance of each score normalization per user in Chapter 6 also illustrated that the normalization should be chosen per user. Furthermore, it was also reported by another author that the performance of each user can change in different ways over time (POH *et al.*, 2015). Another study also suggested that different groups of users may need modifications in the adaptation strategy and parameters (RATTANI; MARCIALIS; ROLI, 2009). In view of all of these findings, the research for this thesis proposed the *modular adaptive biometric system (ModBioS)*, which is able to assign a different adaptation strategy for each user. To the best of the author's knowledge, there is no previous biometric system with this capability.

ModBioS introduces a modular framework, in which each aspect an adaptive biometric system is divided into modules. The behaviour of this new modular biometric system can be changed by choosing different combinations of module implementations. The proposed system can generalize several current adaptive biometric systems proposed in the literature (Section 2.2) as well as all *Usage Control* versions presented in Chapter 4.

According to the experimental results, the best combination of modules can be different among datasets and among users in the same dataset. To the best of the author's knowledge, this is the first study that experimentally illustrated the performance of different *adaptation strategies* to investigate whether it implies that different strategies should be chosen per user. Some methods to choose the modules per user using only the enrollment samples were proposed here. The results have shown that the performance of *ModBioS* is higher than that of most baseline biometric systems. In theory, *ModBioS* should obtain the same or the best overall performance than that of all generalized baselines. However, *ModBioS* could not reach the best performance in all datasets. This is due to a sub-optimal choice of the modules, since the choice is made

using just the enrollment samples. A further investigation in Chapter 7 showed that the change pattern can differ between the enrollment and test samples. Even though, *ModBioS* under this sub-optimal choice of modules attain higher recognition performance than most baseline biometric systems. There is a number of opportunities with the modular framework for future work. They are presented in Section 8.4, along with possible further work for the other aspects of an adaptive biometric system studied throughout this thesis.

8.2 Publications

Several papers have been published during the development of the research for this thesis. Hence, part of the results reported throughout this thesis can be found on these publications. The papers that were published and are related to this thesis are presented next:

- [Conference] PISANI, P. H.; LORENA, A. C.; CARVALHO, A. C. P. L. F. Algoritmos imunológicos adaptativos em dinâmica da digitação: um contexto de fluxo de dados. In: Anais do X Encontro Nacional de Inteligência Artificial e Computacional - ENIAC. p. 1–12, 2013.
- [Journal] PISANI, P. H.; LORENA, A. C. A systematic review on keystroke dynamics. Journal of the Brazilian Computer Society, Springer, v. 19, n. 4, p. 573–587, 2013. ISSN 0104-6500. DOI: 10.1007/s13173-013-0117-7
- [Conference] PISANI, P. H.; LORENA, A. C.; CARVALHO, A. C. P. L. F. Adaptive algorithms in accelerometer biometrics. In: 2014 Brazilian Conference on Intelligent Systems (BRACIS). p. 336–341, IEEE, 2014. DOI: 10.1109/BRACIS.2014.67
- [Journal] PISANI, P. H.; LORENA, A. C.; CARVALHO, A. C. P. L. F. Adaptive positive selection for keystroke dynamics. Journal of Intelligent & Robotic Systems, Springer, v. 80, n. 1, p. 277–293, 2015. ISSN 1573-0409. DOI: 10.1007/s10846-014-0148-0
- [Conference] PISANI, P. H.; LORENA, A. C.; CARVALHO, A. C. P. L. F. Adaptive approaches for keystroke dynamics. In: 2015 International Joint Conference on Neural Networks (IJCNN). p. 1–8, IEEE, 2015. DOI: 10.1109/IJCNN.2015.7280467
- [Conference] PISANI, P. H.; LORENA, A. C.; CARVALHO, A. C. P. L. F. Ensemble of adaptive algorithms for keystroke dynamics. In: 2015 Brazilian Conference on Intelligent Systems (BRACIS). p. 310–315, IEEE, 2015. DOI: 10.1109/BRACIS.2015.29
- [Journal] PISANI, P. H.; GIOT, R.; CARVALHO, A. C. P. L. F.; LORENA, A. C. Enhanced template update: Application to keystroke dynamics. Computers & Security, Elsevier, v. 60, p. 134–153, 2016. ISSN 0167-4048. DOI: 10.1016/j.cose.2016.04.004
- [Journal] PISANI, P. H.; LORENA, A. C.; CARVALHO, A. C. P. L. F. Adaptive algorithms applied to accelerometer biometrics in a data stream context. Intelligent Data Analysis, IOS Press, v. 21, n. 2, p. 353–370, 2017. ISSN 1088-467X. DOI: 10.3233/IDA-150403

In addition, the following papers were submitted and are currently under review (the year refers to the submission date):

- [Journal] PISANI, P. H.; POH, N.; CARVALHO, A. C. P. L. F.; LORENA, A. C. Score Normalization applied to Adaptive Biometric Systems. 2016.
- [Journal] PISANI, P. H.; GIOT, R.; LORENA, A. C.; CARVALHO, A. C. P. L. F. Modular Adaptive Biometric System. 2017.

8.3 Limitations

The results obtained for the seven datasets studied here may differ if a different set of datasets is used, although it is expected that the same tendencies observed here remain. Studying different biometric modalities, like the physical biometrics (e.g. fingerprint, face, iris), can also result in diverse results. The reported results are also dependent on the data processing/feature extraction procedures adopted in this work. Furthermore, the distance computation adopted in this thesis (e.g. cosine distance) can impact the results, since other distance measures could result in different results. The same applies to the adopted evaluation methodology, since distinct methodologies can obtain other results. Nonetheless, it is expected that the same tendencies and conclusions obtained in this thesis hold for other possible methodologies that consider biometrics in a data stream context. It is important to highlight that there is a limited amount of datasets suitable to evaluate adaptive biometric systems, as previously discussed. In fact, several studies in the area assessed the performance on a single dataset (DIDACI; MARCIALIS; ROLI, 2014; KANG; HWANG; CHO, 2007; ULUDAG; ROSS; JAIN, 2004). All datasets used in this thesis are publicly available and can be used to reproduce the reported experiments. The decisions made to define the adopted evaluation methodology are also properly justified in Chapter 3, which describes it in detail to allow the application of the methodology by other researchers.

Although a number of adaptation strategies were considered in the experiments, just a few classification algorithms were assessed: *Self-Detector*, *M2005* and *OCSVM*. Some of the investigated adaptation strategies can be applied to other classification algorithms (e.g. neural networks), which could result in different performance. Nevertheless, the studied adaptation strategies were mainly designed to deal with one-class classification algorithms. Even though, the choice of the classification algorithms, as described in Chapter 3, was based on what has been used in the literature for the biometric modalities in the context of adaptive biometric systems. This limitation is due to the thesis focus on the adaptation process. Such a limitation is also observed in other studies in the area (MHENNI *et al.*, 2016; GIOT; ROSENBERGER; DORIZZI, 2012b; RATTANI; MARCIALIS; ROLI, 2011b; KANG; HWANG; CHO, 2007). It shows an opportunity for future work in the area, to investigate to what extent different classification algorithms can impair the performance of the adaptation strategies. Note that the current *ModBioS* implementation already contains a module that allows the selection of a different classification algorithm for each user.

The thesis did not discuss questions regarding the computational time of adaptive biometric systems. This can be particularly important for the modular system, which performs an exhaustive search over thousands of module combinations. However, this mainly affects the *Combiner* and not the modular framework, which is one of the main contributions of *ModBioS*. More efficient approaches to implement the *Combiner* is a topic for future work. Issues concerning the memory usage were also not deeply studied in this work. Nonetheless, studying these questions regarding the consumption of computer resources (processing and memory) would not result in negative conclusions regarding the usage of adaptation. As observed during the execution of the experiments, most adaptive systems are actually lightweight. The main exception is the current implementation of *Combiner* used by the modular adaptive biometric system. The next section describes several lines for future research, including some alternatives to overcome this limitation.

8.4 Future work

From the research conducted in this thesis, a number of opportunities for future work were opened. Some of them are presented in this section.

This work studied only single modality biometric systems, however, the application and extension of the proposals of this thesis in a multiple modality context could obtain good results. Previous studies on *Co-Update*, which deals with two modalities to improve the adaptation process, have shown that it can obtain good results (ROLI; DIDACI; MARCIALIS, 2007; RATTANI; MARCIALIS; ROLI, 2011b). The combination of the outputs from two biometric modalities can also improve the classification confidence of query samples. It is important to note, however, that the use of two modalities in a biometric system is not feasible in some cases, since it usually represents an additional cost over a single modality biometric system.

In the evaluation methodology adopted here, the set of registered users does not change over time. However, in practical applications, the system can start with a given set of users, which can be changed later by removing or adding registered users. For example, in a company, a new employee could be hired several months after the implementation of a biometric system. This new employee would then be a new registered user in the system. The study of the effect of these changes is a topic for future research. It can also be studied under the context of the Doddington's Zoo phenomenon (DODDINGTON *et al.*, 1998) discussed in Chapter 3. For instance, the introduction of a lamb, which is more likely to have its biometric reference changed by an impostor, can result in performance degradation. Issues concerning the Doddington's Zoo phenomenon can impact all adaptation strategies, particularly ETU, which attempts to model the impostors too. A further investigation of the Doddington's Zoo phenomenon for biometrics in a data stream context is another opportunity for future work.

Another important issue for future work is a proper comparison between physical and behavioural biometric modalities in adaptive biometric systems. This new study should be

conducted along with the question of adaptation to changes due to *time* and *condition* formalized in (POH; RATTANI; ROLI, 2012). As discussed in Chapter 2, some studies which deal with physical modalities seem to mainly adapt the biometric reference to different *conditions* (POH; KITTLER; RATTANI, 2014), while others in the area of behavioural biometrics are mainly concerned with adaptation due to *time* (GIOT; ROSENBERGER; DORIZZI, 2012b). In a face recognition system, for instance, the enrollment could have been performed with samples at a given pose. However, during the test, the queries may contain the genuine user at diverse poses and, therefore, the biometric reference should be adapted to include these new poses. Note that the older poses should not be removed in this case, since the system is just adding new *conditions*. Nevertheless, if the face undergoes changes due to time (e.g. due to ageing), the biometric reference should be adapted to these changes. In this case, the older patterns must be removed from the biometric reference, since they do not represent the current biometric features of the genuine user. To the best of the author's knowledge, there is no adaptive biometric system capable of distinguishing time from condition changes and, therefore, apply different adaptation strategies for each case.

In the experiments, it was observed that some classification algorithms are better at reducing FMR, while others at reducing FNMR. In Machine learning, previous studies have shown that the combination of individual techniques in ensembles can produce more accurate and stable decision models (LUMINI; NANNI, 2009; TORRE *et al.*, 2015). However, to the best of the author's knowledge, there is no prior work using ensemble of one class classifiers for keystroke dynamics in a data stream context. An initial study on ensembles of one-class classifiers for keystroke dynamics was conducted by the research for this thesis (PISANI; LORENA; CARVALHO, 2015c), where two strategies for combining the base adaptive models were employed: voting and stacking. An extended version of this study was conducted later, answering some questions raised after the initial study. Even though, there are still several questions to be explored regarding the application of ensembles in adaptive biometric systems.

The *Enhanced Template Update* framework used only the queries as a source of information to manage the genuine and impostor models. In future research, this framework can be extended to consider a database-aware approach, where the biometric system is able to access data from all registered users. The usage of this additional source of information can improve the reliability of the impostor model, since the data from the users different from the reference can be understood as impostor data.

According to the experimental results, score normalization improved the performance of adaptive biometric systems. Nonetheless, some normalization procedures did not update the normalization terms over time, like F-Norm, for example. By adapting these terms, it is expected that their performance can be increased. However, it is still an open question on how to update them for F-Norm. A possible solution is to use the classified queries to adapt them, although mistakes can be made since the true label is not available. Furthermore, a method to apply score normalization to the impostor model of *ETU* can enhance its performance. The

studied score normalization procedures were designed to be applied on the genuine models. Thus, the implications of their usage to normalize the scores from the impostor model need to be investigated.

Moreover, the application of score normalization in *ModBioS* can be studied in future work. It would be possible by the inclusion of a new score normalization module, making it possible to choose different score normalization procedures for each user. As reported in Chapter 6, the best score normalization can differ among users in the same dataset. In some cases, a user obtained higher balanced accuracy without any score normalization processing. The implementation of such a module is not a trivial task since the score normalization would have to consider the translation between score indexes and thresholds adopted by *ModBioS*. This translation exists to enable *ModBioS* to deal with different classification algorithms, which can output scores at different ranges of values.

Still regarding *ModBioS*, additional implementations of the current modules can be developed in future work. This would result in a higher number of combinations and can increase the amount of adaptive biometric systems generalized by *ModBioS*. Moreover, apart from the new module for score normalization discussed in the last paragraph, the framework can be extended in order to support modules that deal with an impostor model, similarly to *ETU*. This would allow *ModBioS* to generalize *ETU*. An additional extension for the *ModBioS* would be the support for various ensemble configurations (KUNCHEVA, 2014), which could improve the final classification decision.

Another important topic for future research on the modular system is the improvement of the *Combiner*, which chooses the combination of modules per user. The current version managed to obtain combinations that attain higher performance than most baselines. Even though, the choice can be improved. According to the discussion in Chapter 7, the user characteristics, including the *change pattern*, in the enrollment samples may be different from the characteristics in the test data. In view of this fact, additional sources of information should be used to improve the choice of modules. The information obtained over time by the biometric system is a possible source to improve the choice of the combinations. Apart from that, the current *Combiner* performs an exhaustive search and this requires a considerable amount of computer resources, so the application of other search methods, such as by the use of evolutionary computing (CASTRO, 2006), could improve this aspect of the modular system. The application of proposals from *Metalearning* is another possibility to enhance the choice of the modules per user (LEMKE; BUDKA; GABRYS, 2015; BRAZDIL *et al.*, 2009). Another aspect for further research is the study of whether the adaptation strategies should be changed over time depending on the user. It may happen that, not only is a different adaptation needed per user, but it also should be modified over time. All in all, this new modular adaptive biometric system opens a myriad of research lines for future work.

REFERENCES

- AGGARWAL, C. C.; HAN, J.; WANG, J.; YU, P. S. A framework for clustering evolving data streams. In: **Proceedings of the 29th International Conference on Very Large Data Bases - Volume 29**. [S.l.]: VLDB Endowment, 2003. (VLDB '03), p. 81–92. ISBN 0-12-722442-4. (Cited on pages 26 and 56)
- ARTHUR, D.; VASSILVITSKII, S. K-means++: The advantages of careful seeding. In: **Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms**. [S.l.]: Society for Industrial and Applied Mathematics, 2007. (SODA '07), p. 1027–1035. ISBN 978-0-898716-24-5. (Cited on pages 97 and 100)
- AUCKENTHALER, R.; CAREY, M.; LLOYD-THOMAS, H. Score normalization for text-independent speaker verification systems. **Digital Signal Processing**, v. 10, n. 1–3, p. 42–54, 2000. ISSN 1051-2004. (Cited on page 114)
- AYTAR, Y.; ZISSERMAN, A. Tabula rasa: Model transfer for object category detection. In: **2011 International Conference on Computer Vision**. [S.l.]: IEEE, 2011. p. 2252–2259. ISSN 1550-5499. (Cited on page 40)
- BAILLY-BAILLIÈRE, E.; BENGIO, S.; BIMBOT, F.; HAMOUZ, M.; KITTLER, J.; MARIÉTHOZ, J.; MATAS, J.; MESSER, K.; POPOVICI, V.; PORÉE, F.; RUIZ, B.; THIRAN, J.-P. The banca database and evaluation protocol. In: **Proceedings of the 4th international conference on Audio- and video-based biometric person authentication**. [S.l.]: Springer-Verlag, 2003. (AVBPA'03), p. 625–638. ISBN 3-540-40302-7. (Cited on pages 45 and 52)
- BENAVOLI, A.; CORANI, G.; DEMŠAR, J.; ZAFFALON, M. Time for a change: a tutorial for comparing multiple classifiers through bayesian analysis. **CoRR**, abs/1606.04316, 2016. Available at: <<http://arxiv.org/abs/1606.04316>>. (Cited on pages 72, 73, 86, 104, 111, 120, 124, 146, 151, and 161)
- BENAVOLI, A.; CORANI, G.; MANGILI, F. Should we really use post-hoc tests based on mean-ranks? **Journal of Machine Learning Research**, JMLR.org, v. 17, n. 5, p. 1–10, 2016. (Cited on page 73)
- BENGIO, S.; MARIÉTHOZ, J.; KELLER, M. The expected performance curve. In: **International Conference on Machine Learning (ICML), Workshop on ROC Analysis in Machine Learning**. [S.l.: s.n.], 2005. p. 1–8. (Cited on pages 51, 62, and 161)
- BENGIO, S.; MARIÉTHOZ, J. Biometric person authentication is a multiple classifier problem. In: HAINDL, M.; KITTLER, J.; ROLI, F. (Ed.). **Multiple Classifier Systems**. [S.l.]: Springer Berlin Heidelberg, 2007, (Lecture Notes in Computer Science, v. 4472). p. 513–522. ISBN 978-3-540-72481-0. (Cited on page 61)
- BIFET, A.; HOLMES, G.; PFAHRINGER, B.; KIRKBY, R.; GAVALDÀ, R. New ensemble methods for evolving data streams. In: **Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. [S.l.]: ACM, 2009. (KDD '09), p. 139–148. ISBN 978-1-60558-495-9. (Cited on pages 26 and 56)
- BIGGIO, B.; FUMERA, G.; ROLI, F.; DIDACI, L. Poisoning adaptive biometric systems. In: **Proceedings of the 2012 Joint IAPR international conference on Structural, Syntactic, and**

- Statistical Pattern Recognition.** [S.l.]: Springer-Verlag, 2012. (SSPR'12/SPR'12), p. 417–425. ISBN 978-3-642-34165-6. (Cited on pages 36, 45, 46, and 52)
- BISHOP, C. **Pattern Recognition and Machine Learning.** [S.l.]: Springer, 2006. ISBN 978-0-387-31073-2. (Cited on page 97)
- BLUM, A.; CHAWLA, S. Learning from labeled and unlabeled data using graph mincuts. In: **Proceedings of the Eighteenth International Conference on Machine Learning.** [S.l.]: Morgan Kaufmann Publishers Inc., 2001. (ICML '01), p. 19–26. ISBN 1-55860-778-1. (Cited on page 38)
- BLUM, A.; MITCHELL, T. Combining labeled and unlabeled data with co-training. In: **Proceedings of the Eleventh Annual Conference on Computational Learning Theory.** [S.l.]: ACM, 1998. (COLT' 98), p. 92–100. ISBN 1-58113-057-0. (Cited on page 38)
- BRAZDIL, P.; CARRIER, C. G.; SOARES, C.; VILALTA, R. **Metalearning.** [S.l.]: Springer, 2009. ISBN 978-3-540-73263-1. (Cited on pages 158 and 168)
- CASTRO, L. N. de. **Fundamentals of Natural Computing.** [S.l.]: Chapman & Hall/CRC, 2006. ISBN 1-58488-643-9. (Cited on pages 158 and 168)
- CASTRO, L. N. de; TIMMIS, J. **Artificial Immune Systems: A New Computational Intelligence Approach.** [S.l.]: Springer, 2002. ISBN 1-85223-594-7. (Cited on page 76)
- ÇEKER, H.; UPADHYAYA, S. Adaptive techniques for intra-user variability in keystroke dynamics. In: **2016 IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS).** [S.l.]: IEEE, 2016. p. 1–6. (Cited on pages 40 and 45)
- CHANG, C.-C.; LIN, C.-J. LIBSVM: A library for support vector machines. **ACM Transactions on Intelligent Systems and Technology**, v. 2, p. 27:1–27:27, 2011. Software available at <<http://www.csie.ntu.edu.tw/~cjlin/libsvm>>. (Cited on page 72)
- COMMONS Math: The Apache Commons Mathematics Library (3.3). 2014. Available at: <<http://commons.apache.org/proper/commons-math/>>. (Cited on page 97)
- CORANI, G.; BENAVALI, A.; DEMŠAR, J.; MANGILI, F.; ZAFFALON, M. **Statistical comparison of classifiers through Bayesian hierarchical modelling.** [S.l.], 2016. (Cited on pages 72, 73, 86, 104, 111, 120, 124, 146, 151, and 161)
- DEITEL, H. M.; DEITEL, P. J. **Java How to Program.** 4th. ed. [S.l.]: Prentice Hall PTR, 2001. ISBN 0130341517. (Cited on page 32)
- DEMŠAR, J. Statistical comparisons of classifiers over multiple data sets. **Journal of Machine Learning Research**, JMLR.org, v. 7, p. 1–30, 2006. (Cited on page 72)
- _____. On the appropriateness of statistical tests in machine learning. In: **Workshop on Evaluation Methods for Machine Learning in conjunction with ICML.** [S.l.: s.n.], 2008. p. 1–4. (Cited on page 72)
- DERAWI, M.; BOURS, P. Gait and activity recognition using commercial phones. **Computers & Security**, Elsevier, v. 39, Part B, p. 137 – 144, 2013. ISSN 0167-4048. (Cited on page 66)
- DIDACI, L.; MARCIALIS, G. L.; ROLI, F. Analysis of unsupervised template update in biometric recognition systems. **Pattern Recognition Letters**, v. 37, p. 151 – 160, 2014. ISSN 0167-8655. Partially Supervised Learning for Pattern Recognition. (Cited on pages 37 and 165)
- DODDINGTON, G.; LIGGETT, W.; MARTIN, A.; PRZYBOCKI, M.; REYNOLDS, D. Sheep, goats, lambs and wolves a statistical analysis of speaker performance in the nist 1998 speaker recognition evaluation. In: **International Conference on Spoken Language Processing.** [S.l.: s.n.], 1998. (Cited on pages 60 and 166)

- DRUMMOND, C. Machine learning as an experimental science (revisited). In: **Proceedings of the AAAI 2006 Workshop on Evaluation Methods for Machine Learning**. [S.l.: s.n.], 2006. p. 1–5. (Cited on page 72)
- DUSERICK, W. **Whitepaper on Liberty Protocol and Identity Theft**. [S.l.]: Liberty Alliance Project, 2004. (Cited on page 25)
- FENKER, S.; BOWYER, K. Analysis of template aging in iris biometrics. In: **2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)**. [S.l.]: IEEE, 2012. p. 45–51. ISSN 2160-7508. (Cited on pages 52 and 62)
- FLACH, P. **Machine Learning: The Art and Science of Algorithms That Make Sense of Data**. [S.l.]: Cambridge University Press, 2012. ISBN 978-1-107-42222-3. (Cited on pages 57 and 100)
- FRANK, J.; MANNOR, S.; PRECUP, D. **Data Sets: Mobile Phone Gait Recognition Data**. 2010. Available at: <<http://www.cs.mcgill.ca/~jfrank8/data/gait-dataset.html>>. (Cited on pages 66 and 67)
- FRENI, B.; MARCIALIS, G.; ROLI, F. Replacement algorithms for fingerprint template update. In: CAMPILHO, A.; KAMEL, M. (Ed.). **Image Analysis and Recognition**. [S.l.]: Springer Berlin Heidelberg, 2008, (Lecture Notes in Computer Science, v. 5112). p. 884–893. ISBN 978-3-540-69811-1. (Cited on pages 41, 78, 129, and 161)
- GAMA, J. **Knowledge Discovery from Data Streams**. [S.l.]: CRC Press, 2010. ISBN 9781439826119. (Cited on page 56)
- GAMA, J.; ŽLIOBAITĖ, I.; BIFET, A.; PECHENIZKIY, M.; BOUCHACHIA, A. A survey on concept drift adaptation. **ACM Computing Surveys**, ACM, v. 46, n. 4, p. 44:1–44:37, 2014. ISSN 0360-0300. (Cited on page 26)
- GIOT, R.; DORIZZI, B.; ROSENBERGER, C. Analysis of template update strategies for keystroke dynamics. In: **2011 IEEE Workshop on Computational Intelligence in Biometrics and Identity Management (CIBIM)**. [S.l.]: IEEE, 2011. p. 21–28. (Cited on pages 35, 42, 45, 52, and 161)
- GIOT, R.; EL-ABED, M.; HEMERY, B.; ROSENBERGER, C. Unconstrained keystroke dynamics authentication with shared secret. **Computers & Security**, Elsevier, v. 30, n. 6-7, p. 427–445, 2011. ISSN 0167-4048. (Cited on page 59)
- GIOT, R.; EL-ABED, M.; ROSENBERGER, C. Greyc keystroke: a benchmark for keystroke dynamics biometric systems. In: **IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS 2009)**. [S.l.]: IEEE Computer Society, 2009. p. 419–424. (Cited on pages 52, 64, and 65)
- _____. Web-based benchmark for keystroke dynamics biometric systems: A statistical analysis. In: **2012 Eighth International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP)**. [S.l.]: IEEE, 2012. p. 11–15. (Cited on pages 26, 52, 64, and 65)
- GIOT, R.; ROSENBERGER, C.; DORIZZI, B. Can chronological information be used as a soft biometric in keystroke dynamics? In: **2012 Eighth International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP)**. [S.l.]: IEEE, 2012. p. 7–10. (Cited on page 45)
- _____. Hybrid template update system for unimodal biometric systems. In: **2012 IEEE Fifth International Conference on Biometrics: Theory, Applications and Systems (BTAS)**. [S.l.]: IEEE, 2012. p. 1–7. (Cited on pages 11, 26, 36, 42, 48, 49, 51, 60, 62, 68, 129, 133, 165, and 167)

- _____. Performance evaluation of biometric template update. In: **International Biometric Performance Testing Conference**. [S.l.: s.n.], 2012. p. 1–4. (Cited on pages 26, 27, 64, and 160)
- _____. A new protocol to evaluate the resistance of template update systems against zero-effort attacks. In: **2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)**. [S.l.]: IEEE, 2013. p. 131–137. (Cited on pages 48, 49, and 60)
- HAN, J. **Data Mining: Concepts and Techniques, 2nd edition**. [S.l.]: Morgan Kaufmann Publishers Inc./Elsevier, 2006. ISBN 1-55860-901-6. (Cited on page 56)
- HIMAGA, M.; KOU, K. Finger vein authentication technology and financial applications. In: RATHA, N.; GOVINDARAJU, V. (Ed.). **Advances in Biometrics**. [S.l.]: Springer London, 2008. p. 89–105. ISBN 978-1-84628-920-0. (Cited on page 50)
- HOCQUET, S.; RAMEL, J.-Y.; CARDOT, H. User classification for keystroke dynamics authentication. In: _____. **Proceedings of the International Conference on Advances in Biometrics (ICB)**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007. p. 531–539. ISBN 978-3-540-74549-5. (Cited on page 69)
- HOSSEINZADEH, D.; KRISHNAN, S. Gaussian mixture modeling of keystroke patterns for biometric applications. **IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews**, IEEE, v. 38, n. 6, p. 816–826, 2008. (Cited on page 64)
- INTERNATIONAL ORGANIZATION FOR STANDARDIZATION AND INTERNATIONAL ELECTROTECHNICAL COMMISSION. **ISO/IEC 19784-1**: Information technology – biometric application programming interface – part 1: Bioapi specification. [S.l.], 2006. (Cited on page 26)
- JAIN, A.; ROSS, A.; PRABHAKAR, S. An introduction to biometric recognition. **IEEE Transactions on Circuits and Systems for Video Technology**, IEEE, v. 14, n. 1, p. 4–20, 2004. ISSN 1051-8215. (Cited on pages 25, 31, 32, 34, and 159)
- JAIN, A. K.; NANDAKUMAR, K.; ROSS, A. 50 years of biometric research: Accomplishments, challenges, and opportunities. **Pattern Recognition Letters**, Elsevier, v. 79, p. 80 – 105, 2016. ISSN 0167-8655. (Cited on pages 25, 26, 31, 33, 34, and 159)
- KAGGLE. **Accelerometer Biometric Competition**. 2013. Available at: <<https://www.kaggle.com/c/accelerometer-biometric-competition>>. (Cited on pages 27, 63, and 66)
- KANG, P.; HWANG, S.-s.; CHO, S. Continual retraining of keystroke dynamics based authenticator. In: LEE, S.-W.; LI, S. (Ed.). **Advances in Biometrics**. [S.l.]: Springer Berlin / Heidelberg, 2007, (Lecture Notes in Computer Science, v. 4642). p. 1203–1211. ISBN 978-3-540-74548-8. (Cited on pages 40, 42, 44, 45, 68, 77, 129, 161, and 165)
- KILLOURHY, K.; MAXION, R. Why did my detector do that?! predicting keystroke-dynamics error rates. In: JHA, S.; SOMMER, R.; KREIBICH, C. (Ed.). **Recent Advances in Intrusion Detection**. [S.l.]: Springer Berlin / Heidelberg, 2010, (Lecture Notes in Computer Science, v. 6307). p. 256–276. ISBN 978-3-642-15511-6. (Cited on pages 52, 64, and 65)
- KRUENGKRAI, C.; JARUSKULCHAI, C. Using one-class SVMs for relevant sentence extraction. In: **Proceedings of the 3rd International Symposium on Communications and Information Technologies (ISCIT-2003)**. [S.l.: s.n.], 2003. (Cited on page 71)
- KUNCHEVA, L. I. **Combining Pattern Classifiers: Methods and Algorithms**. Second edition. [S.l.]: John Wiley & Sons, Inc., 2014. ISBN 978-1-118-31523-1. (Cited on page 168)
- KWAPISZ, J.; WEISS, G.; MOORE, S. Cell phone-based biometric identification. In: **2010 Fourth IEEE International Conference on Biometrics: Theory Applications and Systems (BTAS)**. [S.l.]: IEEE, 2010. p. 1–7. (Cited on page 66)

- KWAPISZ, J. R.; WEISS, G. M.; MOORE, S. A. Activity recognition using cell phone accelerometers. **ACM SIGKDD Explorations Newsletter**, ACM, v. 12, n. 2, p. 74–82, 2011. ISSN 1931-0145. (Cited on pages 66 and 67)
- LEE, H. joo; CHO, S. Retraining a keystroke dynamics-based authenticator with impostor patterns. **Computers & Security**, Elsevier, v. 26, n. 4, p. 300–310, 2007. (Cited on page 96)
- LEMKE, C.; BUDKA, M.; GABRYS, B. Metalearning: a survey of trends and technologies. **Artificial Intelligence Review**, Springer, v. 44, n. 1, p. 117–130, 2015. ISSN 1573-7462. (Cited on pages 158 and 168)
- LI, F.; WECHSLER, H. Open set face recognition using transduction. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, IEEE, v. 27, n. 11, p. 1686–1697, 2005. ISSN 0162-8828. (Cited on pages 57 and 60)
- LOCKHART, J. W.; WEISS, G. M.; XUE, J. C.; GALLAGHER, S. T.; GROSNER, A. B.; PULICKAL, T. T. Design considerations for the wisdm smart phone-based sensor mining architecture. In: **Proceedings of the Fifth International Workshop on Knowledge Discovery from Sensor Data**. [S.l.]: ACM, 2011. (SensorKDD '11), p. 25–33. ISBN 978-1-4503-0832-8. (Cited on page 67)
- LUMINI, A.; NANNI, L. Ensemble of on-line signature matchers based on overcomplete feature generation. **Expert Systems with Applications**, Elsevier, v. 36, n. 3, Part 1, p. 5291 – 5296, 2009. ISSN 0957-4174. (Cited on page 167)
- LUXBURG, U. von; SCHÖLKOPF, B. Statistical learning theory: Models, concepts, and results. In: GABBAY, D. M.; HARTMANN, S.; WOODS, J. (Ed.). **Inductive Logic**. [S.l.]: Elsevier, 2011, (Handbook of the History of Logic, v. 10). p. 651–706. ISBN 978-0-444-52936-7. (Cited on pages 129 and 163)
- M., A. S.; K., A.; RAJENDRAN, R.; MOHAN, A.; S., A. P.; K., M. S.; AZIZ, F. Efficient on-line and offline template update mechanisms for speaker recognition. **Computers & Electrical Engineering**, Elsevier, v. 50, p. 10 – 25, 2016. ISSN 0045-7906. (Cited on page 52)
- MAGALHÃES, S. T.; REVETT, K.; SANTOS, H. M. D. Password secured sites - stepping forward with keystroke dynamics. In: **Proceedings of the International Conference on Next Generation Web Services Practices**. [S.l.]: IEEE Computer Society, 2005. (NWESP '05), p. 293–298. ISBN 0-7695-2452-4. (Cited on pages 68 and 70)
- MAIO, D.; MALTONI, D.; CAPPELLI, R.; WAYMAN, J. L.; JAIN, A. K. Fvc2002: Second fingerprint verification competition. In: **16th International Conference on Pattern Recognition**. [S.l.]: IEEE, 2002. v. 3, p. 811–814. ISSN 1051-4651. (Cited on page 52)
- MARCIALIS, G. L.; DIDACI, L.; PISANO, A.; GRANGER, E.; ROLI, F. Why template self-update should work in biometric authentication systems? In: **2012 11th International Conference on Information Science, Signal Processing and their Applications (ISSPA)**. [S.l.]: IEEE, 2012. p. 1086–1091. (Cited on page 52)
- MARCIALIS, G. L.; RATTANI, A.; ROLI, F. Biometric template update: An experimental investigation on the relationship between update errors and performance degradation in face verification. In: **Proceedings of the 2008 Joint IAPR International Workshop on Structural, Syntactic, and Statistical Pattern Recognition**. Berlin, Heidelberg: Springer-Verlag, 2008. (SSPR & SPR '08), p. 684–693. ISBN 978-3-540-89688-3. (Cited on page 161)
- MARTINEZ, A.; BENAVENTE, R. **The AR Face Database**. [S.l.], 1998. (Cited on page 52)

- MASSO, M.; VAISMAN, I. I. Accurate and efficient gp120 v3 loop structure based models for the determination of hiv-1 co-receptor usage. **BMC Bioinformatics**, v. 11, n. 1, p. 1–11, 2010. ISSN 1471-2105. (Cited on page 51)
- MATOVSKI, D.; NIXON, M.; MAHMOODI, S.; CARTER, J. The effect of time on the performance of gait biometrics. In: **IEEE International Conference on Biometrics: Theory Applications and Systems (BTAS)**. [S.l.]: IEEE, 2010. p. 1–6. (Cited on page 66)
- MCEWAN, C.; HART, E. Representation in the (artificial) immune system. **Journal of Mathematical Modelling and Algorithms**, Springer Netherlands, v. 8, n. 2, p. 125–149, 2009. ISSN 1570-1166. (Cited on pages 27, 70, 75, 76, and 93)
- MENA-TORRES, D.; AGUILAR-RUIZ, J. S. A similarity-based approach for data stream classification. **Expert Systems with Applications**, Elsevier, v. 41, n. 9, p. 4224 – 4234, 2014. ISSN 0957-4174. (Cited on pages 27, 70, 75, 76, and 93)
- MHENNI, A.; ROSENBERGER, C.; CHERRIER, E.; AMARA, N. E. B. Keystroke template update with adapted thresholds. In: **2016 2nd International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)**. [S.l.]: IEEE, 2016. p. 483–488. (Cited on pages 36, 40, 68, 69, and 165)
- MITCHELL, T. M. **Machine Learning**. 1. ed. [S.l.]: McGraw-Hill, Inc., 1997. ISBN 0070428077, 9780070428072. (Cited on pages 27 and 75)
- MONTALVÃO, J.; FREIRE, E. O.; JR., M. A. B.; GARCIA, R. Contributions to empirical analysis of keystroke dynamics in passwords. **Pattern Recognition Letters**, Elsevier Science Inc., v. 52, n. C, p. 80–86, 2015. ISSN 0167-8655. (Cited on page 40)
- NALDI, M. C.; CAMPELLO, R. J. G. B.; HRUSCHKA, E. R.; CARVALHO, A. C. P. L. F. Efficiency issues of evolutionary k-means. **Applied Soft Computing**, Elsevier Science Publishers B. V., v. 11, n. 2, p. 1938–1952, 2011. ISSN 1568-4946. (Cited on page 97)
- NICKEL, C.; WIRTL, T.; BUSCH, C. Authentication of smartphone users based on the way they walk using k-nn algorithm. In: **2012 Eighth International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP)**. [S.l.]: IEEE, 2012. p. 16–20. (Cited on pages 66 and 69)
- PAGANO, C.; GRANGER, E.; SABOURIN, R.; TUVERI, P.; MARCIALIS, G. L.; ROLI, F. Context-sensitive self-updating for adaptive face recognition. In: _____. **Adaptive Biometric Systems: Recent Advances and Challenges**. [S.l.]: Springer International Publishing, 2015. p. 9–34. ISBN 978-3-319-24865-3. (Cited on pages 11, 47, and 48)
- PEACOCK, A.; KE, X.; WILKERSON, M. Typing patterns: a key to user identification. **IEEE Security Privacy**, IEEE, v. 2, n. 5, p. 40–47, 2004. (Cited on page 64)
- PEDREGOSA, F. Hyperparameter optimization with approximate gradient. In: **Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48**. [S.l.]: JMLR.org, 2016. (ICML'16), p. 737–746. (Cited on page 33)
- PISANI, P. H. **Algoritmos Imunológicos aplicados na Detecção de Intrusões com Dinâmica da Digitação**. Master's Thesis (Master's Thesis) — Universidade Federal do ABC, 2012. (Cited on pages 27, 53, 65, 69, 75, 93, and 161)
- PISANI, P. H.; CARVALHO, A. C. P. L. F. **Simplified function to apply the Bayesian hierarchical statistical test**. [S.l.], 2016. 1 - 7 p. (Cited on page 73)
- PISANI, P. H.; GIOT, R.; CARVALHO, A. C. P. L. F.; LORENA, A. C. Enhanced template update: Application to keystroke dynamics. **Computers & Security**, Elsevier, v. 60, p. 134 – 153, 2016. ISSN 0167-4048. (Cited on pages 63, 65, 96, 97, 98, 100, 105, 113, and 153)

PISANI, P. H.; LORENA, A. C. A systematic review on keystroke dynamics. **Journal of the Brazilian Computer Society**, Springer London, v. 19, n. 4, p. 573–587, 2013. ISSN 0104-6500. (Cited on pages 27, 63, and 65)

_____. Emphasizing typing signature in keystroke dynamics using immune algorithms. **Applied Soft Computing**, Elsevier, v. 34, p. 178 – 193, 2015. ISSN 1568-4946. (Cited on pages 27, 53, 65, 69, 70, 75, 76, 93, and 161)

PISANI, P. H.; LORENA, A. C.; CARVALHO, A. C. P. L. F. Algoritmos imunológicos adaptativos em dinâmica da digitação: um contexto de fluxo de dados. In: **Anais do X Encontro Nacional de Inteligência Artificial e Computacional - ENIAC**. [S.l.: s.n.], 2013. p. 1–12. (Cited on pages 75, 77, and 92)

_____. Adaptive algorithms in accelerometer biometrics. In: **2014 Brazilian Conference on Intelligent Systems (BRACIS)**. [S.l.]: IEEE, 2014. p. 336–341. (Cited on pages 75, 81, 113, and 129)

_____. Adaptive approaches for keystroke dynamics. In: **2015 International Joint Conference on Neural Networks (IJCNN)**. [S.l.]: IEEE, 2015. p. 1–8. ISSN 2161-4393. (Cited on pages 63, 75, 82, and 129)

_____. Adaptive positive selection for keystroke dynamics. **Journal of Intelligent & Robotic Systems**, Springer, v. 80, n. 1, p. 277–293, 2015. ISSN 1573-0409. (Cited on pages 28, 68, 75, 76, 77, 78, 80, 92, 129, 130, 141, 144, 153, 157, and 163)

_____. Ensemble of adaptive algorithms for keystroke dynamics. In: **2015 Brazilian Conference on Intelligent Systems (BRACIS)**. [S.l.]: IEEE, 2015. p. 310–315. (Cited on page 167)

_____. Adaptive algorithms applied to accelerometer biometrics in a data stream context. **Intelligent Data Analysis**, IOS Press, v. 21, n. 2, p. 353–370, 2017. ISSN 1088-467X. (Cited on page 75)

PISANI, P. H.; POH, N.; CARVALHO, A. C. P. L. F.; LORENA, A. C. Score normalization applied to adaptive biometric systems. **SUBMITTED to Computers & Security in 2016**, 2016. (Cited on pages 58, 61, 114, 116, 119, 121, 123, 125, and 191)

POH, N.; BENGIO, S. F-ratio client-dependent normalisation for biometric authentication tasks. In: **2005 IEEE International Conference on Acoustics, Speech, and Signal Processing**. [S.l.]: IEEE, 2005. v. 1, p. 721–724. ISSN 1520-6149. (Cited on page 115)

POH, N.; KITTLER, J.; CHAN, C. H.; PANDIT, M. Algorithm to estimate biometric performance change over time. **IET Biometrics**, IET, v. 4, n. 4, p. 236–245, 2015. ISSN 2047-4938. (Cited on pages 28, 129, 130, 141, 144, 157, and 163)

POH, N.; KITTLER, J.; MARCEL, S.; MATROUF, D.; BONASTRE, J.-F. Model and score adaptation for biometric systems: Coping with device interoperability and changing acquisition conditions. In: **20th International Conference on Pattern Recognition (ICPR)**. [S.l.]: IEEE, 2010. p. 1229–1232. ISSN 1051-4651. (Cited on pages 54, 113, 127, and 162)

POH, N.; KITTLER, J.; RATTANI, A. Handling session mismatch by fusion-based co-training: An empirical study using face and speech multimodal biometrics. In: **2014 IEEE Symposium on Computational Intelligence in Biometrics and Identity Management (CIBIM)**. [S.l.]: IEEE, 2014. p. 81–86. ISSN 2325-4300. (Cited on pages 11, 45, 46, 50, 85, and 167)

POH, N.; MERATI, A.; KITTLER, J. Adaptive client-impostor centric score normalization: A case study in fingerprint verification. In: **IEEE 3rd International Conference on Biometrics: Theory, Applications, and Systems**. [S.l.]: IEEE, 2009. p. 1–7. (Cited on pages 28, 53, 113, 114, 115, and 162)

- POH, N.; RATTANI, A.; ROLI, F. Critical analysis of adaptive biometric systems. **IET Biometrics**, IET, v. 1, n. 4, p. 179–187, 2012. ISSN 2047-4938. (Cited on pages 26, 32, 34, 35, 45, 46, 159, and 167)
- POH, N.; TISTARELLI, M. Customizing biometric authentication systems via discriminative score calibration. In: **2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. [S.l.]: IEEE, 2012. p. 2681–2686. ISSN 1063-6919. (Cited on page 118)
- POH, N.; WONG, R.; KITTLER, J.; ROLI, F. Challenges and research directions for adaptive biometric recognition systems. In: _____. **Proceedings of the Third International Conference on Advances in Biometrics (ICB)**. [S.l.]: Springer Berlin Heidelberg, 2009. p. 753–764. ISBN 978-3-642-01793-3. (Cited on pages 44 and 62)
- POH, N.; WONG, R.; MARCIALIS, G. L. Toward an attack-sensitive tamper-resistant biometric recognition with a symmetric matcher: A fingerprint case study. In: **2014 IEEE Symposium on Computational Intelligence in Biometrics and Identity Management (CIBIM)**. [S.l.]: IEEE, 2014. p. 175–180. ISSN 2325-4300. (Cited on page 60)
- Precise Biometrics. **Understanding Biometric Performance Evaluation**. 2014. Available at: <<http://precisebiometrics.com/wp-content/uploads/2014/11/White-Paper-Understanding-Biometric-Performance-Evaluation.pdf>>. (Cited on page 50)
- PREECE, S. J.; GOULERMAS, J. Y.; KENNEY, L. P. J.; HOWARD, D. A comparison of feature extraction methods for the classification of dynamic activities from accelerometer data. **IEEE Transactions on Biomedical Engineering**, IEEE, v. 56, n. 3, p. 871–879, 2009. (Cited on pages 67 and 68)
- RATTANI, A. Introduction to adaptive biometric systems. In: _____. **Adaptive Biometric Systems: Recent Advances and Challenges**. [S.l.]: Springer International Publishing, 2015. p. 1–8. ISBN 978-3-319-24865-3. (Cited on pages 26 and 32)
- RATTANI, A.; FRENI, B.; MARCIALIS, G. L.; ROLI, F. Template update methods in adaptive biometric systems: A critical review. In: _____. **Third International Conference on Advances in Biometrics (ICB)**. [S.l.]: Springer Berlin Heidelberg, 2009. p. 847–856. ISBN 978-3-642-01793-3. (Cited on pages 37, 113, and 129)
- RATTANI, A.; MARCIALIS, G.; ROLI, F. Self adaptive systems: An experimental analysis of the performance over time. In: **2011 IEEE Workshop on Computational Intelligence in Biometrics and Identity Management (CIBIM)**. [S.l.]: IEEE, 2011. p. 36–43. (Cited on pages 26 and 51)
- _____. Temporal analysis of biometric template update procedures in uncontrolled environment. In: MAINO, G.; FORESTI, G. (Ed.). **Image Analysis and Processing - ICIAP 2011**. [S.l.]: Springer Berlin Heidelberg, 2011, (Lecture Notes in Computer Science, v. 6978). p. 595–604. ISBN 978-3-642-24084-3. (Cited on pages 37, 113, 161, 165, and 166)
- RATTANI, A.; MARCIALIS, G. L.; ROLI, F. Biometric template update using the graph mincut algorithm: A case study in face verification. In: **2008 Biometrics Symposium (BSYM)**. [S.l.]: IEEE, 2008. p. 23–28. (Cited on pages 36, 38, and 129)
- _____. An experimental analysis of the relationship between biometric template update and the doddington's zoo: A case study in face verification. In: _____. **2009 Proceedings of the 15th International Conference on Image Analysis and Processing (ICIAP)**. [S.l.]: Springer Berlin Heidelberg, 2009. p. 434–442. ISBN 978-3-642-04146-4. (Cited on pages 28, 60, 130, 141, 157, and 163)

_____. Biometric system adaptation by self-update and graph-based techniques. **Journal of Visual Languages & Computing**, Elsevier, v. 24, n. 1, p. 1–9, 2013. ISSN 1045-926X. (Cited on pages 11, 26, 37, 38, 40, 46, 47, 48, and 129)

_____. A multi-modal dataset, protocol and tools for adaptive biometric systems: a benchmarking study. **International Journal of Biometrics (IJBM)**, v. 5, n. 3/4, p. 266–287, 2013. (Cited on pages 38, 45, 46, 59, 60, and 160)

RAYKAR, V. C.; SAHA, A. Data split strategies for evolving predictive models. In: _____. **Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2015, Porto, Portugal, September 7-11, 2015, Proceedings, Part I**. [S.l.]: Springer International Publishing, 2015. p. 3–19. ISBN 978-3-319-23528-8. (Cited on page 136)

REYNOLDS, D. Comparison of background normalization methods for text-independent speaker verification. In: **1997 Eurospeech**. [S.l.: s.n.], 1997. p. 963–966. (Cited on page 115)

ROLI, F.; DIDACI, L.; MARCIALIS, G. Template co-update in multimodal biometric systems. In: LEE, S.-W.; LI, S. Z. (Ed.). **Advances in Biometrics**. [S.l.]: Springer Berlin Heidelberg, 2007, (Lecture Notes in Computer Science, v. 4642). p. 1194–1202. ISBN 978-3-540-74548-8. (Cited on pages 38, 44, and 166)

_____. Adaptive biometric systems that can improve with use. In: RATHA, N.; GOVINDARAJU, V. (Ed.). **Advances in Biometrics**. [S.l.]: Springer London, 2008. p. 447–471. ISBN 978-1-84628-920-0. (Cited on pages 26, 31, 32, 34, and 159)

ROLI, F.; MARCIALIS, G. Semi-supervised PCA-based face recognition using self-training. In: YEUNG, D.-Y.; KWOK, J.; FRED, A.; ROLI, F.; RIDDER, D. (Ed.). **Structural, Syntactic, and Statistical Pattern Recognition**. [S.l.]: Springer Berlin Heidelberg, 2006, (Lecture Notes in Computer Science, v. 4109). p. 560–568. ISBN 978-3-540-37236-3. (Cited on pages 35, 37, 42, 44, and 129)

SCHEIDAT, T.; MAKRUSHIN, A.; VIELHAUER, C. **Automatic Template Update Strategies for Biometrics**. [S.l.], 2007. (Cited on pages 41, 42, 53, 78, and 80)

SCHEIRER, W. J.; ROCHA, A. de R.; SAPKOTA, A.; BOULT, T. E. Toward open set recognition. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, IEEE, v. 35, n. 7, p. 1757–1772, 2013. ISSN 0162-8828. (Cited on pages 57 and 60)

SCHÖLKOPF, B.; PLATT, J. C.; SHAWE-TAYLOR, J. C.; WILLIAMSON, R. C. **Estimating the Support of a High-Dimensional Distribution**. [S.l.], 1999. (Cited on pages 69 and 71)

SPRAGER, S.; ZAZULA, D. A cumulant-based method for gait identification using accelerometer data with principal component analysis and support vector machine. **WSEAS Transactions on Signal Processing**, World Scientific and Engineering Academy and Society (WSEAS), v. 5, n. 11, p. 369–378, 2009. ISSN 1790-5022. (Cited on pages 27, 63, and 66)

STIBOR, T.; TIMMIS, J. Is negative selection appropriate for anomaly detection? **ACM GECCO**, ACM, p. 321–328, 2005. (Cited on pages 27, 69, 70, 75, and 76)

TAX, D. M.; DUIN, R. P. Outliers and data descriptions. In: **Proceedings of the Seventh Annual Conference of the Advanced School for Computing and Imaging**. [S.l.: s.n.], 2001. (Cited on page 71)

TAYLOR, M. E.; STONE, P. Transfer learning for reinforcement learning domains: A survey. **Journal of Machine Learning Research**, JMLR.org, v. 10, p. 1633–1685, 2009. (Cited on page 40)

TEH, P. S.; TEOH, A. B. J.; YUE, S. A survey of keystroke dynamics biometrics. **The Scientific World Journal**, Hindawi, p. 1–24, 2013. (Cited on page 65)

- TORRE, M. De-la; GRANGER, E.; SABOURIN, R.; GORODNICHY, D. O. An adaptive ensemble-based system for face recognition in person re-identification. **Machine Vision and Applications**, Springer, v. 26, n. 6, p. 741–773, 2015. ISSN 1432-1769. (Cited on page 167)
- ULUDAG, U.; ROSS, A.; JAIN, A. Biometric template selection and update: a case study in fingerprints. **Pattern Recognition**, Elsevier, v. 37, n. 7, p. 1533 – 1542, 2004. ISSN 0031-3203. (Cited on pages 11, 36, 41, 44, 52, 129, 160, 161, and 165)
- VAPNIK, V. N. **Statistical Learning Theory**. [S.l.]: Wiley, 1998. ISBN 978-0-471-03003-4. (Cited on page 71)
- WINDLEY, P. J. **Digital Identity**. [S.l.]: O'Reilly Media, 2005. ISBN 0596008783. (Cited on page 25)
- YANG, J.; YAN, R.; HAUPTMANN, A. G. Adapting svm classifiers to data with shifted distributions. In: **Seventh IEEE International Conference on Data Mining Workshops (ICDMW 2007)**. [S.l.: s.n.], 2007. p. 69–76. ISSN 2375-9232. (Cited on page 40)
- YU, E.; CHO, S. Novelty detection approach for keystroke dynamics identity verification. In: LIU, J.; CHEUNG, Y.-m.; YIN, H. (Ed.). **Intelligent Data Engineering and Automated Learning**. [S.l.]: Springer Berlin / Heidelberg, 2003, (Lecture Notes in Computer Science, v. 2690). p. 1016–1023. ISBN 978-3-540-40550-4. (Cited on page 72)
- YU, Q.; YIN, Y.; YANG, G.; NING, Y.; LI, Y. Face and gait recognition based on semi-supervised learning. In: LIU, C.-L.; ZHANG, C.; WANG, L. (Ed.). **Pattern Recognition**. [S.l.]: Springer Berlin Heidelberg, 2012, (Communications in Computer and Information Science, v. 321). p. 284–291. ISBN 978-3-642-33505-1. (Cited on page 32)
- ZHANG, Z.; HU, M.; WANG, Y. A survey of advances in biometric gait recognition. In: SUN, Z.; LAI, J.; CHEN, X.; TAN, T. (Ed.). **Biometric Recognition**. [S.l.]: Springer Berlin Heidelberg, 2011, (LNCS, v. 7098). p. 150–158. ISBN 978-3-642-25448-2. (Cited on page 66)
- ZHU, X. **Semi-Supervised Learning Literature Survey**. [S.l.], 2006. (Cited on pages 37 and 38)
- ŽLIOBAITĖ, I. **Learning under Concept Drift: an Overview**. [S.l.], 2010. abs/1010.4784. (Cited on page 26)
- ŽLIOBAITĖ, I.; BIFET, A.; READ, J.; PFAHRINGER, B.; HOLMES, G. Evaluation methods and decision theory for classification of streaming data with temporal dependence. **Machine Learning**, Springer, v. 98, n. 3, p. 455–482, 2015. ISSN 1573-0565. (Cited on pages 61 and 161)

EXPERIMENTAL RESULTS

This appendix presents the full table of results for most experiments performed for this thesis: Tables 24 to 30. The best results for each group are highlighted in bold (standard deviation among runs is shown between parenthesis). Adaptive biometric systems are indicated by the adaptation strategy between parenthesis like, for example, *Self-Detector (Sliding)*. Conversely, non-adaptive biometric systems do not use an adaptation strategy, hence, they do not contain a parenthesis in their names like, for example, *Self-Detector*. Part of these tables are also shown throughout the thesis to discuss the obtained results.

Table 24 – Results - GREYC.

Biometric system	FMR	FNMR	Acc (balanc.)
<i>Self-Detector</i>	0.090 (0.010)	0.165 (0.005)	0.872 (0.006)
<i>Self-Detector (Sliding)</i>	0.092 (0.011)	0.129 (0.004)	0.890 (0.006)
<i>Self-Detector (Growing)</i>	0.105 (0.011)	0.119 (0.005)	0.888 (0.006)
<i>M2005</i>	0.221 (0.019)	0.130 (0.003)	0.824 (0.009)
<i>M2005 (DB)</i>	0.220 (0.019)	0.086 (0.003)	0.847 (0.009)
<i>M2005 (IDB)</i>	0.210 (0.018)	0.092 (0.004)	0.849 (0.008)
<i>M2005 (Adapted Thresholds - Growing)</i>	0.239 (0.019)	0.092 (0.003)	0.835 (0.010)
<i>M2005 (Adapted Thresholds - Sliding)</i>	0.184 (0.017)	0.115 (0.004)	0.851 (0.008)
<i>Self-Detector (Usage Control)</i>	0.091 (0.010)	0.140 (0.005)	0.884 (0.006)
<i>Self-Detector (Usage Control 2)</i>	0.069 (0.009)	0.168 (0.006)	0.882 (0.006)
<i>Self-Detector (Usage Control R)</i>	0.092 (0.010)	0.140 (0.005)	0.884 (0.006)
<i>Self-Detector (Usage Control S)</i>	0.089 (0.010)	0.149 (0.005)	0.881 (0.006)
<i>Self-Detector (ETU 0)</i>	0.104 (0.010)	0.121 (0.005)	0.887 (0.006)
<i>M2005 (ETU 0)</i>	0.189 (0.017)	0.115 (0.004)	0.848 (0.008)
<i>Self-Detector (ETU 1)</i>	0.089 (0.011)	0.144 (0.006)	0.884 (0.006)
<i>Self-Detector (ETU 2)</i>	0.096 (0.010)	0.126 (0.005)	0.889 (0.006)
<i>M2005 (ETU 3)</i>	0.192 (0.017)	0.100 (0.004)	0.854 (0.008)
<i>Self-Detector (ETU 0) PGP</i>	0.104 (0.010)	0.121 (0.005)	0.887 (0.006)
<i>M2005 (ETU 0) PGP</i>	0.189 (0.017)	0.116 (0.004)	0.848 (0.008)
<i>Self-Detector (ETU 1) PGP</i>	0.089 (0.011)	0.144 (0.006)	0.884 (0.006)
<i>Self-Detector (ETU 2) PGP</i>	0.096 (0.010)	0.126 (0.005)	0.889 (0.006)
<i>M2005 (ETU 3) PGP</i>	0.192 (0.017)	0.101 (0.004)	0.854 (0.008)
<i>Self-Detector (ETU - Sliding) PGP</i>	0.092 (0.011)	0.129 (0.004)	0.889 (0.006)
<i>ModBioS (User)</i>	0.153 (0.020)	0.095 (0.011)	0.876 (0.007)
<i>ModBioS (Grouped)</i>	0.077 (0.018)	0.151 (0.012)	0.886 (0.005)
<i>ModBioS - Fusion (User)</i>	0.158 (0.020)	0.086 (0.011)	0.878 (0.007)
<i>ModBioS - Fusion (Grouped)</i>	0.130 (0.107)	0.117 (0.042)	0.877 (0.033)
<i>ModBioS (Hybrid)</i>	0.102 (0.051)	0.132 (0.034)	0.883 (0.010)
<i>ModBioS - Fusion (Hybrid)</i>	0.102 (0.050)	0.123 (0.032)	0.888 (0.010)

Table 25 – Results - CMU.

Biometric system	FMR	FNMR	Acc (balanc.)
<i>Self-Detector</i>	0.287 (0.023)	0.410 (0.016)	0.651 (0.009)
<i>Self-Detector (Sliding)</i>	0.291 (0.031)	0.211 (0.013)	0.749 (0.016)
<i>Self-Detector (Growing)</i>	0.562 (0.039)	0.118 (0.009)	0.660 (0.018)
<i>M2005</i>	0.273 (0.028)	0.451 (0.019)	0.638 (0.013)
<i>M2005 (DB)</i>	0.129 (0.014)	0.373 (0.014)	0.749 (0.010)
<i>M2005 (IDB)</i>	0.122 (0.011)	0.306 (0.008)	0.786 (0.006)
<i>M2005 (Adapted Thresholds - Growing)</i>	0.462 (0.019)	0.113 (0.007)	0.712 (0.009)
<i>M2005 (Adapted Thresholds - Sliding)</i>	0.160 (0.011)	0.295 (0.010)	0.773 (0.007)
<i>Self-Detector (Usage Control)</i>	0.351 (0.033)	0.211 (0.013)	0.719 (0.016)
<i>Self-Detector (Usage Control 2)</i>	0.143 (0.012)	0.323 (0.014)	0.767 (0.009)
<i>Self-Detector (Usage Control R)</i>	0.311 (0.030)	0.220 (0.013)	0.735 (0.015)
<i>Self-Detector (Usage Control S)</i>	0.213 (0.014)	0.275 (0.012)	0.756 (0.008)
<i>Self-Detector (ETU 0)</i>	0.538 (0.016)	0.102 (0.007)	0.680 (0.009)
<i>M2005 (ETU 0)</i>	0.064 (0.008)	0.623 (0.019)	0.656 (0.009)
<i>Self-Detector (ETU 1)</i>	0.251 (0.015)	0.203 (0.017)	0.773 (0.013)
<i>Self-Detector (ETU 2)</i>	0.285 (0.019)	0.207 (0.013)	0.754 (0.013)
<i>M2005 (ETU 3)</i>	0.244 (0.016)	0.143 (0.006)	0.807 (0.009)
<i>Self-Detector (ETU 0) PGP</i>	0.573 (0.018)	0.088 (0.009)	0.670 (0.009)
<i>M2005 (ETU 0) PGP</i>	0.064 (0.009)	0.669 (0.015)	0.633 (0.008)
<i>Self-Detector (ETU 1) PGP</i>	0.271 (0.012)	0.201 (0.015)	0.764 (0.010)
<i>Self-Detector (ETU 2) PGP</i>	0.268 (0.016)	0.224 (0.014)	0.754 (0.011)
<i>M2005 (ETU 3) PGP</i>	0.175 (0.017)	0.243 (0.011)	0.791 (0.010)
<i>Self-Det. (ETU - Sliding) PGP</i>	0.250 (0.021)	0.251 (0.015)	0.750 (0.011)
<i>ModBioS (User)</i>	0.215 (0.015)	0.222 (0.013)	0.781 (0.010)
<i>ModBioS (Grouped)</i>	0.212 (0.071)	0.186 (0.085)	0.801 (0.016)
<i>ModBioS - Fusion (User)</i>	0.213 (0.020)	0.196 (0.021)	0.796 (0.009)
<i>ModBioS - Fusion (Grouped)</i>	0.165 (0.066)	0.236 (0.067)	0.800 (0.004)
<i>ModBioS (Hybrid)</i>	0.214 (0.024)	0.207 (0.028)	0.790 (0.011)
<i>ModBioS - Fusion (Hybrid)</i>	0.216 (0.021)	0.159 (0.017)	0.812 (0.009)

Table 26 – Results - GREYC-Web (Logins).

Biometric system	FMR	FNMR	Acc (balanc.)
<i>Self-Detector</i>	0.066 (0.008)	0.141 (0.005)	0.896 (0.005)
<i>Self-Detector (Sliding)</i>	0.074 (0.011)	0.085 (0.004)	0.920 (0.007)
<i>Self-Detector (Growing)</i>	0.124 (0.015)	0.060 (0.003)	0.908 (0.008)
<i>M2005</i>	0.096 (0.013)	0.245 (0.016)	0.829 (0.008)
<i>M2005 (DB)</i>	0.083 (0.012)	0.179 (0.012)	0.869 (0.008)
<i>M2005 (IDB)</i>	0.095 (0.015)	0.131 (0.011)	0.887 (0.008)
<i>M2005 (Adapted Thresholds - Growing)</i>	0.132 (0.016)	0.303 (0.023)	0.782 (0.015)
<i>M2005 (Adapted Thresholds - Sliding)</i>	0.074 (0.011)	0.394 (0.020)	0.766 (0.010)
<i>Self-Detector (Usage Control)</i>	0.078 (0.011)	0.084 (0.003)	0.919 (0.006)
<i>Self-Detector (Usage Control 2)</i>	0.035 (0.007)	0.148 (0.010)	0.908 (0.007)
<i>Self-Detector (Usage Control R)</i>	0.069 (0.009)	0.086 (0.004)	0.922 (0.006)
<i>Self-Detector (Usage Control S)</i>	0.053 (0.007)	0.123 (0.005)	0.912 (0.005)
<i>Self-Detector (ETU 0)</i>	0.353 (0.030)	0.034 (0.011)	0.807 (0.017)
<i>M2005 (ETU 0)</i>	0.073 (0.011)	0.427 (0.046)	0.750 (0.021)
<i>Self-Detector (ETU 1)</i>	0.065 (0.013)	0.146 (0.020)	0.894 (0.011)
<i>Self-Detector (ETU 2)</i>	0.103 (0.014)	0.071 (0.009)	0.913 (0.010)
<i>M2005 (ETU 3)</i>	0.163 (0.017)	0.118 (0.018)	0.860 (0.013)
<i>Self-Detector (ETU 0) PGP</i>	0.355 (0.032)	0.042 (0.010)	0.802 (0.020)
<i>M2005 (ETU 0) PGP</i>	0.067 (0.010)	0.505 (0.037)	0.714 (0.018)
<i>Self-Detector (ETU 1) PGP</i>	0.064 (0.013)	0.151 (0.019)	0.893 (0.010)
<i>Self-Detector (ETU 2) PGP</i>	0.099 (0.014)	0.078 (0.011)	0.911 (0.011)
<i>M2005 (ETU 3) PGP</i>	0.120 (0.016)	0.151 (0.018)	0.865 (0.013)
<i>Self-Detector (ETU - Sliding) PGP</i>	0.067 (0.010)	0.106 (0.008)	0.913 (0.008)
<i>ModBioS (User)</i>	0.131 (0.021)	0.060 (0.011)	0.904 (0.008)
<i>ModBioS (Grouped)</i>	0.070 (0.019)	0.115 (0.030)	0.907 (0.009)
<i>ModBioS - Fusion (User)</i>	0.137 (0.020)	0.055 (0.006)	0.904 (0.009)
<i>ModBioS - Fusion (Grouped)</i>	0.090 (0.019)	0.084 (0.019)	0.913 (0.007)
<i>ModBioS (Hybrid)</i>	0.084 (0.016)	0.099 (0.021)	0.908 (0.006)
<i>ModBioS - Fusion (Hybrid)</i>	0.082 (0.020)	0.092 (0.027)	0.913 (0.010)

Table 27 – Results - GREYC-Web (Passwords).

Biometric system	FMR	FNMR	Acc (balanc.)
<i>Self-Detector</i>	0.388 (0.014)	0.180 (0.002)	0.716 (0.007)
<i>Self-Detector (Sliding)</i>	0.330 (0.021)	0.205 (0.008)	0.733 (0.012)
<i>Self-Detector (Growing)</i>	0.468 (0.017)	0.123 (0.002)	0.704 (0.008)
<i>M2005</i>	0.329 (0.035)	0.251 (0.015)	0.710 (0.024)
<i>M2005 (DB)</i>	0.251 (0.026)	0.240 (0.014)	0.754 (0.018)
<i>M2005 (IDB)</i>	0.247 (0.023)	0.190 (0.006)	0.781 (0.013)
<i>M2005 (Adapted Thresholds - Growing)</i>	0.273 (0.030)	0.348 (0.023)	0.690 (0.022)
<i>M2005 (Adapted Thresholds - Sliding)</i>	0.146 (0.022)	0.465 (0.022)	0.694 (0.018)
<i>Self-Detector (Usage Control)</i>	0.383 (0.021)	0.171 (0.006)	0.723 (0.011)
<i>Self-Detector (Usage Control 2)</i>	0.186 (0.013)	0.396 (0.012)	0.709 (0.010)
<i>Self-Detector (Usage Control R)</i>	0.360 (0.022)	0.188 (0.006)	0.726 (0.012)
<i>Self-Detector (Usage Control S)</i>	0.255 (0.017)	0.296 (0.009)	0.725 (0.009)
<i>Self-Detector (ETU 0)</i>	0.483 (0.018)	0.115 (0.012)	0.701 (0.010)
<i>M2005 (ETU 0)</i>	0.138 (0.020)	0.583 (0.026)	0.640 (0.013)
<i>Self-Detector (ETU 1)</i>	0.441 (0.019)	0.140 (0.014)	0.710 (0.011)
<i>Self-Detector (ETU 2)</i>	0.364 (0.020)	0.171 (0.009)	0.733 (0.011)
<i>M2005 (ETU 3)</i>	0.294 (0.016)	0.177 (0.021)	0.765 (0.013)
<i>Self-Detector (ETU 0) PGP</i>	0.494 (0.018)	0.118 (0.012)	0.694 (0.010)
<i>M2005 (ETU 0) PGP</i>	0.134 (0.018)	0.614 (0.025)	0.626 (0.011)
<i>Self-Detector (ETU 1) PGP</i>	0.448 (0.019)	0.141 (0.014)	0.705 (0.011)
<i>Self-Detector (ETU 2) PGP</i>	0.331 (0.021)	0.213 (0.011)	0.728 (0.012)
<i>M2005 (ETU 3) PGP</i>	0.247 (0.019)	0.268 (0.017)	0.742 (0.014)
<i>Self-Detector (ETU - Sliding) PGP</i>	0.292 (0.023)	0.254 (0.011)	0.727 (0.012)
<i>ModBioS (User)</i>	0.254 (0.025)	0.254 (0.020)	0.746 (0.012)
<i>ModBioS (Grouped)</i>	0.290 (0.016)	0.231 (0.011)	0.739 (0.010)
<i>ModBioS - Fusion (User)</i>	0.226 (0.018)	0.266 (0.020)	0.754 (0.010)
<i>ModBioS - Fusion (Grouped)</i>	0.292 (0.022)	0.219 (0.018)	0.745 (0.009)
<i>ModBioS (Hybrid)</i>	0.263 (0.036)	0.230 (0.030)	0.754 (0.019)
<i>ModBioS - Fusion (Hybrid)</i>	0.295 (0.017)	0.214 (0.012)	0.745 (0.010)

Table 28 – Results - McGill.

Biometric system	FMR	FNMR	Acc (balanc.)
<i>Self-Detector</i>	0.104 (0.018)	0.545 (0.023)	0.675 (0.019)
<i>Self-Detector (Sliding)</i>	0.166 (0.018)	0.282 (0.036)	0.776 (0.023)
<i>Self-Detector (Growing)</i>	0.490 (0.036)	0.116 (0.029)	0.697 (0.019)
<i>OCSVM</i>	0.323 (0.031)	0.486 (0.025)	0.596 (0.019)
<i>OCSVM (Growing)</i>	0.323 (0.031)	0.486 (0.025)	0.596 (0.019)
<i>OCSVM (Sliding)</i>	0.037 (0.006)	0.894 (0.006)	0.534 (0.005)
<i>Self-Detector (Usage Control)</i>	0.294 (0.035)	0.230 (0.037)	0.738 (0.027)
<i>Self-Detector (Usage Control 2)</i>	0.077 (0.007)	0.356 (0.038)	0.784 (0.021)
<i>Self-Detector (Usage Control R)</i>	0.233 (0.028)	0.250 (0.036)	0.758 (0.024)
<i>Self-Detector (Usage Control S)</i>	0.115 (0.016)	0.422 (0.035)	0.732 (0.023)
<i>Self-Detector (ETU 0)</i>	0.569 (0.023)	0.141 (0.034)	0.645 (0.017)
<i>Self-Detector (ETU 1)</i>	0.523 (0.030)	0.155 (0.039)	0.661 (0.015)
<i>Self-Detector (ETU 2)</i>	0.209 (0.026)	0.315 (0.044)	0.738 (0.033)
<i>Self-Detector (ETU 0) PGP</i>	0.585 (0.031)	0.168 (0.037)	0.623 (0.014)
<i>Self-Detector (ETU 1) PGP</i>	0.539 (0.036)	0.192 (0.040)	0.635 (0.014)
<i>Self-Detector (ETU 2) PGP</i>	0.186 (0.019)	0.361 (0.049)	0.727 (0.031)
<i>Self-Detector (ETU - Sliding) PGP</i>	0.134 (0.013)	0.435 (0.045)	0.715 (0.025)
<i>ModBioS (User)</i>	0.350 (0.038)	0.211 (0.029)	0.719 (0.016)
<i>ModBioS (Grouped)</i>	0.320 (0.066)	0.168 (0.048)	0.756 (0.014)
<i>ModBioS - Fusion (User)</i>	0.370 (0.037)	0.214 (0.044)	0.708 (0.023)
<i>ModBioS - Fusion (Grouped)</i>	0.386 (0.145)	0.162 (0.045)	0.726 (0.054)
<i>ModBioS (Hybrid)</i>	0.355 (0.090)	0.166 (0.045)	0.740 (0.031)
<i>ModBioS - Fusion (Hybrid)</i>	0.360 (0.104)	0.158 (0.035)	0.741 (0.037)

Table 29 – Results - WISDM 1.1.

Biometric system	FMR	FNMR	Acc (balanc.)
<i>Self-Detector</i>	0.183 (0.013)	0.242 (0.013)	0.788 (0.009)
<i>Self-Detector (Sliding)</i>	0.227 (0.028)	0.186 (0.021)	0.794 (0.016)
<i>Self-Detector (Growing)</i>	0.349 (0.024)	0.123 (0.014)	0.764 (0.016)
<i>OCSVM</i>	0.183 (0.017)	0.426 (0.024)	0.695 (0.013)
<i>OCSVM (Growing)</i>	0.183 (0.017)	0.426 (0.024)	0.696 (0.013)
<i>OCSVM (Sliding)</i>	0.062 (0.013)	0.633 (0.014)	0.653 (0.010)
<i>Self-Detector (Usage Control)</i>	0.261 (0.030)	0.163 (0.024)	0.788 (0.017)
<i>Self-Detector (Usage Control 2)</i>	0.123 (0.015)	0.213 (0.021)	0.832 (0.011)
<i>Self-Detector (Usage Control R)</i>	0.227 (0.027)	0.174 (0.024)	0.799 (0.017)
<i>Self-Detector (Usage Control S)</i>	0.152 (0.014)	0.237 (0.015)	0.805 (0.010)
<i>Self-Detector (ETU 0)</i>	0.416 (0.022)	0.143 (0.024)	0.720 (0.012)
<i>Self-Detector (ETU 1)</i>	0.389 (0.030)	0.155 (0.027)	0.728 (0.015)
<i>Self-Detector (ETU 2)</i>	0.217 (0.028)	0.191 (0.021)	0.796 (0.017)
<i>Self-Detector (ETU 0) PGP</i>	0.441 (0.025)	0.155 (0.024)	0.702 (0.013)
<i>Self-Detector (ETU 1) PGP</i>	0.407 (0.033)	0.163 (0.025)	0.715 (0.015)
<i>Self-Detector (ETU 2) PGP</i>	0.210 (0.026)	0.195 (0.021)	0.798 (0.016)
<i>Self-Detector (ETU - Sliding) PGP</i>	0.202 (0.025)	0.195 (0.021)	0.801 (0.015)
<i>ModBioS (User)</i>	0.129 (0.015)	0.244 (0.026)	0.814 (0.015)
<i>ModBioS (Grouped)</i>	0.166 (0.056)	0.201 (0.028)	0.817 (0.021)
<i>ModBioS - Fusion (User)</i>	0.130 (0.017)	0.228 (0.024)	0.821 (0.012)
<i>ModBioS - Fusion (Grouped)</i>	0.128 (0.025)	0.203 (0.023)	0.835 (0.011)
<i>ModBioS (Hybrid)</i>	0.146 (0.025)	0.195 (0.023)	0.829 (0.009)
<i>ModBioS - Fusion (Hybrid)</i>	0.129 (0.023)	0.200 (0.023)	0.835 (0.010)

Table 30 – Results - WISDM 2.0.

Biometric system	FMR	FNMR	Acc (balanc.)
<i>Self-Detector</i>	0.168 (0.006)	0.220 (0.007)	0.806 (0.005)
<i>Self-Detector (Sliding)</i>	0.203 (0.015)	0.150 (0.008)	0.824 (0.008)
<i>Self-Detector (Growing)</i>	0.302 (0.018)	0.116 (0.009)	0.791 (0.008)
<i>OCSVM</i>	0.090 (0.006)	0.381 (0.007)	0.765 (0.006)
<i>OCSVM (Growing)</i>	0.090 (0.006)	0.381 (0.007)	0.765 (0.006)
<i>OCSVM (Sliding)</i>	0.028 (0.003)	0.572 (0.008)	0.700 (0.005)
<i>Self-Detector (Usage Control)</i>	0.227 (0.016)	0.136 (0.008)	0.819 (0.008)
<i>Self-Detector (Usage Control 2)</i>	0.123 (0.007)	0.176 (0.009)	0.851 (0.006)
<i>Self-Detector (Usage Control R)</i>	0.198 (0.014)	0.144 (0.009)	0.829 (0.007)
<i>Self-Detector (Usage Control S)</i>	0.156 (0.007)	0.185 (0.010)	0.829 (0.006)
<i>Self-Detector (ETU 0)</i>	0.437 (0.009)	0.096 (0.006)	0.734 (0.005)
<i>Self-Detector (ETU 1)</i>	0.422 (0.011)	0.103 (0.007)	0.738 (0.006)
<i>Self-Detector (ETU 2)</i>	0.206 (0.010)	0.154 (0.007)	0.820 (0.006)
<i>Self-Detector (ETU 0) PGP</i>	0.462 (0.011)	0.093 (0.006)	0.723 (0.006)
<i>Self-Detector (ETU 1) PGP</i>	0.443 (0.012)	0.100 (0.007)	0.728 (0.006)
<i>Self-Detector (ETU 2) PGP</i>	0.197 (0.010)	0.167 (0.009)	0.818 (0.006)
<i>Self-Detector (ETU - Sliding) PGP</i>	0.179 (0.011)	0.178 (0.009)	0.822 (0.006)
<i>ModBioS (User)</i>	0.152 (0.009)	0.167 (0.009)	0.841 (0.006)
<i>ModBioS (Grouped)</i>	0.130 (0.015)	0.172 (0.007)	0.849 (0.008)
<i>ModBioS - Fusion (User)</i>	0.147 (0.011)	0.164 (0.012)	0.844 (0.007)
<i>ModBioS - Fusion (Grouped)</i>	0.122 (0.017)	0.178 (0.029)	0.850 (0.008)
<i>ModBioS (Hybrid)</i>	0.141 (0.007)	0.176 (0.013)	0.842 (0.008)
<i>ModBioS - Fusion (Hybrid)</i>	0.112 (0.012)	0.187 (0.025)	0.851 (0.008)

EXPERIMENTAL RESULTS - SCORE NORMALIZATION

This appendix presents the full table of results for the experiments using score normalization in Tables 31 to 37. The best results for each group are highlighted in bold (standard deviation among runs is shown between parenthesis).

Table 31 – Results using score normalization - GREYC.

Biometric system	T-Norm			Z-Norm		
	FMR	FNMR	Acc (balanced)	FMR	FNMR	Acc (balanced)
<i>Self-Detector</i>	0.111 (0.010)	0.097 (0.006)	0.896 (0.005)	0.111 (0.011)	0.146 (0.007)	0.871 (0.005)
<i>Self-Detector (Sliding)</i>	0.103 (0.009)	0.071 (0.008)	0.913 (0.006)	0.114 (0.013)	0.118 (0.007)	0.884 (0.006)
<i>Self-Detector (Growing)</i>	0.104 (0.009)	0.072 (0.008)	0.912 (0.006)	0.133 (0.014)	0.108 (0.007)	0.880 (0.006)
<i>Self-Detector (Usage Control)</i>	0.110 (0.009)	0.077 (0.006)	0.907 (0.005)	0.115 (0.012)	0.128 (0.006)	0.878 (0.005)
<i>Self-Detector (Usage Control R)</i>	0.109 (0.009)	0.077 (0.006)	0.907 (0.005)	0.114 (0.012)	0.128 (0.006)	0.879 (0.005)
<i>Self-Detector (Usage Control S)</i>	0.110 (0.010)	0.083 (0.006)	0.904 (0.005)	0.109 (0.012)	0.134 (0.006)	0.879 (0.005)
<i>Self-Detector (Usage Control 2)</i>	0.108 (0.010)	0.073 (0.008)	0.910 (0.006)	0.085 (0.011)	0.146 (0.007)	0.884 (0.005)
<i>M2005</i>	0.138 (0.009)	0.156 (0.010)	0.853 (0.006)	0.098 (0.014)	0.223 (0.010)	0.839 (0.008)
<i>M2005 (DB)</i>	0.141 (0.010)	0.110 (0.007)	0.875 (0.006)	0.098 (0.013)	0.188 (0.011)	0.857 (0.008)
<i>M2005 (IDB)</i>	0.141 (0.009)	0.132 (0.013)	0.864 (0.008)	0.100 (0.012)	0.166 (0.008)	0.867 (0.006)
Biometric system	F-Norm			Adaptive F-Norm		
	FMR	FNMR	Acc (balanced)	FMR	FNMR	Acc (balanced)
<i>Self-Detector</i>	0.107 (0.011)	0.134 (0.005)	0.879 (0.005)	0.104 (0.010)	0.112 (0.006)	0.892 (0.006)
<i>Self-Detector (Sliding)</i>	0.109 (0.012)	0.105 (0.005)	0.893 (0.006)	0.108 (0.011)	0.086 (0.006)	0.903 (0.006)
<i>Self-Detector (Growing)</i>	0.126 (0.013)	0.093 (0.005)	0.891 (0.006)	0.114 (0.011)	0.079 (0.005)	0.903 (0.006)
<i>Self-Detector (Usage Control)</i>	0.108 (0.011)	0.114 (0.005)	0.889 (0.005)	0.105 (0.010)	0.092 (0.005)	0.901 (0.006)
<i>Self-Detector (Usage Control R)</i>	0.108 (0.012)	0.113 (0.005)	0.889 (0.005)	0.105 (0.010)	0.092 (0.006)	0.902 (0.006)
<i>Self-Detector (Usage Control S)</i>	0.105 (0.011)	0.123 (0.005)	0.886 (0.005)	0.106 (0.010)	0.099 (0.004)	0.898 (0.006)
<i>Self-Detector (Usage Control 2)</i>	0.081 (0.011)	0.140 (0.006)	0.890 (0.006)	0.093 (0.010)	0.098 (0.006)	0.904 (0.006)
<i>M2005</i>	0.141 (0.018)	0.152 (0.004)	0.853 (0.009)	0.177 (0.013)	0.125 (0.003)	0.849 (0.007)
<i>M2005 (DB)</i>	0.125 (0.018)	0.120 (0.004)	0.878 (0.009)	0.172 (0.014)	0.086 (0.003)	0.871 (0.007)
<i>M2005 (IDB)</i>	0.123 (0.018)	0.111 (0.005)	0.883 (0.008)	0.170 (0.014)	0.091 (0.005)	0.869 (0.006)

Table 32 – Results using score normalization - CMU.

Biometric system	T-Norm			Z-Norm		
	FMR	FNMR	Acc (balanced)	FMR	FNMR	Acc (balanced)
<i>Self-Detector</i>	0.209 (0.007)	0.324 (0.016)	0.734 (0.009)	0.158 (0.016)	0.541 (0.023)	0.651 (0.008)
<i>Self-Detector (Sliding)</i>	0.217 (0.005)	0.125 (0.008)	0.829 (0.005)	0.183 (0.026)	0.345 (0.019)	0.736 (0.011)
<i>Self-Detector (Growing)</i>	0.195 (0.004)	0.136 (0.006)	0.834 (0.003)	0.376 (0.043)	0.252 (0.019)	0.686 (0.013)
<i>Self-Detector (Usage Control)</i>	0.217 (0.006)	0.133 (0.006)	0.825 (0.004)	0.213 (0.028)	0.352 (0.020)	0.718 (0.010)
<i>Self-Detector (Usage Control R)</i>	0.220 (0.004)	0.126 (0.007)	0.827 (0.004)	0.189 (0.026)	0.357 (0.021)	0.727 (0.011)
<i>Self-Detector (Usage Control S)</i>	0.233 (0.005)	0.140 (0.008)	0.814 (0.004)	0.131 (0.011)	0.425 (0.022)	0.722 (0.008)
<i>Self-Detector (Usage Control 2)</i>	0.225 (0.005)	0.139 (0.006)	0.818 (0.004)	0.087 (0.010)	0.462 (0.020)	0.725 (0.009)
<i>M2005</i>	0.196 (0.006)	0.420 (0.016)	0.692 (0.009)	0.220 (0.020)	0.447 (0.007)	0.667 (0.010)
<i>M2005 (DB)</i>	0.203 (0.007)	0.205 (0.015)	0.796 (0.010)	0.167 (0.015)	0.392 (0.026)	0.720 (0.011)
<i>M2005 (IDB)</i>	0.205 (0.007)	0.184 (0.015)	0.806 (0.009)	0.190 (0.016)	0.309 (0.026)	0.750 (0.011)
Biometric system	F-Norm			Adaptive F-Norm		
	FMR	FNMR	Acc (balanced)	FMR	FNMR	Acc (balanced)
<i>Self-Detector</i>	0.218 (0.022)	0.476 (0.018)	0.653 (0.008)	0.266 (0.021)	0.359 (0.022)	0.687 (0.007)
<i>Self-Detector (Sliding)</i>	0.213 (0.030)	0.259 (0.014)	0.764 (0.013)	0.264 (0.027)	0.152 (0.013)	0.792 (0.015)
<i>Self-Detector (Growing)</i>	0.454 (0.049)	0.166 (0.013)	0.690 (0.019)	0.552 (0.048)	0.101 (0.009)	0.674 (0.021)
<i>Self-Detector (Usage Control)</i>	0.259 (0.034)	0.266 (0.015)	0.737 (0.013)	0.305 (0.033)	0.158 (0.010)	0.769 (0.015)
<i>Self-Detector (Usage Control R)</i>	0.230 (0.029)	0.272 (0.015)	0.749 (0.011)	0.266 (0.023)	0.153 (0.011)	0.790 (0.010)
<i>Self-Detector (Usage Control S)</i>	0.156 (0.013)	0.338 (0.017)	0.753 (0.006)	0.213 (0.013)	0.181 (0.011)	0.803 (0.005)
<i>Self-Detector (Usage Control 2)</i>	0.100 (0.011)	0.391 (0.014)	0.755 (0.007)	0.182 (0.012)	0.203 (0.011)	0.808 (0.006)
<i>M2005</i>	0.233 (0.018)	0.453 (0.012)	0.657 (0.013)	0.181 (0.017)	0.531 (0.011)	0.644 (0.009)
<i>M2005 (DB)</i>	0.108 (0.013)	0.347 (0.029)	0.772 (0.014)	0.096 (0.009)	0.387 (0.018)	0.759 (0.009)
<i>M2005 (IDB)</i>	0.097 (0.010)	0.295 (0.022)	0.804 (0.010)	0.084 (0.007)	0.357 (0.021)	0.779 (0.010)

Table 33 – Results using score normalization - GREYC-Web (Logins).

Biometric system	T-Norm			Z-Norm		
	FMR	FNMR	Acc (balanced)	FMR	FNMR	Acc (balanced)
<i>Self-Detector</i>	0.095 (0.008)	0.048 (0.011)	0.929 (0.007)	0.104 (0.012)	0.104 (0.013)	0.896 (0.005)
<i>Self-Detector (Sliding)</i>	0.092 (0.008)	0.024 (0.004)	0.942 (0.005)	0.152 (0.021)	0.061 (0.010)	0.894 (0.009)
<i>Self-Detector (Growing)</i>	0.094 (0.008)	0.023 (0.006)	0.942 (0.005)	0.262 (0.034)	0.038 (0.007)	0.850 (0.014)
<i>Self-Detector (Usage Control)</i>	0.096 (0.008)	0.024 (0.005)	0.940 (0.005)	0.169 (0.021)	0.058 (0.008)	0.886 (0.010)
<i>Self-Detector (Usage Control R)</i>	0.094 (0.008)	0.024 (0.005)	0.941 (0.006)	0.137 (0.019)	0.058 (0.008)	0.902 (0.009)
<i>Self-Detector (Usage Control S)</i>	0.098 (0.007)	0.027 (0.005)	0.937 (0.004)	0.089 (0.009)	0.081 (0.011)	0.915 (0.006)
<i>Self-Detector (Usage Control 2)</i>	0.094 (0.007)	0.028 (0.005)	0.939 (0.005)	0.062 (0.009)	0.116 (0.016)	0.911 (0.009)
<i>M2005</i>	0.096 (0.006)	0.160 (0.011)	0.872 (0.006)	0.113 (0.011)	0.173 (0.013)	0.857 (0.008)
<i>M2005 (DB)</i>	0.102 (0.008)	0.094 (0.010)	0.902 (0.007)	0.139 (0.017)	0.130 (0.016)	0.866 (0.013)
<i>M2005 (IDB)</i>	0.104 (0.007)	0.090 (0.009)	0.903 (0.007)	0.156 (0.018)	0.101 (0.015)	0.872 (0.012)
Biometric system	F-Norm			Adaptive F-Norm		
	FMR	FNMR	Acc (balanced)	FMR	FNMR	Acc (balanced)
<i>Self-Detector</i>	0.081 (0.012)	0.112 (0.008)	0.903 (0.005)	0.054 (0.007)	0.111 (0.007)	0.918 (0.004)
<i>Self-Detector (Sliding)</i>	0.092 (0.013)	0.063 (0.006)	0.922 (0.007)	0.068 (0.008)	0.055 (0.003)	0.938 (0.004)
<i>Self-Detector (Growing)</i>	0.157 (0.021)	0.041 (0.004)	0.901 (0.009)	0.090 (0.010)	0.043 (0.003)	0.933 (0.005)
<i>Self-Detector (Usage Control)</i>	0.098 (0.014)	0.063 (0.005)	0.919 (0.007)	0.067 (0.008)	0.058 (0.004)	0.937 (0.005)
<i>Self-Detector (Usage Control R)</i>	0.087 (0.012)	0.062 (0.006)	0.925 (0.006)	0.065 (0.007)	0.055 (0.004)	0.940 (0.004)
<i>Self-Detector (Usage Control S)</i>	0.067 (0.008)	0.097 (0.009)	0.918 (0.005)	0.055 (0.006)	0.079 (0.007)	0.933 (0.005)
<i>Self-Detector (Usage Control 2)</i>	0.044 (0.007)	0.128 (0.011)	0.914 (0.007)	0.044 (0.006)	0.090 (0.006)	0.933 (0.005)
<i>M2005</i>	0.084 (0.011)	0.196 (0.010)	0.860 (0.008)	0.092 (0.008)	0.196 (0.009)	0.856 (0.006)
<i>M2005 (DB)</i>	0.071 (0.011)	0.134 (0.012)	0.898 (0.009)	0.089 (0.008)	0.109 (0.009)	0.901 (0.007)
<i>M2005 (IDB)</i>	0.064 (0.012)	0.113 (0.009)	0.911 (0.006)	0.094 (0.010)	0.094 (0.005)	0.906 (0.006)

Table 34 – Results using score normalization - GREYC-Web (Passwords).

Biometric system	T-Norm			Z-Norm		
	FMR	FNMR	Acc (balanced)	FMR	FNMR	Acc (balanced)
<i>Self-Detector</i>	0.410 (0.016)	0.149 (0.013)	0.720 (0.008)	0.288 (0.016)	0.289 (0.029)	0.711 (0.012)
<i>Self-Detector (Sliding)</i>	0.414 (0.016)	0.097 (0.010)	0.744 (0.010)	0.240 (0.021)	0.329 (0.026)	0.716 (0.013)
<i>Self-Detector (Growing)</i>	0.418 (0.019)	0.127 (0.015)	0.727 (0.012)	0.347 (0.026)	0.263 (0.030)	0.695 (0.012)
<i>Self-Detector (Usage Control)</i>	0.436 (0.016)	0.095 (0.011)	0.735 (0.009)	0.272 (0.022)	0.302 (0.027)	0.713 (0.013)
<i>Self-Detector (Usage Control R)</i>	0.430 (0.014)	0.093 (0.010)	0.738 (0.009)	0.256 (0.022)	0.314 (0.028)	0.715 (0.014)
<i>Self-Detector (Usage Control S)</i>	0.411 (0.014)	0.103 (0.009)	0.743 (0.008)	0.177 (0.012)	0.401 (0.027)	0.711 (0.013)
<i>Self-Detector (Usage Control 2)</i>	0.411 (0.015)	0.106 (0.011)	0.742 (0.009)	0.143 (0.013)	0.476 (0.021)	0.691 (0.010)
<i>M2005</i>	0.279 (0.012)	0.259 (0.016)	0.731 (0.012)	0.274 (0.035)	0.277 (0.014)	0.724 (0.023)
<i>M2005 (DB)</i>	0.276 (0.014)	0.152 (0.015)	0.786 (0.011)	0.147 (0.024)	0.366 (0.019)	0.743 (0.011)
<i>M2005 (IDB)</i>	0.277 (0.012)	0.138 (0.012)	0.792 (0.008)	0.149 (0.021)	0.313 (0.018)	0.769 (0.012)
Biometric system	F-Norm			Adaptive F-Norm		
	FMR	FNMR	Acc (balanced)	FMR	FNMR	Acc (balanced)
<i>Self-Detector</i>	0.398 (0.016)	0.173 (0.004)	0.715 (0.006)	0.439 (0.011)	0.121 (0.003)	0.720 (0.006)
<i>Self-Detector (Sliding)</i>	0.352 (0.024)	0.192 (0.007)	0.728 (0.012)	0.410 (0.017)	0.108 (0.009)	0.741 (0.010)
<i>Self-Detector (Growing)</i>	0.494 (0.021)	0.116 (0.004)	0.695 (0.010)	0.520 (0.013)	0.070 (0.003)	0.705 (0.007)
<i>Self-Detector (Usage Control)</i>	0.405 (0.024)	0.159 (0.007)	0.718 (0.011)	0.450 (0.018)	0.094 (0.007)	0.728 (0.010)
<i>Self-Detector (Usage Control R)</i>	0.378 (0.024)	0.174 (0.006)	0.724 (0.011)	0.432 (0.019)	0.099 (0.007)	0.735 (0.011)
<i>Self-Detector (Usage Control S)</i>	0.271 (0.020)	0.278 (0.011)	0.726 (0.009)	0.368 (0.016)	0.141 (0.009)	0.745 (0.009)
<i>Self-Detector (Usage Control 2)</i>	0.210 (0.016)	0.365 (0.012)	0.712 (0.010)	0.362 (0.016)	0.147 (0.012)	0.746 (0.010)
<i>M2005</i>	0.246 (0.030)	0.320 (0.024)	0.717 (0.026)	0.320 (0.026)	0.230 (0.014)	0.725 (0.019)
<i>M2005 (DB)</i>	0.131 (0.022)	0.379 (0.017)	0.745 (0.015)	0.249 (0.019)	0.177 (0.018)	0.787 (0.014)
<i>M2005 (IDB)</i>	0.116 (0.019)	0.348 (0.008)	0.768 (0.012)	0.251 (0.017)	0.162 (0.012)	0.794 (0.012)

Table 35 – Results using score normalization - McGill.

Biometric system	T-Norm			Z-Norm		
	FMR	FNMR	Acc (balanced)	FMR	FNMR	Acc (balanced)
<i>Self-Detector</i>	0.154 (0.016)	0.608 (0.053)	0.619 (0.023)	0.156 (0.033)	0.545 (0.084)	0.649 (0.026)
<i>Self-Detector (Sliding)</i>	0.180 (0.031)	0.434 (0.069)	0.693 (0.034)	0.340 (0.086)	0.273 (0.130)	0.694 (0.034)
<i>Self-Detector (Growing)</i>	0.181 (0.038)	0.501 (0.100)	0.659 (0.038)	0.616 (0.130)	0.139 (0.105)	0.622 (0.018)
<i>Self-Detector (Usage Control)</i>	0.243 (0.033)	0.326 (0.068)	0.715 (0.027)	0.458 (0.115)	0.209 (0.121)	0.667 (0.023)
<i>Self-Detector (Usage Control R)</i>	0.196 (0.035)	0.382 (0.076)	0.711 (0.029)	0.400 (0.104)	0.230 (0.126)	0.685 (0.027)
<i>Self-Detector (Usage Control S)</i>	0.203 (0.036)	0.389 (0.078)	0.704 (0.024)	0.161 (0.039)	0.399 (0.121)	0.720 (0.044)
<i>Self-Detector (Usage Control 2)</i>	0.210 (0.024)	0.357 (0.050)	0.717 (0.024)	0.154 (0.028)	0.322 (0.127)	0.762 (0.053)
Biometric system	F-Norm			Adaptive F-Norm		
	FMR	FNMR	Acc (balanced)	FMR	FNMR	Acc (balanced)
<i>Self-Detector</i>	0.106 (0.017)	0.560 (0.019)	0.667 (0.013)	0.072 (0.019)	0.663 (0.019)	0.632 (0.006)
<i>Self-Detector (Sliding)</i>	0.171 (0.024)	0.283 (0.035)	0.773 (0.019)	0.172 (0.091)	0.389 (0.063)	0.720 (0.022)
<i>Self-Detector (Growing)</i>	0.495 (0.053)	0.110 (0.027)	0.698 (0.021)	0.603 (0.079)	0.169 (0.048)	0.614 (0.019)
<i>Self-Detector (Usage Control)</i>	0.304 (0.042)	0.208 (0.034)	0.744 (0.022)	0.289 (0.096)	0.313 (0.056)	0.699 (0.028)
<i>Self-Detector (Usage Control R)</i>	0.242 (0.034)	0.246 (0.036)	0.756 (0.021)	0.225 (0.093)	0.358 (0.060)	0.708 (0.024)
<i>Self-Detector (Usage Control S)</i>	0.107 (0.014)	0.406 (0.046)	0.744 (0.024)	0.074 (0.014)	0.563 (0.038)	0.681 (0.015)
<i>Self-Detector (Usage Control 2)</i>	0.076 (0.007)	0.345 (0.033)	0.789 (0.017)	0.049 (0.011)	0.484 (0.037)	0.734 (0.015)

Table 36 – Results using score normalization - WISDM 1.1.

Biometric system	T-Norm			Z-Norm		
	FMR	FNMR	Acc (balanced)	FMR	FNMR	Acc (balanced)
<i>Self-Detector</i>	0.242 (0.030)	0.280 (0.054)	0.739 (0.020)	0.182 (0.075)	0.346 (0.113)	0.736 (0.029)
<i>Self-Detector (Sliding)</i>	0.336 (0.046)	0.132 (0.066)	0.766 (0.024)	0.297 (0.145)	0.282 (0.133)	0.710 (0.019)
<i>Self-Detector (Growing)</i>	0.282 (0.031)	0.132 (0.055)	0.793 (0.019)	0.371 (0.161)	0.230 (0.121)	0.700 (0.032)
<i>Self-Detector (Usage Control)</i>	0.316 (0.038)	0.112 (0.036)	0.786 (0.020)	0.310 (0.147)	0.267 (0.129)	0.711 (0.031)
<i>Self-Detector (Usage Control R)</i>	0.326 (0.042)	0.128 (0.061)	0.773 (0.023)	0.288 (0.139)	0.282 (0.122)	0.715 (0.025)
<i>Self-Detector (Usage Control S)</i>	0.312 (0.022)	0.174 (0.034)	0.757 (0.012)	0.170 (0.074)	0.332 (0.112)	0.749 (0.025)
<i>Self-Detector (Usage Control 2)</i>	0.345 (0.053)	0.142 (0.050)	0.756 (0.018)	0.200 (0.096)	0.322 (0.127)	0.739 (0.021)
Biometric system	F-Norm			Adaptive F-Norm		
	FMR	FNMR	Acc (balanced)	FMR	FNMR	Acc (balanced)
<i>Self-Detector</i>	0.038 (0.006)	0.412 (0.030)	0.775 (0.014)	0.036 (0.007)	0.418 (0.032)	0.773 (0.014)
<i>Self-Detector (Sliding)</i>	0.041 (0.009)	0.312 (0.018)	0.824 (0.009)	0.060 (0.013)	0.263 (0.023)	0.838 (0.009)
<i>Self-Detector (Growing)</i>	0.068 (0.011)	0.252 (0.025)	0.840 (0.014)	0.059 (0.012)	0.246 (0.023)	0.847 (0.012)
<i>Self-Detector (Usage Control)</i>	0.044 (0.010)	0.274 (0.018)	0.841 (0.011)	0.057 (0.011)	0.257 (0.019)	0.843 (0.009)
<i>Self-Detector (Usage Control R)</i>	0.041 (0.009)	0.316 (0.017)	0.821 (0.008)	0.057 (0.012)	0.274 (0.019)	0.835 (0.008)
<i>Self-Detector (Usage Control S)</i>	0.033 (0.008)	0.347 (0.018)	0.810 (0.008)	0.047 (0.010)	0.303 (0.021)	0.825 (0.008)
<i>Self-Detector (Usage Control 2)</i>	0.024 (0.007)	0.354 (0.016)	0.811 (0.007)	0.043 (0.011)	0.291 (0.022)	0.833 (0.009)

Table 37 – Results using score normalization - WISDM 2.0.

Biometric system	T-Norm			Z-Norm		
	FMR	FNMR	Acc (balanced)	FMR	FNMR	Acc (balanced)
<i>Self-Detector</i>	0.200 (0.038)	0.206 (0.071)	0.797 (0.018)	0.231 (0.052)	0.264 (0.044)	0.753 (0.009)
<i>Self-Detector (Sliding)</i>	0.378 (0.042)	0.060 (0.007)	0.781 (0.020)	0.404 (0.069)	0.210 (0.049)	0.693 (0.012)
<i>Self-Detector (Growing)</i>	0.215 (0.034)	0.179 (0.085)	0.803 (0.029)	0.457 (0.072)	0.198 (0.050)	0.672 (0.012)
<i>Self-Detector (Usage Control)</i>	0.255 (0.031)	0.111 (0.042)	0.817 (0.014)	0.409 (0.070)	0.210 (0.052)	0.690 (0.011)
<i>Self-Detector (Usage Control R)</i>	0.324 (0.040)	0.073 (0.018)	0.801 (0.015)	0.388 (0.069)	0.213 (0.052)	0.700 (0.010)
<i>Self-Detector (Usage Control S)</i>	0.314 (0.049)	0.100 (0.023)	0.793 (0.015)	0.241 (0.054)	0.243 (0.046)	0.758 (0.008)
<i>Self-Detector (Usage Control 2)</i>	0.490 (0.031)	0.045 (0.007)	0.733 (0.014)	0.293 (0.060)	0.229 (0.051)	0.739 (0.009)
Biometric system	F-Norm			Adaptive F-Norm		
	FMR	FNMR	Acc (balanced)	FMR	FNMR	Acc (balanced)
<i>Self-Detector</i>	0.071 (0.006)	0.300 (0.007)	0.815 (0.004)	0.079 (0.006)	0.268 (0.008)	0.826 (0.005)
<i>Self-Detector (Sliding)</i>	0.077 (0.010)	0.192 (0.006)	0.866 (0.006)	0.103 (0.012)	0.166 (0.004)	0.866 (0.006)
<i>Self-Detector (Growing)</i>	0.127 (0.013)	0.151 (0.004)	0.861 (0.006)	0.134 (0.012)	0.134 (0.006)	0.866 (0.006)
<i>Self-Detector (Usage Control)</i>	0.084 (0.010)	0.188 (0.003)	0.864 (0.005)	0.105 (0.012)	0.158 (0.004)	0.868 (0.005)
<i>Self-Detector (Usage Control R)</i>	0.077 (0.009)	0.194 (0.004)	0.865 (0.005)	0.098 (0.010)	0.161 (0.005)	0.871 (0.004)
<i>Self-Detector (Usage Control S)</i>	0.065 (0.007)	0.241 (0.004)	0.847 (0.004)	0.087 (0.007)	0.195 (0.006)	0.859 (0.006)
<i>Self-Detector (Usage Control 2)</i>	0.045 (0.005)	0.240 (0.006)	0.858 (0.004)	0.071 (0.008)	0.193 (0.005)	0.868 (0.005)

SCORE NORMALIZATION - PERFORMANCE PER USER

Additional graphs for the performance of the score normalization procedures per user is shown in this appendix: Figures 43 to 49. There are two types of graphs. The first one is a heat map of the performance per user for each score normalization in the experiments. The stronger the green, the better is the performance (balanced accuracy). The second graph just highlights which score normalization performed best per user.

Figure 43 – Score normalization performance per user - GREYC. The first column presents the balanced accuracy per user (green is better), while the second column highlights (black) the best score normalization for each user.

Figure 44 – Score normalization performance per user - CMU. The first column presents the balanced accuracy per user (green is better), while the second column highlights (black) the best score normalization for each user. The plots for M2005 (IDB) were submitted to (PISANI *et al.*, 2016).

Figure 45 – Score normalization performance per user - GREYC-Web (Logins). The first column presents the balanced accuracy per user (green is better), while the second column highlights (black) the best score normalization for each user.

Figure 46 – Score normalization performance per user - GREYC-Web (Passwords). The first column presents the balanced accuracy per user (green is better), while the second column highlights (black) the best score normalization for each user.

Figure 47 – Score normalization performance per user - McGill. The first column presents the balanced accuracy per user (green is better), while the second column highlights (black) the best score normalization for each user.

Figure 48 – Score normalization performance per user - WISDM 1.1. The first column presents the balanced accuracy per user (green is better), while the second column highlights (black) the best score normalization for each user.

Figure 49 – Score normalization performance per user - WISDM 2.0. The first column presents the balanced accuracy per user (green is better), while the second column highlights (black) the best score normalization for each user.

COMBINATIONS PERFORMANCE PER USER GROUPING

All graphs for the performance of the combinations per user grouping are shown in this appendix: Figures 50 to 56. These plots consider the combinations of adaptation strategy and classification algorithm for the best threshold values, as discussed in Section 7.3.1. There are 84 combinations for keystroke dynamics datasets and 80 combinations for the accelerometer-based gait biometrics datasets (single gallery). Each one is divided into five parts, corresponding to each user grouping of the user cross-validation for biometric data streams methodology, as described in Section 3.2. Heatmaps on the left present the balanced accuracy on the test data (the greener the better). The color scale is the same for all heatmaps, meaning that the average performance in GREYC-Web is higher than in CMU, for example. The plots on the right just highlight the best combination. When a draw occurs, more than one combination is highlighted.

Figure 50 – Performance of combinations per user grouping - GREYC.

Figure 51 – Performance of combinations per user grouping - CMU.

Figure 52 – Performance of combinations per user grouping - GREYC-Web (Logins).

Figure 53 – Performance of combinations per user grouping - GREYC-Web (Passwords).

Figure 54 – Performance of combinations per user grouping - McGill.

Figure 55 – Performance of combinations per user grouping - WISDM 1.1.

Figure 56 – Performance of combinations per user grouping - WISDM 2.0.

COMBINATIONS PERFORMANCE PER USER

All graphs for the combinations performance per user are shown in this appendix: Figures 57 to 63. These plots consider the combinations of adaptation strategy and classification algorithm for the best threshold values, as discussed in Section 7.3.2. There are 84 combinations for keystroke dynamics datasets and 80 combinations for the accelerometer-based gait biometrics datasets (single gallery). Each one is divided into five parts, corresponding to each user grouping of the user cross-validation for biometric data streams methodology, as described in Section 3.2. Heatmaps in the top present the balanced accuracy on the test data (the greener the better). The color scale is the same for all heat maps, meaning that the average performance in GREYC-Web is higher than in CMU, for example. The plots in the bottom just highlight the best combination. When a draw occurs, more than one combination is highlighted.

(a) Balanced accuracy.

(b) Best combination.

Figure 57 – Performance of combinations per user - GREYC.

(a) Balanced accuracy.

(b) Best combination.

Figure 58 – Performance of combinations per user - CMU.

(a) Balanced accuracy.

(b) Best combination.

Figure 59 – Performance of combinations per user - GREYC-Web (Logins).

(a) Balanced accuracy.

(b) Best combination.

Figure 60 – Performance of combinations per user - GREYC-Web (Passwords).

(a) Balanced accuracy.

(b) Best combination.

Figure 61 – Performance of combinations per user - McGill.

(a) Balanced accuracy.

(b) Best combination.

Figure 62 – Performance of combinations per user - WISDM 1.1.

(a) Balanced accuracy.

(b) Best combination.

Figure 63 – Performance of combinations per user - WISDM 2.0.