UNIVERSIDADE DE SÃO PAULO
ESCOLA DE ARTES, CIÊNCIAS E HUMANIDADES
PROGRAMA DE PÓS-GRADUAÇÃO EM SISTEMAS DE INFORMAÇÃO

RICARDO ALVES FEITOSA

**Estudo de representações multidimensionais para segmentação das fases do
gesto**

São Paulo

2018

RICARDO ALVES FEITOSA

Estudo de representações multidimensionais para segmentação das fases do gesto

Dissertação apresentada à Escola de Artes, Ciências e Humanidades da Universidade de São Paulo para obtenção do título de Mestre em Ciências pelo Programa de Pós-graduação em Sistemas de Informação.

Área de concentração: Metodologia e Técnicas da Computação

Versão corrigida contendo as alterações solicitadas pela comissão julgadora em 17 de abril de 2018. A versão original encontra-se em acervo reservado na Biblioteca da EACH-USP e na Biblioteca Digital de Teses e Dissertações da USP (BDTD), de acordo com a Resolução CoPGr 6018, de 13 de outubro de 2011.

Orientador: Profa. Dra. Sarajane Marques Peres

Coorientador: Prof. Dr. Clodoaldo Aparecido de Moraes Lima

São Paulo

2018

Autorizo a reprodução e divulgação total ou parcial deste trabalho, por qualquer meio convencional ou eletrônico, para fins de estudo e pesquisa, desde que citada a fonte.

CATALOGAÇÃO-NA-PUBLICAÇÃO

(Universidade de São Paulo. Escola de Artes, Ciências e Humanidades. Biblioteca)
CRB 8 - 4936

Feitosa, Ricardo Alves

Estudo de representações multidimensionais para segmentação das fases do gesto / Ricardo Alves Feitosa ; orientadora, Sarajane Marques Peres ; coorientador, Clodoaldo Aparecido de Moraes Lima. – 2018.

104 f. : il

Dissertação (Mestrado em Ciências) - Programa de Pós-Graduação em Sistemas de Informação, Escola de Artes, Ciências e Humanidades, Universidade de São Paulo.

Versão corrigida.

1. Inteligência artificial. 2. Reconhecimento de padrões. 3. Gestos - Análise. 4. Aprendizado computacional. I. Peres, Sarajane Marques, orient. II. Lima, Clodoaldo Aparecido de Moraes, coorient. III. Título

CDD 22.ed.– 006.3

Dissertação de autoria de Ricardo Alves Feitosa, sob o título “**Estudo de representações multidimensionais para segmentação das fases do gesto**”, apresentada à Escola de Artes, Ciências e Humanidades da Universidade de São Paulo, para obtenção do título de Mestre em Ciências pelo Programa de Pós-graduação em Sistemas de Informação, na área de concentração Metodologia e Técnicas da Computação, aprovada em 17 de abril de 2018 pela comissão julgadora constituída pelos doutores:

Profa. Dra. Sarajane Marques Peres

Escola de Artes, Ciências e Humanidades - Universidade de São Paulo
Presidente

Prof. Dr. Emílio Del Moral Hernandez

Escola Politécnica - Universidade de São Paulo

Prof. Dr. Leandro Augusto da Silva

Universidade Presbiteriana Mackenzie

Prof. Dr. Ivandre Paraboni

Escola de Artes, Ciências e Humanidades - Universidade de São Paulo

Agradecimentos

Em primeiro lugar agradeço a Deus por me proporcionar os meios para que eu alcance os meus objetivos. Apesar do caminho ser árduo, sempre me sustentou.

A Universidade de São Paulo e a todos os professores, profissionais e alunos que contribuíram com a minha formação. Em especial a minha orientadora, a Profa. Dra. Sarajane Marques Peres, que teve muita paciência, sempre mostrando o caminho com entusiasmo e ética, com amor a profissão e a quem admiro desde a época da graduação. Muito obrigado por tudo.

A Camila, que sempre está ao meu lado, escutando minhas reclamações e que hoje me entende melhor que do eu.

A toda minha família. Em especial a minha mãe, minha primeira grande professora e que sempre está por perto me acolhendo com seu amor.

E aos meus amigos que sempre me ajudaram a rir das adversidades com as quais nos deparamos durante a ótima vida que temos.

Resumo

FEITOSA, Ricardo Alves. **Estudo de representações multidimensionais para segmentação das fases do gesto**. 2018. 104 f. Dissertação (Mestrado em Ciências) – Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, São Paulo, 2018.

Sistemas de análise de gestos têm se destacado por suas contribuições para a interação entre humanos, humanos e máquinas, e humanos e ambiente. Nessa interação, a gesticulação natural é vista como parte do sistema linguístico que suporta a comunicação, e qualquer sistema de informação que objetiva usar interação para suporte à decisão deveria ser capaz de interpretá-la. Essa interpretação pode ser realizada por meio da segmentação das fases do gesto. Para resolver essa tarefa, o estabelecimento de uma representação de dados eficiente para os gestos é um ponto crítico. A representação escolhida e sua associação a técnicas de análise podem ou não favorecer a solução sob implementação. Neste trabalho, formas de representação de gestos são submetidas aos algoritmos de reconhecimento de padrões MLP e SOM para elaborar um ambiente propício à identificação das representações mais discriminantes, quais aspectos as diferentes representações descrevem com eficiência, e como elas podem ser combinadas para melhorar a segmentação das fases do gesto. Para construção das representações multidimensionais são usados aspectos espaciais e temporais combinados com a normalização dos dados e a aplicação do filtro *wavelet* na busca pela representação mais discriminante para o reconhecimento das fases do gesto. Ambos os algoritmos alcançaram bons resultados com o uso dos aspectos temporais. O MLP conseguiu classificar todas as fases do gesto em configurações de representação contendo dados sobre todos os membros monitorados. O SOM apresentou boa capacidade para formar grupos contendo dados de uma mesma fase do gesto mesmo com o uso de poucas características na construção da representação, porém não foi possível identificar a proposta de uma nova fase do gesto com o aprendizado não supervisionado.

Palavras-chaves: Representação de Gestos. Segmentação das Fases do Gesto. Reconhecimento de Padrões. Aprendizado de Máquina. Perceptron Multicamadas. Mapas Auto Organizáveis.

Abstract

FEITOSA, Ricardo Alves. **Study of multidimensional representations for the gesture phases segmentation**. 2018. 104 p. Dissertation (Master of Science) – School of Arts, Sciences and Humanities, University of São Paulo, São Paulo, 2018.

Gestures analysis systems have stood out for their contributions to the interaction between humans, humans and machines, and humans and environments. In this interaction, natural gesticulation is seen as part of a linguistic system that supports the communication, and all information systems aiming at the use of such an interaction in making decisions should be able to interpret it. Such an interpretation can be carried out through the gesture phases segmentation. In order to solve this task, the establishment of an efficient data representation for gestures is a critical issue. The chosen representation as well as its combination with techniques for analysis can or can not favor the solution being developed. In this work, different forms representation for gestures are applied to pattern recognition algorithms MLP and SOM to create an adequate environment to identify the more discriminative representations, which aspect the different representations describe with more efficiency, and how they can be combined in order to improve gesture phases segmentation. To construct the multidimensional representations we use spatial and temporal aspects combined with the normalization of the data and the application of the wavelet filter in the search for the most discriminating representation for the recognition of the gesture phases. Both algorithms achieved good results with the use of temporal aspects. MLP was able to classify all gesture phases using representation settings containing data about all monitored members. SOM presented good ability to form groups containing data of the same gesture phase even with the use of few characteristics in the construction of the representation, but it was not possible to identify the proposal of a new gesture phase with unsupervised learning.

Keywords: Representation of Gestures. Segmentation of Gesture Phases. Pattern Recognition. Machine Learning. Multilayer Perceptron. Self Organizing Maps.

Lista de figuras

Figura 1 – Ilustração de unidade gestual.	18
Figura 2 – Ilustração das fases do gesto de uma unidade gestual.	28
Figura 3 – Esquema de Teoria dos Gestos.	30
Figura 4 – Modelo de RNA.	38
Figura 5 – Arquitetura gráfica de um MLP com duas camadas ocultas.	39
Figura 6 – Modelo de Kohonen.	43
Figura 7 – Matriz-U gerada sobre a base de dados Iris.	46
Figura 8 – Sinal referente à velocidade escalar da mão direita durante a realização de um gesto: original e com filtro	50
Figura 9 – Sinal original - Velocidade escalar da mão direita ao longo dos <i>frames</i> . .	52
Figura 10 – Sinal filtrado - Velocidade escalar da mão direita ao longo dos <i>frames</i> . .	52
Figura 11 – Exemplo de geração de janela retangular de dimensão 10.	53
Figura 12 – Esquema da implementação <i>feedforwardnet</i>	57
Figura 13 – Estratégia fixa: <i>k-fold</i> com $k = 10$	58
Figura 14 – Estratégia cíclica: <i>k-fold</i> com $k = 10$	58
Figura 15 – Organização de testes para o MLP.	60
Figura 16 – Organização de testes para o SOM.	62
Figura 17 – Matriz-U do resultado #1-2.	79
Figura 18 – Matriz-U do resultado #1-8.	80
Figura 19 – Matriz-U do resultado #2-1.	81
Figura 20 – Matriz-U do resultado #2-3.	82
Figura 21 – Matriz-U do resultado #2-4.	82
Figura 22 – Matriz-U do resultado #2-6.	83
Figura 23 – Matriz-U dos resultados #3-1 e #3-2.	84
Figura 24 – Matriz-U dos resultados #3-5 e #3-6.	85
Figura 25 – Matriz-U dos resultados #4-1 e #4-2.	87
Figura 26 – Matriz-U do resultado #4-10.	88
Figura 27 – Matriz-U dos resultados #4-8 e #4-12.	89
Figura 28 – Matriz-U do resultado #4-13.	90
Figura 29 – Matriz-U do resultado #4-3.	90

Figura 30 – Matriz-U dos resultados #4-11 e #4-6.	91
Figura 31 – Matriz-U do resultado #4-16.	92

Lista de tabelas

Tabela 1 – Dados sobre os vídeos que compõe a base de dados; coeficiente de concordância entre os rotuladores humanos.	48
Tabela 2 – Relação entre o formato matricial e os aspectos do gesto nele representados.	64
Tabela 3 – Melhores resultados identificados quanto a acurácia total por vídeo. . .	65
Tabela 4 – Melhores resultados identificados quanto a acurácia total para o vídeo a1.	66
Tabela 5 – Melhores resultados identificados quanto a acurácia por fase para o vídeo a1.	67
Tabela 6 – Melhores resultados identificados quanto ao F1-Score por fase seguindo a ordenação por acurácia total.	68
Tabela 7 – Melhores resultados identificados quanto ao F1-Score para as fases de <i>stroke</i> , descanso e <i>hold</i>	68
Tabela 8 – Acurácias e F1-Scores dos melhores resultados identificados quanto a acurácia total para modelos de representação com menos de 18 aspectos para o gesto.	69
Tabela 9 – Acurácias e F1-Scores dos melhores resultados identificados quanto a acurácia total para modelos de representação com um ou dois aspectos para o gesto.	70
Tabela 10 – Precisão e revocação dos melhores resultados identificados quanto ao F1-Score para as fases de <i>stroke</i> , descanso e <i>hold</i>	72
Tabela 11 – Acurácias e F1-Scores dos melhores resultados identificados quanto a acurácia total para representações com uma ou duas características e janela de tamanho 1.	73
Tabela 12 – Acurácias e F1-Scores dos melhores resultados identificados quanto a acurácia total para as demais variações de configuração do MLP. . . .	74
Tabela 13 – Melhores resultados identificados quanto ao erro de quantização por representação temporal.	77
Tabela 14 – Melhores resultados identificados quanto aos erros de quantização e topológicos isoladamente.	78

Tabela 15 – Matriz de confusão entre as fases de cada dado definida pelo especialista e a fase mais representativa associada aos BMUs para os resultados #1-1 e #1-11.	78
Tabela 16 – Melhores resultados identificados quanto aos erros de quantização para mais de um aspecto temporal.	80
Tabela 17 – Melhores resultados identificados quanto aos erros de quantização para tamanho de mapa diferente de 20x20 e resultados com janelas de 20 e 40 mesmo em mapas 20x20.	83
Tabela 18 – Matriz de confusão entre as fases de cada dado definida pelo especialista e a fase mais representativa associada aos BMUs para os resultados #3-3 e #3-4.	84
Tabela 19 – Melhores resultados identificados para cada caso de teste quanto aos erros de quantização para modelos formados com aspectos espaciais. . .	86

Lista de abreviaturas e siglas

Acc	Acurácia.
AspG	Aspectos do gesto.
BMU	<i>Best Matching Unit.</i>
DivDados	Método de divisão dos dados entre treinamento, validação e testes.
DWT	Transformada Wavelet Discreta – <i>Discrete Wavelet Transform.</i>
Ffnet: L-M	<i>Feedforwardnet</i> com Levenberg-Marquardt.
FSc	F1-Score.
GD Adp+Mon	Gradiente descendente com taxa de aprendizado adaptativa e <i>momentum</i> .
HMM	Modelo Oculto de Markov – <i>Hidden Markov Models.</i>
Jal	Tamanho da janela de dados.
MLP	Perceptron Multicamadas – <i>Multilayer Perceptron.</i>
MS	Microsoft.
NeOc	Número de neurônios na camada escondida.
Norm	Identificação de normalização dos dados.
Qt_Erro	Erro de quantização.
RNA	Redes Neurais Artificiais.
SOM	Mapas Auto Organizáveis – <i>Self Organizing Maps.</i>
SVM	Máquinas de Vetores de Suporte – <i>Support Vector Machines.</i>
TamMapa	Tamanho do mapa do SOM.
TipoSOM	Tipo do algoritmo SOM usado para correção de erro.
Tp_Erro	Erro topológico.
TxA	Taxa de aprendizado.

Lista de símbolos

(x, y, z)	Coordenadas espaciais.
α	Taxa de aprendizado.
π	Letra grega pi adotada para representar a dízima periódica 3,1415...
$\Re$	Domínio dos números reais.
b	Índice do BMU.
C	Conjunto de classes.
cA	Coeficiente de aproximação.
cD	Coeficiente de detalhe.
d_t	Cálculo da diferença entre as coordenadas de um membro do corpo humano entre os <i>frames</i> f e $f + 1$.
$dlem$	Número de dados de entrada no SOM.
E_q	Erro de quantização.
E_t	Erro de topológico.
$e_j(n)$	Erro do neurônio j no instante n .
EQM	Erro Quadrático Médio.
f	<i>Frames</i> .
G	Conjunto de grupos.
$G(z)$	Filtro passa-baixa.
h	Função de vizinhança centrada ao redor do neurônio vencedor.
$H(z)$	Filtro passa-alta.
j	Janela retangular de dados.
K	Número de <i>folds</i> .

$K\alpha$	Coefficiente de concordância.
mpd	Número de unidades do mapa SOM dividido pelo número de dados usados para treinamento do SOM.
$mUnits$	Número de unidade do mapa SOM.
N	Número de entradas.
n	Iteração atual.
R	Conjunto de rótulos.
r	Raio de vizinhança entre os neurônios.
t	Época de treinamento.
V	Vídeo composto por uma sequência de <i>frames</i> .
v_j	Soma ponderada das entradas mais o bias para o MLP.
w	Vetor de pesos do neurônio.
y_j	Saída do neurônio para o MLP.

Sumário

1	Introdução	16
1.1	<i>Apresentação do problema</i>	20
1.2	<i>Objetivos</i>	21
1.3	<i>Método</i>	23
1.4	<i>Organização deste documento</i>	26
2	Segmentação das fases do gesto	27
2.1	<i>Teoria de Gestos</i>	27
2.2	<i>Segmentação das fases do gesto</i>	30
2.2.1	Segmentação manual	30
2.2.2	Segmentação automática	31
3	Representação de dados gestuais	34
3.1	<i>Representações orientadas à informação espacial</i>	35
3.2	<i>Representações orientadas à informação temporal</i>	36
3.3	<i>Estratégias de combinação de características</i>	37
4	Aprendizado de Máquina	38
4.1	<i>Aprendizado de Máquina Supervisionado com MLP</i>	38
4.2	<i>Aprendizado de Máquina Não Supervisionado com SOM</i>	42
5	Experimentação, resultados e discussões	47
5.1	<i>Estratégia de experimentação</i>	47
5.1.1	Base de dados	47
5.1.2	Pré-processamento	49
5.1.3	Construção de representações multidimensionais	53
5.1.4	Descrição dos experimentos	55
5.2	<i>Experimentos com MLP</i>	63
5.2.1	Aspectos temporais	64
5.2.2	Aspectos espaciais	71
5.3	<i>Experimentos com SOM</i>	76
5.3.1	Aspectos temporais	77

5.3.2	Aspectos espaciais	86
6	Conclusões	93
6.1	<i>Trabalhos publicados</i>	95
6.2	<i>Limitações e trabalhos futuros</i>	96
	Referências¹	99

¹ De acordo com a Associação Brasileira de Normas Técnicas. NBR 6023.

1 Introdução

A análise de gestos é uma tarefa comumente realizada em contextos nos quais se pretende compreender o significado de um gesto e, a partir dessa compreensão, fazer uso desse significado em algum tipo de tomada de decisão. Considerando a automação da tarefa de análise de gestos, tem-se uma aplicação para a área de reconhecimento de padrões. O padrão a ser reconhecido depende do contexto sob o qual o gesto está sendo analisado, e um desses contextos é o uso e a compreensão dos gestos como um elemento do sistema linguístico e da relação dele com outros elementos desse sistema. Na comunicação ou interação entre pessoas, a relação entre gestos e língua se estabelece de forma natural, seja na gesticulação natural ou na composição de uma língua gestual (por exemplo, uma língua de sinais). Num contexto no qual um sistema de informação necessita mediar a interação entre humanos, ou estabelecer a interação entre humanos e máquinas ou ambientes, faz-se útil que os gestos sejam considerados como elementos que carregam informação. Um dos exemplos de interação entre humanos e máquinas é o realizado por dispositivos com telas sensíveis a toques (BARALDI; BIMBO; LANDUCCI, 2008). Considerados como sistemas com um vocabulário finito, é uma classe de algoritmos que tem sido explorada também para o reconhecimento de línguas de sinais, como em Ong e Ranganath (2005).

Porém, quando se trata de gesticulação natural, não há vocabulário finito e bem definido para ser analisado. Os gestos são incertos e dependem da diversidade local e cultural do contexto no qual são executados (KIM; SONG; KIM, 2007). Nesse cenário, os trabalhos desenvolvidos estudam os movimentos de diferentes partes do corpo para analisar o comportamento humano. Por exemplo o trabalho de Drosou et al. (2012), que explora a análise dos gestos para o uso em biometria.

O contexto de análise de gestos explorado nesse trabalho está delimitado pela gesticulação natural, embutida no discurso executado durante um tipo de comunicação. Dentro desse contexto, a Teoria de Gestos permite diferentes tipos de análises, considerando diferentes tipos de padrões de gestos. Uma destas análises é aquela estudada por Kendon (1980), McNeill (1992) e McNeill (2015), que propõem que um gesto está estruturado em fases, e permite implementar uma análise de como os gestos estão estruturados no contexto em que se manifestam. De acordo com esses autores, as seguintes fases podem ser consideradas:

- *preparação*: fase que compreende o momento em que os membros do corpo, por exemplo as mãos, se deslocam de uma posição de descanso (um período com ausência de gesticulação) em direção ao *stroke*;
- *pre-stroke hold*: fase que representa uma pausa no movimento das mãos entre a *preparação* e o *stroke*;
- *stroke*: é o gesto em si, ou o período da gesticulação que carrega alguma informação que tem um significado importante no contexto de execução;
- *post-stroke hold*: fase que representa uma pausa no movimento das mãos entre o *stroke* e a *retração*;
- *retração*: fase que compreende o tempo que a mão leva para voltar à posição de descanso (*rest*);
- *hold*: uma fase inserida nesse trabalho para representar um período em que há ausência de movimento, porém com presença de informação ou significado, e geralmente ocorre entre a *preparação* e a *retração*.

Ainda seguindo o disposto na Teoria de Gestos, o período compreendido entre o momento em que a mão sai da posição de *rest* e retorna para ela mesma é chamado de unidade gestual (KITA; GIJN; HULST, 1998), e a frase gestual (*gesture phrase* ou *G-phrase*) é um segundo elemento dentro de uma unidade gestual composta por uma ou mais fases do gesto com significado dentro do contexto de gesticulação. O *stroke* é a única fase obrigatória dentro da unidade gestual. A figura 1 exemplifica as fases do gesto distribuindo-as em uma configuração possível ao longo de uma unidade gestual que se manifesta na gesticulação natural realizada por uma pessoa. Para compor esta figura foram escolhidos alguns *frames* de um vídeo no qual a pessoa executa um gesto relacionado a uma ação (“torcer”, nesse caso). Cada *frame* foi retirado do período no vídeo referente a uma fase do gesto.

Figura 1 – Ilustração de unidade gestual.

Fonte: Madeo, Peres e Lima (2016)

A análise de gestos considerando o escopo de suas fases é conhecido como a segmentação das fases do gesto, e é por meio da identificação das fases e posterior relação dessas fases com outros elementos linguísticos (fala, por exemplo) que se torna possível o estudo do significado dos gestos e da gesticulação natural dentro do discurso executado durante um período de comunicação. A automação desta tarefa vem recebendo notória atenção nos últimos anos. Entre os estudos recentes correlatos a essa tarefa está o trabalho de Wagner et al. (2014a), que focou no problema de segmentação das fases do gesto, mais especificamente na identificação da unidade gestual. Assim como o trabalho de Yin e Davis (2014) que tratou do reconhecimento de todo o gesto, a relevância de cada fase e a utilização da segmentação para suportar outras tarefas, utilizando *Hidden Markov Models* (HMMs) para identificar fases do gesto de vídeos obtidos com a câmera e os sensores do MS Kinect¹ com um total de 840 gestos. Também há o trabalho de Gebre, Wittenburg e Lenkiewicz (2012) onde o foco foi a identificação de quais *frames* de vídeo pertenciam à fase de *stroke* a partir da posição e velocidade das mãos. Há ainda o trabalho de Ramakrishnan e Neff (2013), que utilizou heurísticas e Máquinas de Vetores de Suporte (*Support Vector Machines* – SVM) para identificar as fases do gesto a partir de vídeos capturados de diferentes usuários que foram filmados em diferentes condições de humor.

¹ <https://developer.microsoft.com/pt-br/windows/kinect>.

Há ainda os trabalhos de Ong e Ranganath (2005) no reconhecimento de línguas de sinais e Drosou et al. (2012), na análise dos gestos para o uso em biometria. A segmentação das fases do gesto é uma técnica utilizada para análise dos gestos e que pode beneficiar profissionais em diferentes tarefas nesse contexto.

O desenvolvimento de soluções para automação da segmentação das fases do gesto depara-se com a necessidade da escolha de uma representação computacional para os dados sob análise – os gestos. Essa é uma escolha tão crítica quanto a escolha das técnicas de análise de dados que serão aplicadas. Ainda, o domínio sob o qual os gestos são interpretados traz particularidades que devem estar bem representadas para que técnicas computacionais sejam capazes de tratá-las adequadamente. Dentro de um sistema linguístico, a execução e interpretação de um gesto pode considerar diferentes aspectos, sendo os mais comuns: aspectos espaciais que embutem informação sobre forma, amplitude ou direção; aspectos temporais que incorporam tanto a informação sobre velocidade e aceleração, quanto informação sobre frequência e periodicidade; aspectos estruturais que carregam a informação estrutural e de relação entre os gestos e suas partes constituintes (SMITH, 2011; DIETTERICH, 2002), ou entre os gestos e os demais elementos do sistema.

O presente trabalho se configura como uma extensão do estudo realizado por Madeo, Peres e Lima (2016). Nele, foi utilizada uma representação de aspectos temporais do gesto para implementar a segmentação das suas fases usando SVM, em sua versão clássica e em versões alternativas que incorporam processamento temporal. Outros autores exploraram o mesmo problema ou problemas similares, fazendo uso de diferentes abordagens para representar os gestos e diferentes técnicas para implementar a segmentação das fases. Uma revisão de literatura neste tópico é apresentada por Madeo, Wagner e Peres (2013).

Em um primeiro momento é possível afirmar que especialistas da área de linguística são os principais beneficiados com a análise baseada na segmentação das fases do gesto. Mas várias outras aplicações podem ser delineadas, aplicações essas tão sofisticadas quanto as já citadas. Dentre as aplicações com vocabulário finito, pode-se citar ainda os trabalhos de Spano, Cisternino e Paternò (2012), que utiliza traços básicos de um gesto para compor sub-gestos e, com o uso de operadores, reconhecer gestos mais complexos. Nesse grupo também estão os trabalhos realizados por Lee, Yeh e Hsiao (2016) e Madeo et al. (2010), na qual são exploradas as configurações e movimentos das mãos. Outro trabalho recente com aplicação da segmentação das fases do gesto é o de Freitas et al. (2014), no qual expressões faciais são identificadas.

Outro exemplo, agora sobre gesticulação natural, é o realizado por Salem et al. (2012). Em seu trabalho as fases do gesto são utilizadas para analisar os gestos humanos e desenvolver uma gesticulação natural para robôs, além de investigar a utilização dessa gesticulação na interação dos robôs com os humanos.

Na gesticulação natural, provavelmente, é onde se encontram grande parte dos desafios na análise dos gestos por conta da infinidade de gestos e nuances que os caracterizam. Tanto que surgem trabalhos como o de Jacob e Wachs (2014), no qual as fases do gesto são utilizadas para determinar se um gesto foi intencional ou não.

1.1 Apresentação do problema

Madeo, Peres e Lima (2016) focaram seu trabalho em explorar a segmentação das fases do gesto por meio de técnicas de aprendizado de máquina, concluindo que é necessário um trabalho extenso de ajuste de parâmetros dos algoritmos para que se alcance resultados satisfatórios. Porém, os autores não realizaram uma exploração abrangente de representações para os gestos. Essa lacuna no estudo citado aliada às diferentes formas possíveis de representar os aspectos que definem um gesto constituem o problema que é tratado no presente trabalho. O foco então é a busca por representações multidimensionais discriminantes para os gestos, considerando que seus aspectos espaciais, temporais e estruturais precisam estar representados de forma que a informação relacionada a eles possa ser usada na segmentação das fases, se necessário. Tais representações serão caracterizadas por multidimensionalidade tanto porque incorporaram diferentes aspectos descritivos dos gestos, quanto porque o gesto será tratado na forma de uma série temporal representada por um conjunto de pontos em altas dimensões.

Na literatura encontra-se tanto iniciativas que usam representações que expressam apenas algum aspecto dos gestos, quanto iniciativas que procuram combinar, de forma explícita ou implícita, a informação referente a diferentes aspectos. Contudo, tais estudos delimitam áreas de aplicação específicas, com características próprias, que dificultam a transferência de conhecimento entre diferentes domínios.

O domínio de segmentação das fases do gesto traz desafios para a área de reconhecimento de padrões. Em uma análise simples e empírica poder-se-ia dizer que desde que a gesticulação se dá principalmente com a mão dominante, e que entre as fases

de um gesto percebe-se alterações da intensidade do movimento (*stroke* seria a fase de maior intensidade e o período de descanso, considerando-o como uma fase, a de menor intensidade), então uma representação adequada poderia ser a velocidade escalar da mão direita. Ou, imaginando que o período de descanso ocorre quando as mãos estão abaixo do nível da cintura e a unidade gestual ocorre acima do nível da cintura, o uso da informação referente à coordenada espacial das mãos deveria fazer parte da representação. Entretanto, análises detalhadas mostram que o uso isolado dessas informações gera uma segmentação equivocada pois um *hold*, assim como o descanso, possui ausência de movimento, e o descanso pode ocorrer na posição de braços cruzados acima da cintura. Além disso, já no estudo de Madeo, Peres e Lima (2016) constatou-se que a representação dos gestos precisa considerar um período de tempo, de forma que a análise de *frames* isolados de um vídeo não teria sucesso na representação de fases. Esta constatação também é confirmada de forma empírica, já que a análise de uma imagem (um *frame*) não traz informação suficiente para indicar qual fase do gesto ocorre naquele momento específico ali representado.

Sendo assim, como principal problema tratado por esse trabalho está a definição de como deve ser uma representação de gestos para que ela tenha poder discriminativo para permitir que algoritmos de reconhecimento de padrões produzam bons resultados quando aplicados ao problema de segmentação das fases do gesto.

1.2 Objetivos

A partir dessa análise empírica do problema, das constatações já obtidas por Madeo, Peres e Lima (2016) e também com base no estudo da literatura correlata, pode-se adotar como hipótese nesse trabalho que utilizar uma representação computacional unidimensional, que considere apenas um aspecto descritivo do gesto, ou um momento isolado da gesticulação, não é suficiente para obter bons resultados de segmentação de suas fases. Há, portanto, a necessidade de:

- embutir informações de naturezas diferentes na representação cobrindo a descrição dos aspectos espaciais, temporais e estruturais de um gesto;
- propiciar a representação de um período de tempo.

Partindo desta hipótese, delineia-se como objetivo geral deste trabalho demonstrar que diferentes aspectos dos gestos são importantes para representação adequada das

fases do gesto, e que também se faz necessário o uso de representações que incorporem explicitamente a informação referente a um período, ou janela, de tempo. A demonstração de tais necessidades será realizada por meio da comparação de resultados de segmentação de fases obtidos por algoritmos de reconhecimento de padrões, mais especificamente considerando o paradigma de aprendizado de máquina, a partir do uso de diferentes representações para os gestos.

A partir do tema, hipótese e objetivo geral apresentados para esse trabalho, é possível ainda delinear alguns objetivos específicos:

- verificar qual a influência de cada um dos diferentes aspectos do gesto sobre a capacidade discriminativa das representações adotadas;
- verificar a possibilidade de eliminação do viés especialista da segmentação automática das fases do gesto com o uso de algoritmos de aprendizado de máquina não supervisionados. Algoritmos de aprendizado supervisionado realizam ajustes nos pesos durante as iterações com o uso de uma classificação prévia realizada por um especialista. Sendo assim, o resultado do algoritmo é influenciado por um viés especialista que é absorvido na etapa de treinamento desse tipo de algoritmo;
- verificar a possibilidade de gerar novo conhecimento sobre as fases do gesto, a partir do uso de algoritmos de aprendizado não supervisionado, oferecendo aporte analítico para especialistas da área de linguística que estudam a necessidade de reestruturar as fases do gesto para um esquema diferente daquele apresentado por Kendon (1980), McNeill (1992) e McNeill (2015).

A análise de gestos geralmente exige o desenvolvimento de procedimentos para: (1) identificar os movimentos realizados pelos usuários, com a captação de dados sobre os gestos executados por uma ou mais pessoas a partir do uso de sensores ou outras formas de captação; (2) identificar, representar e reconhecer os movimentos com base em um modelo de análise específico, sendo que esses procedimentos podem ser executados manualmente ou de forma automática, por exemplo, com apoio de algoritmos de reconhecimento de padrões; e (3) avaliação dos resultados. Cada um desses procedimentos pode ser resolvido de diversas formas (KYRIAZIS et al., 2016; OZTURK et al., 2015; SUAREZ; MURPHY, 2012).

As atividades referentes aos procedimentos 2 e 3, quando executadas manualmente por um especialista humano, seguindo a metodologia de divisão de vídeos em *frames* e

usando a imagem do *frame* como informação para tomada de decisão sobre o reconhecimento ou identificação dos movimentos, pode demandar cerca de 2 horas, considerando um vídeo de 1 minuto e a observação apenas das fases do gesto. Se diferentes elementos do contexto precisam ser considerados, por exemplo, se além das fases do gesto, tipos de gestos também precisam ser identificados, o tempo dispendido será ainda maior.

Sendo assim, o conhecimento sobre quais aspectos do gesto são mais discriminantes para construção de uma representação computacional para o gesto, de forma que se obtenha bons resultados na tarefa de reconhecimento de padrões sobre segmentação das fases do gesto, resulta na automação de uma atividade manual que despende um tempo que pode se tornar proibitivo para ser realizada em escala de forma eficiente.

Além da questão de custo em termos de tempo e eficiência, a realização dessa tarefa de forma manual está sujeita à subjetividade do categorizador humano (KITA; GIJN; HULST, 1998), o que implica em problemas de eficácia. Um especialista pode concluir que um *frame* pode representar uma preparação para o *stroke* enquanto outro pode concluir que esse *frame* já representa o *stroke*. Além disso, o cansaço proveniente da execução da tarefa de análise manual por tanto tempo, faz também com que o especialista humano se torne incoerente, de forma que o que foi considerado uma preparação no início de um vídeo, pode, ao final da análise, ser considerado como um *stroke*, pelo mesmo especialista. Assim, a automação da tarefa, considerando a possibilidade de obter bons resultados dado o encontro de boas representações para os gestos, possui potencial para agregar valor aos profissionais que se dedicam à análise de gestos e análise de discursos de forma a reduzir ou eliminar a subjetividade associada à tarefa.

1.3 Método

O alcance dos objetivos delineados para esse trabalho exigiu, em um primeiro momento, um estudo exploratório sobre as alternativas apresentadas na literatura para a construção de representações para os gestos, e em um segundo momento, a aplicação dessas e outras representações combinadas de diferentes maneiras junto a algoritmos de reconhecimento de padrões para que a segmentação das fases do gesto seja executada. Assim, este trabalho se configura como uma pesquisa exploratória, pois objetivou analisar uma série de combinações de características descritivas dos gestos e o impacto delas sobre

o resultado do reconhecimento de padrões, enumerando as combinações que tornam a representação mais discriminante e propiciam a geração dos melhores resultados. E, desde que espera-se auxiliar os profissionais que se dedicam à análise de gestos, esse trabalho é também de natureza aplicada.

A experimentação foi implementada a partir de técnicas de reconhecimento de padrões e executada em um contexto controlado, sobre o qual se tem conhecimento das particularidades da gesticulação e das fases do gesto executado. Sendo assim, esse trabalho se enquadra no gênero de pesquisa empírica, por se basear na observação dos resultados gerados a partir de diferentes combinações de características dos gestos, constituindo a representação do gesto dentro de um contexto conhecido pelos pesquisadores. Porém, não apenas a observação foi utilizada para gerar as conclusões derivadas desse trabalho. A teoria foi necessária para explicar os comportamentos dos algoritmos de reconhecimento de padrões e para avaliá-los, agregando profundidade ao trabalho.

Centrado na objetividade, esse trabalho se desenvolveu seguindo uma abordagem quantitativa, analisando os diferentes elementos envolvidos no problema a partir de uma linguagem matemática e verificando a relação entre as variáveis que compõem as características dos gestos, as configurações dos algoritmos de reconhecimento de padrões utilizados e os resultados obtidos. Teve-se a intenção de gerar análises de forma neutra e direcionada para destacar os resultados mais expressivos matematicamente evidenciados.

Diante do exposto, diz-se que esse trabalho seguiu duas fases em sua metodologia, ou seja, dois procedimentos. O primeiro guiado por uma pesquisa bibliográfica e o segundo guiado por uma pesquisa experimental.

No primeiro procedimento, foi realizado um estudo exploratório referente às áreas de análise de gestos e segmentação das fases do gesto, principalmente com base na linha de trabalho já seguida por pesquisas que antecederam esse estudo. Também, foi realizada a revisão do estado da arte, na forma de uma Revisão Sistemática, para levantamento de alternativas para representação dos gestos, considerando aspectos sintáticos de análise de gestos. A revisão sistemática é um método de pesquisa científica estruturado para coletar e analisar a pesquisa relevante sobre um determinado problema (BIOLCHINI et al., 2005). Uma revisão bibliográfica exploratória foi conduzida para complementar o levantamento das possibilidades existentes para construção da representação para os gestos, de procedimentos de pré-processamento necessários, de como os algoritmos de reconhecimento de padrões

podem ser configurados, de quais procedimentos de pós-processamento podem ser aplicados e quais as melhores métricas para avaliação dos resultados.

No segundo procedimento, experimentos foram planejados e executados para: compor combinações das diferentes características dos gestos e gerar representações multidimensionais, submeter tais representações a algoritmos de reconhecimento de padrões, e avaliar os resultados para identificação de quão discriminante cada representação é na tarefa de segmentação das fases do gesto.

Sobre a aplicação de algoritmos de reconhecimento de padrões, duas abordagens foram seguidas. Uma abordagem está baseada no uso de aprendizado de máquina supervisionado, no qual o algoritmo possui uma correção do erro baseado em uma classificação prévia dos dados realizada por um especialista. Nesse contexto procurou-se avaliar o quão discriminante uma representação computacional dos gestos pode ser, a fim de reproduzir uma classificação dos gestos de acordo com as fases do gesto esperadas pelos especialistas. A rede neural *Multilayer Perceptron* (MLP) foi usada para esse propósito pois é um aproximador de funções universal e tem um treinamento relativamente rápido. Para avaliação dos resultados obtidos nesta abordagem, foram usadas métricas de precisão, revocação, acurácia e F1-Score da classificação e a análise das matrizes de confusão formada pela quantidade de dados presentes em cada classe obtida contra a quantidade de dados em cada classe esperada para cada modelo de execução.

A segunda abordagem está baseada no uso de aprendizado de máquina não supervisionado, no qual o algoritmo não depende de um especialista para efetuar ajustes em sua execução, eliminando um possível viés produzido pelo especialista. A rede neural *Self Organizing Maps* (SOM) foi usada para esse propósito, por conta de sua capacidade de redução de dimensionalidade e consequente possibilidade de análise visual da estrutura topológica dos grupos descobertos. A avaliação dos resultados produzidos com essa abordagem foram avaliados de duas formas: a partir da aplicação de métricas de qualidade de agrupamento externas, verificando se o resultado do agrupamento é coerente com o esperado (métricas externas) e mostrando que é possível obter uma segmentação das fases do gesto sem a necessidade de pré-categorização dos dados ou parte deles; a partir da aplicação de métricas de qualidade do mapeamento produzido pelo SOM (erro de quantização e de distorção topológica), proporcionando segurança para a análise visual da topologia dos grupos encontrados e para a possível geração de informação nova sobre as fases do gesto.

1.4 Organização deste documento

A fim de documentar o resultado dos estudos bibliográficos e experimentais realizados, apresentar discussões sobre os resultados obtidos, as conclusões e oportunidades de trabalhos futuros, este texto de dissertação está organizado da seguinte forma:

- no capítulo 2 é apresentada a teoria dos gestos, com detalhes sobre a visão da área de análise linguística de como estudar e analisar os gestos, a explicação de como se dá a segmentação das fases do gesto de forma manual e de forma automática e a apresentação de algumas aplicações possíveis a partir da realização de tal segmentação;
- as diferentes características do gesto que podem ser exploradas para criação da representação multidimensional são discutidas no capítulo 3, com seções dedicadas a cada aspecto de representação (espacial e temporal) e possibilidades de combinações;
- os algoritmos de aprendizado de máquina MLP e SOM são descritos no capítulo 4;
- no capítulo 5, está a definição detalhada do problema, a descrição da estratégia de experimentação, com o detalhe do que compõe a base de dados, o que foi realizado como pré-processamento, o que foi considerado para a construção das representações multidimensionais dos gestos, quais experimentos foram realizados e, por fim, apresenta os resultados e análises das execuções dos algoritmos MLP e SOM;
- no capítulo 6 são apresentadas as conclusões às quais é possível chegar com base no que foi apresentado ao longo do trabalho, as possibilidades a serem exploradas em trabalhos futuros e as considerações finais sobre o trabalho;
- na sequência estão as referências bibliográficas.

2 Segmentação das fases do gesto

O estudo dos gestos examina o uso de partes do corpo na comunicação. A análise apresentada por esse trabalho se baseia nos estudos realizados por Kendon (1980), McNeill (1992) e McNeill (2015). O estudo desses pesquisadores, assim como o de outros que exploram essa área, utiliza como dados vídeos de pessoas falando e gesticulando e os converte em representações que facilitam a análise dos gestos. Esse capítulo se dedica a detalhar a abordagem proposta por Kendon (1980), McNeill (1992) e McNeill (2015).

2.1 Teoria de Gestos

McNeill (1992) abordou em seus estudos tipos funcionais de gestos, definindo cinco grandes categorias com características que os gestos podem representar. A saber:

- **icônicos (*iconics*)**. Gestos realizados em associação com objetos ou eventos.
- **metafóricos (*metaphorics*)**. Gestos também associados a objetos ou eventos, mas relacionados a ideias abstratas.
- **rítmicos (*beats*)**. Gestos que marcam o ritmo do discurso e possuem apenas duas fases de execução: direita/esquerda; para cima e para baixo etc.
- **coesivos (*cohesives*)**. Gestos que ligam 2 assuntos semelhantes, mas que foram separados no discurso por algum comentário adicional.
- **dêiticos (*deictics*)**. Gestos de indicação, apontam para objetos concretos, eventos ou locais, abstratos ou não.

Essas categorias podem estar combinadas em um discurso e há pesquisadores que propõem categorias adicionais. Como Gibbon (2009) que adiciona à definição de McNeill (1992) duas categorias: a de gestos emblemáticos (*emblems*), que são aqueles convencionais ou específicos de uma cultura, por exemplo o polegar levantado sinalizando uma resposta positiva; e a de gestos afetivos (*affective*) relacionados a estados emocionais.

Kendon (1980) propôs uma hierarquia de movimentos para descrever os gestos. Kendon (1980), McNeill (1992) e McNeill (2015) definiram essa hierarquia em termos de unidade gestual e fases do gesto. Uma unidade gestual é o período compreendido entre um membro do corpo sair da posição de descanso e retornar para a posição de descanso. Dentro de uma unidade gestual, até 5 fases do gesto podem acontecer. São elas: preparação

(*preparation*), pausa pré-movimento (*pre-stroke hold*), movimento (*stroke*), pausa pós-movimento (*post-stroke hold*) e retração (*retraction*). Todas definidas no Capítulo 1. Kita, Gijn e Hulst (1998) utilizaram essa hierarquia para descrever os gestos como uma série de fases com movimentos diferentes.

Ao utilizar essa hierarquia, Kita, Gijn e Hulst (1998) definiram que as fases *pre-stroke hold*, *stroke* e *post-stroke hold* representam uma fase expressiva do gesto. Também definiram que essa fase expressiva pode ser representada por um *hold* independente, assim chamada por ser considerada uma fase com significado próprio apesar de ser caracterizada por uma pausa de movimento. Devido a ter o significado próprio, pode substituir a fase de *stroke*. Conforme mencionando no Capítulo 1, a fase *hold* independente foi considerada na abordagem utilizada pelo presente trabalho. A figura 2 apresenta exemplos do período de descanso (*rest*) e de cada uma das fases do gesto consideradas.

Figura 2 – Ilustração das fases do gesto de uma unidade gestual.

Fonte: Ricardo Alves Feitosa, 2018

No descanso, as mãos permanecem sem ou com pouca movimentação na altura do abdômen. Esse período, geralmente, antecede a ocorrência de uma preparação (*preparation*), exemplificada na figura 2 pelo deslocamento das mãos em direção ao *stroke*, fase do gesto

mais expressiva representando os movimentos das mãos em uma corrida no exemplo apresentado. O *stroke* costuma ser precedido por um *hold*, uma retração (*retraction*), ou ainda, outra preparação para um novo *stroke*. O *hold* apresentado no exemplo mostra uma pausa nos movimentos da mão ainda em uma posição com significado expressivo para o gesto. Enquanto a retração ocorre levando as mãos de volta para a posição de descanso, que agora ocupa uma região pouco acima da inicialmente destacada.

No Capítulo 1 foi apresentada a figura 1 com um exemplo trivial mostrando as fases do gesto sobre uma ação de torção. Porém também há gestos em que as fases não estão tão facilmente discriminadas, onde uma retração não necessariamente é seguida pelo fim da unidade gestual, como, por exemplo, em passos de dança. Nessa atividade movimentos contínuos podem ser executados de forma a intercalar diversas fases, *strokes* seguidos por *holds*, retrações e preparações para, então, gesticular um novo *stroke*.

Há ainda a dificuldade em identificar fases com características semelhantes de configuração dos membros. Por exemplo, a dificuldade em diferenciar um *hold* de um descanso, uma vez que ambas estão associadas a redução da intensidade do movimento. Ou ainda a dificuldade de identificar as transições entre as fases, mesmo em gestos triviais como a da figura 1. Identificando o momento e o *frame* exato no qual o gesto sai da fase de preparação e entra na fase de *stroke*, ou quando o movimento de torção saiu da fase de *post-stroke hold* para entrar na fase de retração.

A figura 3 elenca os principais conceitos e suas relações dentro da Teoria dos Gestos. Como características para representar os gestos podem ser utilizadas a localização, a trajetória e a configuração dos membros do corpo que, quando em gesticulação, estão dentro de um dos tipos de gesto apresentados no início dessa seção relacionados a uma função específica para o gesto estar acontecendo.

A fala tem relação direta com o que está em execução pelo gesto, de forma que o gesto mais expressivo, o *stroke*, ocorre em sincronia com o recorte mais expressivo do discurso. Uma frase completa do gesto (*G-phrase*) apresenta um significado completo e, sob o aspecto de uma análise semântica na qual é estudado o significado da mensagem transmitida pelo sistema de línguas naturais, o gesto se apresenta como integrante desse sistema linguístico, complementando a mensagem que o discurso falado está transmitindo. E é nessa definição que o reconhecimento das fases que compõem as frases do gesto auxilia na análise da linguagem natural, posto que a alteração entre as frases do gesto altera os padrões de gesto e fala utilizadas pelos interlocutores, e facilitando a análise do significado

do gesto dentro de um discurso e, inclusive, auxiliando a análise semântica da própria fala ao interpretar as diferentes intensidades das fases do gesto dentro das próprias frases do gesto.

Figura 3 – Esquema de Teoria dos Gestos.

Fonte: Sarajane Marques Peres - Notas de aula, 2017

2.2 Segmentação das fases do gesto

Apresentadas as fases do gesto propostas por Kendon (1980), McNeill (1992) e McNeill (2015), uma nova tarefa se delimita para ser tratada. Como segmentar as fases do gesto? Essa tarefa implica em definir como é possível segmentar as fases do gesto com o objetivo de interpretá-lo. A segmentação das fases do gesto pode ser feita tanto de forma manual quanto de forma automática e esta seção se dedica ao detalhamento dessas duas técnicas.

2.2.1 Segmentação manual

Como mencionado no Capítulo 1, a análise de gestos geralmente envolve as atividades de: (1) identificar os movimentos realizados pelos usuários; (2) representar e reconhecer os movimentos com base em um modelo de análise específico; e (3) avaliação dos resultados. Tomando como fontes de dados os gestos capturados como vídeos e dividindo esses vídeos

em *frames*, focando nas atividades 2 e 3, um especialista humano precisa analisar *frame* a *frame* da base de dados para chegar a um veredito sobre a qual fase cada *frame* pertence quando a tarefa de segmentação é realizada manualmente.

Essa atividade manual precisa ser executada por um especialista que conheça a arquitetura de gestos proposta por Kendon (1980), McNeill (1992) e McNeill (2015), conhecendo as fases e o seu significado para poder derivar esse conceito em rotulações para cada *frame*. Além do mais, a pessoa que está executando essa atividade precisa conhecer o contexto no qual os gestos foram capturados e a sequência de execução desses gestos para poder determinar a qual fase do gesto um determinado *frame* pertence. Já que um *frame* isoladamente é uma imagem estática, e que, sem conhecer as informações mencionadas, é difícil determinar se essa imagem estática é um *stroke*, uma preparação para o *stroke* ou outra fase do gesto.

Ao exposto, soma-se a dificuldade de consenso entre especialistas humanos nessa tarefa. Interpretar e classificar um *frame* é uma atividade subjetiva, conforme já observado por Kita, Gijn e Hulst (1998), de forma que, mesmo dispondo das mesmas informações, dois especialistas podem chegar a conclusões diferentes sobre a qual fase do gesto um *frame* pertence.

Por esses motivos a atividade manual de segmentação das fases do gesto possui um tempo elevado para ser concluída e baixa eficiência. Além da impressão empírica apresentada no Capítulo 1, na qual um vídeo de 1 minuto levou cerca de 2 horas para ter suas fases segmentadas. Um cenário ainda mais extremo está no trabalho de Quek et al. (2002) que levou 10 dias para segmentar manualmente um vídeo de 1 minuto, realizando tanto a análise dos gestos quanto a transcrição da fala.

2.2.2 Segmentação automática

A segmentação das fases do gesto também pode ser executada de forma automática, com o uso de diferentes técnicas, reduzindo e, até, eliminando os problemas inerentes à segmentação manual. A segmentação automática geralmente parte das mesmas fontes de dados da segmentação manual, porém a decisão de a qual fase do gesto cada *frame* pertence fica a cargo de técnicas automatizadas como heurísticas, algoritmos de aprendizado de máquina, gramáticas etc. A aplicação dessas técnicas pode exigir que os dados de

gestos coletados sejam pré-processados e transformados em uma representação de dados específica para ser apresentada à técnica. Por exemplo, pode exigir a transcrição das características dos gestos em uma representação vetorial para ser processada por um algoritmo de reconhecimento de padrões.

Como exemplos de técnicas aplicadas à segmentação automática de gestos há o HMM, aplicado no trabalho de Yin e Davis (2014) na identificação das fases do gesto. Em Popa et al. (2008) também foi feita a utilização de HMMs para identificação de *strokes* a partir de uma representação dos gestos com a informação da trajetória da mão. HMMs têm apresentado frequente utilização para resolver o problema de segmentação das fases do gesto, como já realizado por Martell e Kroll (2007) para identificar as fases de preparação, *stroke*, *hold* e retração dentro de uma unidade gestual sobre vídeos de uma palestra.

Em Ramakrishnan e Neff (2013), foi feita a utilização de heurísticas e SVM. Inicialmente as heurísticas foram usadas para identificar a posição de descanso e pontos de interesse na base de dados composta por vídeos. Em seguida, SVMs foram usadas para classificar esses pontos de interesse em preparação, *stroke* e retração. Heurísticas também foram utilizadas por Bryll, Quek e Esposito (2001), em sua análise da organização do discurso a partir dos movimentos das mãos, na qual o foco foi a identificação de *holds*. Há ainda o trabalho de Gebre, Wittenburg e Lenkiewicz (2012), que empregou logísticas de regressão para detecção de *strokes* da mão a partir de *frames* com as informações dos gestos coletados.

Segmentar as fases do gesto significa dividir o período de execução de um gesto em fases, de acordo com o padrão de execução dos movimentos no tempo e dentro do modelo de fases adotado – neste caso o modelo desenvolvido por Kendon (1980), McNeill (1992) e McNeill (2015). Neste trabalho, os gestos foram analisados a partir da gravação de vídeos nos quais há uma pessoa gesticulando. Assim, tomando um vídeo V como uma sequência de *frames* f_i , $V = \{f_1, f_2, f_3, \dots, f_n\}$ com $i = 1..n$, trata-se a tarefa de segmentação dos gestos como um problema de rotulação de f_i de acordo com um conjunto de rótulos R .

O problema de rotulação pode ser resolvido considerando-o como um problema de classificação, no qual se conhece quais são as classes envolvidas e tem-se um conjunto de dados rotulado para treinamento dos classificadores, sendo que o conjunto de rótulos R é um conjunto finito de classes C . No presente trabalho, $C = \{c_1, \dots, c_5\}$, sendo que $c_i \in \{D, P, S, H, R\}$, no qual cada letra no conjunto correspondente a, respectivamente,

período de descanso¹ e as fases preparação, *stroke*, *hold* e retração. Assim, trata-se de um problema de classificação multiclasse.

Por outro lado, o problema de rotulação pode ainda ser resolvido considerando-o como um problema de agrupamento, no qual não se estabelece *a priori* quantos e quais são os grupos (ou rótulos) que serão buscados na análise. Nesse caso, os *frames* serão particionados em subconjuntos (ou grupos de *frames*) que dizem respeito a diferentes fases do gesto. Os grupos serão estabelecidos a partir de *frames* que, considerando a representação adotada para os gestos, compartilham propriedades suficientemente similares para indicar que pertencem a uma mesma fase do gesto. Assim, o conjunto de rótulos R é um conjunto finito de grupos $G = \{g_1, \dots, g_k\}$, sendo que k é um parâmetro a ser definido em cada modelo de agrupamento criado. Consequentemente, o resultado de segmentação obtido com a tarefa de agrupamento pode ou não respeitar as cinco fases propostas por Kendon (1980), McNeill (1992) e McNeill (2015).

Isto posto, tem-se o problema de segmentação das fases do gesto definido como tarefas de classificação e tarefas de agrupamento, de forma que podem ser submetidos a algoritmos especializados para resolução.

Considerando o problema de encontrar representações multidimensionais adequadas para alcançar segmentações adequadas para as fases do gesto, faz-se necessária a condução de experimentos que propiciem a aplicação dos algoritmos de reconhecimento de padrões que modelam a segmentação sobre os dados, e a análise dos resultados obtidos a partir da construção de modelos de classificação e modelos de agrupamento.

¹ Neste estudo, o período de descanso é considerado uma fase do gesto, por conveniência para apresentação das análises dos experimentos.

3 Representação de dados gestuais

A execução e interpretação de um gesto assume diferentes aspectos dentro de um sistema linguístico, sendo os mais comuns: aspectos espaciais que embutem informação sobre forma, amplitude ou direção; aspectos temporais que incorporam tanto informação sobre velocidade e aceleração, quanto informação sobre frequência e periodicidade; aspectos estruturais que carregam a informação estrutural e de relação entre os gestos e suas partes constituintes (SMITH, 2011; DIETTERICH, 2002).

Na segmentação automática dos gestos, normalmente é necessário efetuar um pré-processamento dos dados gestuais para a construção de uma representação adequada à técnica de segmentação automatizada que será utilizada. Por exemplo, para utilização de algoritmos de reconhecimento de padrões, como SVM e MLP, é necessário representar os gestos como um vetor de características para serem processados pelos algoritmos. Em sua maioria, os estudos que descrevem as representações de dados consideram que eles são derivados de um pré-processamento realizado sobre vídeos obtidos na internet ou capturados com o uso de câmeras do tipo RGB ou sensores. Esses sensores são capazes de obter as coordenadas espaciais de pontos de interesse. Em alguns estudos a fonte de informação não é um vídeo, mas sim uma imagem estática (LIU et al., 2014b; LIU et al., 2015; ZHANG et al., 2014; ZHAO et al., 2015). O uso de imagens estáticas ocorre quando o objetivo é realizar uma análise semântica da cena, analisando poses ou configuração assumida por partes do corpo.

Esse capítulo se dedica a apresentar como essas representações são formadas a partir dos aspectos dos gestos, com exemplos e estratégias de combinação dessas representações, formando possíveis respostas para o problema de pesquisa tratado pelo presente trabalho.

Os aspectos espaciais, temporais e estruturais utilizam representações dos gestos formadas a partir de dados brutos capturados de câmeras e sensores ou por consolidações desses dados. Esses aspectos são muito utilizados para a construção de formas de representação sobre os gestos, independente do uso de imagens estáticas ou dinâmicas (vídeos).

3.1 Representações orientadas à informação espacial

Na construção de representações orientadas à informação espacial são utilizadas características como as coordenadas (x, y, z) da posição espacial dos membros monitorados, a angulação entre membros, o contorno, entre outras.

As representações de dados mais empregadas geralmente combinam esses aspectos. Como em Kyan et al. (2015) onde a angulação em relação ao tronco é usada na representação do posicionamento de mãos, pulsos, cotovelos, ombros, centro dos ombros, quadris, centro dos quadris, joelhos, pés, tornozelos, espinha e cabeça.

A matriz de cores das imagens também é outra informação que vem sendo utilizada para construir representações. Liu et al. (2014b) utilizaram essa característica para resolver uma tarefa de identificação do posicionamento de partes do corpo. Já Dardas, Silva e El-Saddik (2014), Dardas e Georganas (2011) e Xu e Lee (2011) utilizaram a matriz de cores das imagens para resolver tarefas de reconhecimento de gestos, encontrando o contorno das mãos.

Em Zhu et al. (2011) a informação de contorno é utilizada para selecionar pontos ao redor do corpo sob análise. Esses pontos são utilizados na representação dos dados a serem usados na análise de gestos. O contorno de membros também pode ser representado por meio de informações extraídas da correspondência de *pixels*, como feito por Liang, Yuan e Thalmann (2014) e por Liu et al. (2015), que utilizaram apenas essa informação para realizar o reconhecimento de gestos e a identificação do corpo, respectivamente. Correspondência de *pixels* é um termo utilizado por Liu et al. (2015) para representar uma técnica que analisa as imagens à procura de *pixels* com características semelhantes quanto às coordenadas espaciais e tonalidade RGB da cor do *pixel*.

Liu et al. (2014a) utilizaram 79 pontos do rosto e o movimento dos olhos para reconhecer expressões faciais. A mesma tarefa é realizada por Nguyen e Ranganath (2012) e Caridakis, Asteriadis e Karpouzis (2014) que combinaram os mesmos 79 pontos do rosto e o movimento dos olhos com o formato de sobrancelhas, da boca e suas coordenadas espaciais.

3.2 Representações orientadas à informação temporal

Para a construção de representações orientadas à informação temporal, as características utilizadas são a velocidade com a qual os membros monitorados se deslocam, a aceleração, a trajetória e outras informações para as quais a ordem dos *frames* e imagens estáticas importam.

A utilização de informações relacionadas à aceleração e velocidade do gesto para a representação é empregada por Bailador e Triviño (2010) que associaram à aceleração coordenadas espaciais para representar o movimento das mãos e classificá-las utilizando uma lógica Fuzzy. Khan, Bailey e Gupta (2012) utilizaram a informação de velocidade associada à aceleração para representar o movimento, também das mãos, a fim de se identificar *holds* em sequências de movimentos. Kojima, Tamura e Fukunaga (2002) combinaram coordenadas espaciais com aceleração na construção da representação de dados para os gestos. Há ainda a representação de dados gerada por Madeo, Wagner e Peres (2013), que utilizou a velocidade e a aceleração das mãos e dos pulsos.

Não são comuns trabalhos que utilizam apenas informações temporais. Normalmente essas características estão combinadas com os aspectos espaciais. Como o de Hachaj e Ogiela (2014) que utiliza as coordenadas espaciais e angulações de 15 pontos representando mãos, cotovelos, ombros, tórax, quadris, joelhos, pés, espinha e cabeça e a informação temporal (um selo de tempo associado a cada *frame*). Em Caramiaux, Wanderley e Bevilacqua (2012) apenas a posição das mãos e braços é combinada com a angulação e a trajetória dos membros inferiores e superiores do corpo na tarefa de identificação dos gestos.

A informação da trajetória das mãos e da posição de partes do corpo também foi explorada por Abid, Petriu e Amjadian (2015) na tarefa de reconhecimento e classificação de gestos em tempo real. A utilização da informação de trajetória, junto com o posicionamento de objetos com o qual as pessoas estão interagindo nas cenas capturadas, compuseram a representação de dados usada por Lücking et al. (2013) e Rosani, Conci e Natale (2014) também na tarefa de reconhecimento e classificação de gestos em tempo real.

3.3 Estratégias de combinação de características

A partir dos diferentes aspectos do gesto e da sua utilização nas mais diversas tarefas citadas, há também trabalhos que fazem a combinação desses aspectos a fim de obter melhores resultados em sua área de aplicação. Muitos já foram citados, como Bailador e Triviño (2010), Kojima, Tamura e Fukunaga (2002) e Hachaj e Ogiela (2014), que combinaram aspectos espaciais com aspectos temporais.

Essas combinações geraram bons resultados nas tarefas em que foram aplicadas. Hachaj e Ogiela (2014) conseguiu acurácia de 98%, enquanto Bailador e Triviño (2010) e Kojima, Tamura e Fukunaga (2002) chegaram a alcançar 100% de acurácia, sendo as abordagens que realizaram combinações de características as que alcançaram os melhores resultados em termos de acurácia. Dentre as baixas taxas de acurácia estão a utilização de dados apenas sobre posição e angulação dos membros, como em Gaglio et al. (2013), ou apenas sobre velocidade e aceleração, como em Khan, Bailey e Gupta (2012), ambos com 88% de acurácia.

Tais cenários mostram o quão promissor a combinação de características é, inclusive para o reconhecimento de gestos. Nesse contexto é plausível que a plotagem da velocidade de um membro do corpo, medida ao longo dos *frames* de um vídeo, a aceleração ou até aspectos espaciais como as coordenadas x , y e z podem ser tratadas como uma série temporal e, conseqüentemente, estudadas sob esse aspecto, abrindo possibilidades de, por exemplo, visualizar essa série temporal como um sinal sobre o qual um filtro possa ser aplicado a fim de tornar os aspectos usados para representação discriminantes o suficiente para que o gesto possa mais facilmente ter suas fases identificadas por algoritmos de reconhecimento de padrões.

4 Aprendizado de Máquina

Redes Neurais Artificiais (RNA) são modelos matemáticos cujo projeto é inspirado na estrutura neural de organismos inteligentes, capazes de adquirir conhecimento por meio da experiência. As RNA se fundamentam nos seguintes conceitos (FAUSETT, 1994; HAYKIN, 2009): *a)* Os neurônios são os elementos responsáveis por processar as informações. RNAs são modelos compostos por uma camada de entrada (neurônios de entrada), que representam os dados a serem processados, e uma camada de saída (neurônios de saída), que representam a classe/grupo ao qual cada valor de entrada foi associado. *b)* Através de conexões, ou sinapses, os sinais são transmitidos entre os neurônios. *c)* Cada conexão está associada a um peso que influencia o sinal transmitido. *d)* A saída de cada neurônio é determinada por uma função de ativação aplicada a sua entrada. A figura 4 apresenta esses conceitos no modelo apresentado por Haykin (2009).

Figura 4 – Modelo de RNA.

Fonte: Haykin (2009)

As estruturas acima chamadas de “neurônio” também são modelos matemáticos inspirados no modelo de funcionamento de um neurônio biológico. Várias destas estruturas organizadas dentro de uma arquitetura caracterizam, de forma simplificada, uma RNA.

4.1 Aprendizado de Máquina Supervisionado com MLP

O aprendizado supervisionado consiste em fornecer dados de entrada a um algoritmo, como o de uma RNA, e comparar a saída gerada com uma saída desejada. Esse processo é conhecido como treinamento e, através dele, é possível ajustar parâmetros da arquitetura da rede neural para que a saída gerada se aproxime do resultado desejado (FAUSETT,

1994; HAYKIN, 2009). O treinamento de uma RNA com aprendizado supervisionado é realizado de tal forma que os neurônios de saída são previamente associados a rótulos de classes de dados. Por meio do algoritmo de treinamento, as sinapses que ligam neurônios de entrada com neurônios de saída são definidas de forma a permitir o estabelecimento de um mapeamento dos elementos do conjunto de dados para as classes desejadas.

Perceptron Multicamadas (*Multilayer Perceptron* – MLP) é uma RNA amplamente utilizada em problemas de classificação não linearmente separáveis e se caracteriza por seu aprendizado supervisionado. É composta por uma camada de entrada, formada por nós fonte que serão a origem dos dados a serem processados, uma ou mais camadas ocultas e uma camada de saída, essas duas últimas formadas pelos neurônios que processarão a rede e determinarão o valor resultante (HAYKIN, 2009). No processamento dessa rede, o sinal é propagado da camada de entrada para uma ou mais camadas ocultas e seguindo adiante para a camada de saída, caracterizando uma rede alimentada adiante (*feedforward*). A esse tipo de rede é inserido, opcionalmente, um neurônio chamado de bias na camada de entrada, normalmente inicializado com 1, para aumentar o grau de liberdade e adaptação da rede neural durante o aprendizado (SILVA; PERES; BOSCAROLI, 2017). A figura 5 ilustra uma MLP com duas camadas ocultas.

Figura 5 – Arquitetura gráfica de um MLP com duas camadas ocultas.

Fonte: Haykin (2009)

O aprendizado dessa rede utiliza a técnica supervisionada de retropropagação de erro (*backpropagation*), baseada na regra de aprendizagem por correção de erro. Essa técnica é formada por duas fases: a propagação e a retropropagação. Na propagação um

vetor de entrada, contendo os dados a classificar, é apresentado à camada de entrada e seu efeito se propaga camada a camada de neurônios sensoriais da rede, produzindo um conjunto de saídas. Durante essa fase, os pesos sinápticos dos neurônios possuem valores fixos. Na sequência, a retropropagação ajusta os valores dos pesos sinápticos com base em um erro calculado como a diferença entre a resposta desejada e a obtida. Esse ajuste é feito da camada de saída até a camada de entrada, ou seja, no sentido inverso da fase de propagação. Essa técnica de aprendizado faz com que a resposta da rede se mova em direção a resposta esperada (HAYKIN, 2009).

Um MLP possui basicamente 3 características importantes:

1. O modelo de uma rede MLP utiliza uma função não linear de ativação de seus neurônios. Essa não-linearidade é suave e diferenciável em qualquer ponto, normalmente conferida por uma *sigmóide* representada pela função logística 1. Onde v_j é a soma ponderada das entradas mais o bias do neurônio j e y_j é a saída do neurônio.

$$y_j = \frac{1}{1 + \exp(-v_j)} \quad (1)$$

2. A rede contém uma ou mais camadas ocultas de neurônios que atuam no aprendizado de tarefas complexas, extraindo as características mais significativas dos dados de entrada.
3. A rede possui um alto grau de conectividade entre seus neurônios. Qualquer alteração dessa conectividade deve implicar na revisão do número de neurônios ou de seus pesos.

Como apresentado, o erro é calculado a partir da diferença entre a resposta desejada e a obtida, conforme equação 2, onde $e_j(n)$ é o erro do neurônio j no instante n .

$$e_j(n) = d_j(n) - y_j(n) \quad (2)$$

Assim, durante a retropropagação do erro, os novos valores dos pesos da rede podem ser definidos por regras que considerem esse erro, como o procedimento definido pela Regra Delta. Na Regra Delta o valor do vetor de pesos w no instante $n + 1$ é definido por:

$$w(n + 1) = w(n) + \Delta w(n)$$

Sendo que:

$$\Delta w_{ij}(n) = w_{ij} - \alpha \frac{\partial EQM}{\partial W_{ij}}$$

$$EQM_{total} = \frac{1}{N} \sum_N^{n=1} EQM(n)$$

$$EQM(n) = \frac{1}{2} \sum_{ns}^{k=1} e_k^2(n)$$

Em que α é a taxa de aprendizado, N é o número de dados de entradas equivalente ao número de exemplares do conjunto de treinamento e EQM representa o erro quadrático médio.

O ajuste dos pesos pode ser realizado de duas formas: Padrão a padrão, também conhecido como sequencial ou *on-line*; e batelada ou lote, igualmente chamado de modo batch.

No modo sequencial, o ajuste dos pesos sinápticos é feito sempre que um dado é apresentado ao MLP, enquanto no modo em lote o cálculo do gradiente em relação a cada peso é realizado a cada iteração, mas a atualização dos pesos só é realizada depois que todos os dados são apresentados à rede, constituindo o que é conhecido como época de treinamento.

Para otimização dos resultados de um MLP, podem ser utilizadas técnicas sobre o *backpropagation*. Em sua configuração padrão, a Regra Delta apresentada aplica um gradiente descendente sobre o quadrado da soma do erro. Uma otimização possível é a inclusão da taxa de aprendizado adaptativa, que é atualizada a cada atualização dos pesos. Essa técnica pode ou não ser combinada com a adição de uma constante à taxa de aprendizado, chamada de *momentum*, com o objetivo de controlar a oscilação sobre o erro do classificador (HAYKIN, 2009).

Outra otimização que pode ser utilizada é o método Levenberg–Marquardt, que combina duas abordagens: o método de Newton, com o objetivo de convergir rapidamente para próximo de um mínimo local ou mínimo global, mas sem a garantia de que isso possa acontecer; e o gradiente descendente, que possui a certeza de convergência, mas de forma mais lenta (HAYKIN, 2009).

Para avaliação dos resultados apresentados por um MLP usualmente são utilizadas as medidas de precisão, revocação, acurácia e F1-Score da classificação conforme definição a seguir:

- **precisão.** Identifica a taxa de dados classificados corretamente dentre todos os dados identificados. Por exemplo, no reconhecimento de fases do gesto em uma cena são identificados sete *strokes*, porém desses apenas quatro são realmente *strokes* de acordo com o especialista, logo a precisão é de 4/7.
- **revocação.** Identifica a taxa de dados classificados corretamente dentre todos os dados que poderiam ser identificados. Usando mesmo exemplo da precisão, foram identificados corretamente apenas quatro *strokes*, porém a base continha 50 *strokes*, assim a revocação é de 4/50.
- **acurácia.** É uma média ponderada entre os dados da precisão e da revocação, calculada de acordo com a equação:

$$Acurácia = \frac{VP + VN}{VP + VN + FP + FN}$$

Sendo:

- VP = Verdadeiros positivos;
 - VN = Verdadeiros negativos;
 - FP = Falsos positivos;
 - FN = Falsos negativos.
- **F1-Score.** Combina a precisão e a revocação em uma média harmônica, caracterizando-se como uma excelente métrica para analisar as medidas de precisão e revocação. Definido pela equação:

$$F1_Score = 2 \times \frac{precisão \times revocação}{precisão + revocação}$$

4.2 Aprendizado de Máquina Não Supervisionado com SOM

O aprendizado não supervisionado não dispõe de uma saída desejada com a qual se possa comparar o resultado da execução do algoritmo. Dessa forma não há como definir previamente os rótulos para os neurônios de saída (FAUSETT, 1994; HAYKIN, 2009). Nesse tipo de aprendizado, o treinamento auto-organiza os neurônios de saída (ou suas sinapses) de forma a criar um mapeamento que associa os elementos do conjunto de dados a grupos formados de acordo com as semelhanças encontradas.

Mapas Auto Organizáveis (*Self Organizing Maps* — SOM) consistem numa RNA de aprendizado competitivo e não supervisionado, classicamente usada para implementar

a tarefa de agrupamento. Assim como o MLP, o SOM possui uma camada de entrada e uma camada de saída, onde cada neurônio de saída está associado a um vetor de pesos sinápticos que lhes confere uma posição no espaço vetorial. Também há a ideia de vizinhança, onde cada neurônio de saída está conectado à neurônios adjacentes, seguindo uma lei de vizinhança topológica. Essa estrutura de vizinhança permite que o neurônio vencedor exerça influência sobre seus vizinhos, de forma a atualizar também os pesos (ou posicionamento vetorial) dos tais vizinhos (SILVA; PERES; BOSCARIOLI, 2017).

Uma das principais propriedades do SOM é a redução da dimensionalidade dos dados de entrada. Essa redução ocorre através de uma projeção não linear de um espaço de dados de alta dimensionalidade em um conjunto discreto de neurônios, dispostos em um mapa na forma de vetor ou matriz n-dimensional, conforme exemplificado na figura 6. Se a dimensão topológica do mapa é determinada para 2, é possível realizar uma análise simplificada (e visual) da proximidade topológica existente entre os dados. Vale ressaltar que a dimensão topológica do mapa pode assumir qualquer valor menor ou igual à dimensão do espaço dos dados, mas, geralmente utiliza-se um mapa de dimensão 2 (KOHONEN, 2001).

Figura 6 – Modelo de Kohonen.

Fonte: Haykin (2009)

No SOM há duas métricas de distâncias envolvendo as relações entre os neurônios e dados: a distância vetorial, aplicada entre os vetores de pesos e os vetores de dados (neurônios de entrada); e a distância matricial entre os índices da matriz que organiza o mapa. A distância entre os vetores de pesos é utilizada para calcular o neurônio vencedor, verificando qual neurônio de saída mais se aproxima do dado de entrada durante o treinamento. A segunda é utilizada para calcular o raio r dentro do qual os vizinhos do neurônio vencedor serão atualizados.

Na forma mais comum, a atualização dos neurônios vizinhos é proporcional a distância entre eles e o neurônio vencedor. Assim, quanto mais distante um neurônio estiver do neurônio vencedor, menor será a atualização em seu vetor de pesos, sendo que os neurônios fora do raio r , normalmente, não sofrem atualizações para fins de eficiência computacional. A atualização é feita segundo a equação 3.

$$w_j(n+1) = w_j(n) + \alpha(n)h_{j,i(x)}(n)(x(n) - w_j(n)) \quad (3)$$

Em que $\alpha(s)$ é a taxa de aprendizado e h é a função de vizinhança centrada ao redor do neurônio vencedor j .

Sendo assim, para formação de um mapa auto-organizável, três passos são necessários segundo Haykin (2009):

- 1. Competição.** Para cada valor de entrada, os neurônios do mapa calculam o seu respectivo valor com base em uma função discriminante, que pode ser, por exemplo, a distância entre o neurônio e o dado. O neurônio com melhor valor discriminante, vence a competição e é conhecido como *Best Matching Unit* (BMU).
- 2. Cooperação.** O neurônio vencedor determina um raio de vizinhança topológica de neurônios que também sofrerão atualização. A vizinhança pode ser definida por uma Gaussiana, de forma a controlar o grau de participação no processo de aprendizado dos neurônios excitados na vizinhança do BMU. A equação 4 é a proposta por Haykin (2009) para esse fim, onde: $h_{i,j(x)}$ é a função de vizinhança do neurônio vencedor $i(x)$ para o neurônio vizinho j , $d_{j,i}$ é a distância do neurônio vencedor $i(x)$ para o neurônio vizinho j e n é o tempo, a iteração atual. O termo $\sigma(n)$ é definido pela equação 5 onde σ_0 é o sigma inicial definido como parâmetro do SOM, t_1 é definido pela equação 6 e n é o tempo, a iteração atual. Na equação 6, *epochs* é o número de épocas total utilizadas.

$$h_{j,i(x)} = \exp\left(-\frac{d_{j,i}^2}{2\sigma^2(n)}\right) \quad (4)$$

$$\sigma(n) = \sigma_0 \exp\left(-\frac{n}{t_1}\right) \quad (5)$$

$$t_1 = \left(-\frac{epochs}{\log \sigma_0}\right) \quad (6)$$

3. Adaptação sináptica. Permitem que os neurônios vencedores e seus vizinhos cobertos pelo raio de vizinhança melhorem seus valores de função discriminante em relação às entradas. Assim a resposta de um neurônio vencedor à apresentação subsequente de uma entrada similar é melhorada.

Seguindo o proposto por Haykin (2009), a implementação do algoritmo SOM possui uma fase de ordenação e uma fase de convergência. Na fase de ordenação é realizada a ordenação topológica dos pesos do mapa. O aconselhável para essa fase é um valor próximo à 0,1 para a taxa de aprendizado, um valor alto de abrangência de vizinhança e cerca de 1000 épocas. Na fase de convergência é aplicado o ajuste fino no mapa, ou seja, é realizado um ajuste mais preciso dos neurônios sobre os dados de entrada. Nesse ponto são consideradas a taxa de aprendizado, o grau de abrangência da vizinhança e o número de épocas ajustados, sendo que o aconselhável para essa fase é um valor próximo a 0,01 para a taxa de aprendizado, um valor baixo para o grau de abrangência, para que mantenha a atualização em um raio pequeno em torno do neurônio vencedor e épocas correspondentes a 500 vezes o número de neurônios na grade, segundo Haykin (2009).

Também é aconselhável que a taxa de aprendizado utilizada para ajuste dos pesos do mapa sofra decaimento com o tempo durante a execução do algoritmo (HAYKIN, 2009). Esse decaimento é definido pela equação 7.

$$a(n) = a_0 \exp\left(-\frac{n}{t_2}\right) \quad (7)$$

Onde a_0 é a taxa de aprendizado inicial, t_2 é o número de épocas totais, somando-se a quantidade de épocas da fase de ordenação e convergência, e n é o tempo, ou seja, a iteração atual.

Para avaliação dos resultados apresentados pelo SOM podem ser utilizadas as medidas relacionadas ao erro de quantização e ao erro topológico. O erro de quantização é uma medida relativa à resolução do mapa. O número de neurônios no mapa deve ser menor que a quantidade de dados do conjunto de entrada. Nessa quantização, há a perda de informação e um erro é produzido como consequência da padronização das particularidades dos dados. Esse erro é definido segundo a equação 8, onde N é o número de dados de entrada, b é o índice do BMU e $\|.\|$ é o cálculo da distância entre o dado e o BMU.

$$E_q = \frac{1}{N} \sum_{i=1}^N \|x_i - w_{bi}\| \quad (8)$$

O erro topológico é uma medida utilizada para avaliar o quanto a estrutura resultante do mapa aproxima os padrões do espaço de entrada. Esse erro está relacionado à perda de informação que ocorre com a redução de dimensionalidade. A equação 9 apresenta como esse erro é calculado, onde $u(x_i) = 0$ quando o primeiro BMU e o segundo BMU são adjacentes no espaço de saída, caso contrário $u(x_i) = 1$.

$$E_t = \frac{1}{N} \sum_N^{i=1} u(x_i) \quad (9)$$

Outra forma de avaliar o resultado da aplicação do SOM é com o uso da matriz-U, construída para permitir a inspeção visual dos grupos formados pela organização topológica dos neurônios. Sua elaboração resulta em uma imagem $f(x, y)$ na qual as coordenadas de cada *pixel* são derivadas das coordenadas dos neurônios do mapa treinado pelo SOM. A intensidade da cor de cada pixel corresponde à distância entre os neurônios.

A figura 7 ilustra uma matriz-U resultante da aplicação do SOM sobre os elementos da base de dados Iris¹. Essa base é composta por 150 elementos com dados sobre três tipos de flores. O resultado final do SOM foi plotado na matriz-U do exemplo e os hexágonos com cores de tons mais azulados representam concentrações de neurônios, denotando um possível grupo, que são delimitadas por picos, identificados pelos hexágonos com tons mais claros indo em direção à cor amarela.

Figura 7 – Matriz-U gerada sobre a base de dados Iris.

Fonte: Ricardo Alves Feitosa, 2018

¹ Disponível em <https://archive.ics.uci.edu/ml/datasets/iris>.

5 Experimentação, resultados e discussões

Para solucionar o problema apresentado, no qual espera-se identificar representações multidimensionais adequadas para permitir a segmentação automática das fases do gesto, alguns procedimentos devem ser considerados. Inicialmente, o problema será caracterizado, de forma a deixar claro quais são os aspectos que serão analisados e como eles definem o problema em questão, e como os experimentos serão realizados de forma a permitir a verificação da hipótese e alcance dos objetivos delineados para este trabalho. Na estratégia de experimentação estão descritos: a base de dados que suportará os experimentos, os procedimentos de pré-processamento, os procedimentos de parametrização dos algoritmos que implementam as redes neurais artificiais MLP e SOM, e como será realizada a avaliação de resultados e indicação das características que formam as melhores representações multidimensionais para os gestos.

Neste capítulo ainda são apresentados os resultados obtidos com as execuções dos testes para cada algoritmo e as análises sobre esses resultados, avaliando o quão discriminante cada representação foi para a obtenção das classificações e agrupamentos.

5.1 *Estratégia de experimentação*

A estratégia de experimentação tem o objetivo de criar um ambiente de análise de diferentes representações multidimensionais para os gestos, considerando os resultados de segmentação das fases do gesto obtidos a partir do uso dessas representações junto aos algoritmos MLP e SOM. Esta seção, portanto, apresenta a base de dados a ser usada nos experimentos, os procedimentos de pré-processamento e de extração de características utilizados para a execução dos experimentos e a estratégia de combinação dessas características a fim de compor as diferentes representações multidimensionais que serão estudadas.

5.1.1 Base de dados

Os experimentos foram executados em um contexto controlado, sobre o qual se conhece as particularidades da gesticulação e das fases do gesto executado, de forma a proporcionar uma análise assertiva sobre o resultado gerado e maior domínio sobre os

pré-processamentos a serem executados. Seguindo essas diretrizes, foi escolhida a base de dados *Gesture Phase Segmentation Data Set*¹, desenvolvida pelo grupo de pesquisa correlato ao presente trabalho.

A base de dados citada é composta por sete vídeos referentes à gesticulação de três pessoas, (A, B e C), durante uma atividade de contação de histórias (um tipo específico de contexto de discurso caracterizado por um monólogo). Trata-se de uma base de dados rotulada a partir do modelo de fases do gesto, já discutido neste texto. Dados gerais sobre os vídeos e a distribuição de *frames* por fases do gesto são apresentados na tabela 1.

Tabela 1 – Dados sobre os vídeos que compõe a base de dados; coeficiente de concordância entre os rotuladores humanos. # - quantidade; sec - segundos (unidade de tempo); fr - *frames*; D - Posição de descanso; G - Gesto; $K\alpha$ - coeficiente de concordância α (KRIPPENDORFF, 2012); Ug - Unidade gestual; P - Preparação; S - *Stroke*; H - *Hold*; R - Retração; F - Fases.

Usuário	História	# sec	# fr	# D	Unidade gestual G	$K\alpha$ Ug	# P	Fase do gesto S	H	R	$K\alpha$ F
A	1	60s	1747	698	1049	0,90	163	656	39	191	0,83
	2	40s	1264	493	771	0,90	200	431	54	86	0,80
	3	61s	1834	662	1172	0,85	279	535	150	208	0,78
B	1	36s	1073	74	999	0,62	412	287	217	83	0,59
	3	48s	1424	194	1230	0,67	469	390	201	170	0,62
C	1	37s	1111	286	825	0,80	236	262	193	134	0,67
	3	46s	1448	362	1086	0,89	338	389	144	215	0,73

Fonte: Adaptado de Wagner et al. (2014b) e Madeo, Peres e Lima (2016)

Os vídeos foram gravados com o suporte do sensor MS Kinect, usando a sua capacidade de rastrear o corpo humano presente na imagem capturada, e fornecer coordenadas (x, y, z) de pontos do interesse no corpo. Seis pontos de interesse foram rastreados: mão esquerda, mão direita, pulso esquerdo, pulso direito, cabeça e tronco. A partir desses dados, foram extraídas características temporais representadas pela velocidade vetorial e escalar, e aceleração vetorial e escalar. Esses dados estão organizados na base de dados sob uma organização de arquivos, para cada um dos sete vídeos.

No arquivo de extensão *.raw* encontram-se os dados espaciais extraídos de cada *frame*, ou seja, as coordenadas (x, y, z) . A coordenada x , dada em *pixels*, diz respeito à posição, no sentido vertical, do ponto de interesse no *frame*; a coordenada y , também dada em *pixels*, diz respeito à posição, no sentido horizontal, do ponto de interesse no *frame*; a coordenada z é dada na forma de distância em milímetros entre o sensor e o ponto de interesse monitorado. Nesse arquivo também há um selo de tempo que marca o momento

¹ Disponível em <https://archive.ics.uci.edu/ml/datasets/Gesture+Phase+Segmentation>.

em que o *frame* foi obtido e o rótulo do *frame* referente à fase do gesto que ocorre nele, segundo um especialista humano.

No arquivo de extensão *.va3* encontram-se as medidas de velocidade e aceleração, escalares e vetoriais, das mãos direita e esquerda e dos pulsos direito e esquerdo. A velocidade pôde ser obtida a partir do deslocamento dos pontos de interesse no tempo de três *frames*. A aceleração pôde ser obtida a partir das medidas de velocidade.

Nesse trabalho foram adicionadas à base de dados as características de angulação entre as mãos e o tronco e a angulação entre o ponto central das mãos e seus respectivos pulsos. O cálculo dessas angulações foi realizado utilizando as coordenadas espaciais já presentes na base de dados e obtendo-se o arco cosseno do cosseno calculado na equação 10, que tomou como exemplo dois vetores genéricos $v = (p_1, q_1)$ e $u = (p_2, q_2)$.

$$\cos(\Theta) = \frac{\langle v, u \rangle}{|v| \cdot |u|} \quad (10)$$

O ângulo foi calculado seguindo a equação 11 onde *arccos* é o arco cosseno do cosseno de Θ . Nesse trabalho os pontos x e y dos membros do corpo utilizados para cálculo da angulação foram considerados como vetores com origem em $(0, 0)$.

$$\hat{Angulo} = \frac{\arccos(\Theta) \cdot 180}{\pi} \quad (11)$$

5.1.2 Pré-processamento

Para submeter os dados referentes a gestos aos algoritmos de reconhecimento de padrões, foram realizados pré-processamentos de normalização, para anular efeitos de deslocamento do gesticulador em relação ao sensor, e de eliminação de ruído a partir da aplicação de um filtro.

A normalização teve como objetivo eliminar as diferenças nas coordenadas (x, y, z) entre *frames* em decorrência da movimentação do gesticulador para frente e para trás em relação ao sensor, mas sem movimentar os membros monitorados com relação ao seu tronco. Considerando o seguinte exemplo: em um *frame* f o gesticulador está posicionado a m metros do sensor com as mãos posicionadas nas coordenadas (x, y, z) ; em um *frame* $f + 1$ o gesticulador se movimenta alguns centímetros em direção ao sensor sem alterar a distância das mãos em relação ao tronco. Sem a normalização, o algoritmo de reconhecimento

de padrões irá considerar que houve movimentação das mãos entre os *frames* por haver diferença entre as coordenadas das mãos entre os *frames*. Com a normalização essa diferença é eliminada.

A normalização foi calculada tomando o tronco como referência e seguindo as etapas de: cálculo da diferença d_t entre as coordenadas do tronco no *frame* f e $f + 1$; e subtração de d_t das coordenadas dos membros monitorados no *frame* $f + 1$. Para a coordenada x e y há ainda os passos que envolvem o cálculo da distância entre as coordenadas x e y do tronco e da cabeça no *frame* f e $f + 1$, o cálculo da razão entre essas distâncias e a multiplicação dessas referências pelas coordenadas dos membros monitorados em $f + 1$.

Sobre a eliminação de ruído, o gesto, no contexto deste trabalho, é caracterizado pela excursão das mãos pelo espaço, e a descrição dessa excursão se dá por medições realizadas a cada *frame*. Dessa forma, a descrição do gesto pode ser entendida como um sinal contínuo variante no tempo. No quadro superior da figura 8 é mostrado o sinal formado pela velocidade escalar da mão direita ao longo de um vídeo.

Figura 8 – Sinal referente à velocidade escalar da mão direita (eixo y) durante a realização de um gesto ao longo dos *frames* (eixo x): original e com filtro *wavelet*.

Fonte: Ricardo Alves Feitosa, 2018

Os dados relacionados aos aspectos espaciais e temporais podem ser visualizados como dados contínuos, quando os plotamos *frame* a *frame*. Sendo assim, é possível a aplicação de um filtro. Porém, para serem processados por computadores digitais, precisam ser tratados como sinais discretos, gerados a partir da amostragem de sinais contínuos (SEMMLOW; GRIFFEL, 2014). O efeito da aplicação de filtros sobre o sinal original em termos de melhoria para discriminação das fases do gesto é objeto de estudo neste trabalho.

O filtro *wavelet* foi aplicado ao sinal de gestos como um estudo referente à eliminação de ruído. O filtro *wavelet* consiste em uma transformada chamada *Discrete Wavelet Transform* (DWT) capaz de obter características nos domínios de tempo e frequência, ou seja, é adequada para trabalhar um sinal cuja frequência se altera ao longo do tempo, como é o caso do sinal gerado a partir dos gestos.

A *Discrete Wavelet Transform* (DWT) é a transformada utilizada para sinais discretos, implementada com o uso de dois filtros em um banco de filtros, um filtro passa-baixa $G(z)$ e um filtro passa-alta $H(z)$ (SEMMLOW; GRIFFEL, 2014), dados por:

$$G(z)G(z^{-1}) + G(-z)G(-z^{-1}) = 1$$

$$H(z) = zG(-z^{-1})$$

Coefficientes de aproximação cA e de detalhe cD são gerados pelos filtros passa-baixa e passa-alta, respectivamente (SEMMLOW; GRIFFEL, 2014). Normalmente, uma análise *wavelet* é realizada utilizando-se mais de um nível de decomposição do sinal. Ou seja, o sinal é inicialmente decomposto em cA e cD , em seguida cA é decomposto em novos cA e cD , e assim por diante conforme a quantidade de níveis desejados. Para esse trabalho foram considerados três níveis de decomposição do sinal. Por fim, para obter uma nova representação do sinal, utiliza-se a média e/ou o desvio padrão dos cA do último nível de decomposição e dos cD de todos os níveis de decomposição analisados (LIMA; COELHO, 2011; SUBASI, 2007). No quadro inferior da figura 8 é mostrado o sinal formado pela velocidade escalar da mão direita submetido ao filtro *wavelet*.

Uma visão mais detalhada do efeito da aplicação do filtro *wavelet* pode ser observada nas figuras 9 e 10, as quais trazem um detalhamento sobre as fases do gesto (cada letra plotada sobre o sinal indica o rótulo do *frame* referente àquele ponto do sinal). Visualmente é possível observar a suavização do sinal, no entanto, tal suavização pode também esconder comportamentos que caracterizam uma fase do gesto.

Figura 9 – Sinal original - Velocidade escalar da mão direita (eixo y) ao longo dos *frames* (eixo x).

Fonte: Ricardo Alves Feitosa, 2018

Figura 10 – Sinal filtrado (filtro *wavelet*) - Velocidade escalar da mão direita (eixo y) ao longo dos *frames* (eixo x).

Fonte: Ricardo Alves Feitosa, 2018

5.1.3 Construção de representações multidimensionais

O sinal que representa o gesto é composto por medidas escalares e pode ser visto como uma série temporal com cada escalar sendo um dado da série. Na análise de séries temporais, é possível fazer estudos usando o conceito de espaço de fases, o que também proporciona uma análise que considera que sequências de medidas escalares dependem de estados anteriores no sinal (ou no sistema) (KANTZ; SCHREIBER, 2004).

Neste trabalho, a reconstrução do espaço de fases será considerada com atraso de um *frame* de forma que o sinal do gesto para cada *frame* será na realidade representado por uma janela retangular simples de tamanho j , em que j será parametrizado para 1, 20 e 40 nos experimentos. O parâmetro j indica a dimensão do espaço de fases e também caracteriza a representação de gestos resultante como uma representação multidimensional.

Com o uso do conceito de janelas retangulares espera-se criar condições para que os algoritmos MLP e SOM considerem informação anterior ou posterior ao *frame* em análise, conferindo um caráter temporal à representação do gesto. A figura 11 exemplifica como se dá a geração da representação de uma janela retangular de dimensão 10 *frames*, considerando a representação do gesto em um espaço bidimensional (duas características, como por exemplo, a velocidade e a aceleração escalares da mão direita). Cada *frame* é um dado do conjunto de dados e, nesse exemplo, um novo dado foi gerado a partir das características de 10 *frames* a partir da concatenação desses *frames*.

Figura 11 – Exemplo de geração de janela retangular de dimensão 10.

Fonte: Ricardo Alves Feitosa, 2018

Essa janela pode ser usada para representar, por exemplo, o *frame* #1, ou o *frame* #10 ou o *frame* #5, a depender se a informação temporal que se deseja considerar é anterior, posterior ou ambas. O *frame* representado pela janela é dito o *frame* de interesse.

A multidimensionalidade da representação para os gestos estudada neste trabalho também está relacionada com o uso de diferentes características descritivas para aspectos espaciais e temporais. Combinações dessas características comporão as diferentes representações multidimensionais a serem analisadas neste trabalho. Para isso, a seguinte estratégia de construção de representações foi adotada para suportar o estudo:

1. Representação utilizando apenas aspectos espaciais ou temporais. Nesse formato os diferentes aspectos não serão combinados entre si. Serão realizados testes para verificar a efetividades de cada aspecto isoladamente, com variações sobre o número de pontos de interesse do corpo submetidos ao algoritmo, variação da dimensão das janelas retangulares e aplicação da normalização e de filtros.
2. Representação utilizando combinação de aspectos espaciais e temporais existentes na base. Na geração dessas representações serão feitas combinações entre os aspectos já existentes na base, como por exemplo, a combinação das coordenadas x e y das mãos com as coordenadas x e y da cabeça, com variações sobre o número de pontos de interesse do corpo submetidos ao algoritmo, variação da dimensão das janelas retangulares e aplicação da normalização e de filtros.
3. Representação utilizando combinação de aspectos espaciais existentes na base com aspectos adicionais gerados. Nessa bateria de testes serão feitas combinações de coordenadas espaciais existentes na base com uma informação derivada desses dados, a angulação dos membros, com variações sobre o número de pontos de interesse do corpo submetidos ao algoritmo, variação da dimensão das janelas retangulares e aplicação da normalização e de filtros.

Além da representação do gesto, por considerar a modelagem do problema de segmentação das fases como uma tarefa de classificação, é necessário preparar a representação a ser usada para a rotulação de cada *frame*. Assim, após a decisão sobre o *frame* de interesse, associa-se a cada janela (ou seja, a cada dado que será processado pelo algoritmo classificador) o rótulo que está associado ao *frame* de interesse. A decisão sobre o *frame* de interesse pode variar entre escolher o *frame* inicial, o *frame* médio ou o último *frame*.

5.1.4 Descrição dos experimentos

Para a experimentação foram aplicados os dois algoritmos de aprendizado de máquina, MLP e SOM. Para os algoritmos, foram usadas implementações disponíveis no ambiente MATLAB[®], e para as demais rotinas (pré-processamento, construção das representações vetoriais e rotinas de avaliação de resultados) foram usadas rotinas implementadas pelo autor e também por membros do grupo de pesquisa no qual este trabalho está inserido.

Tanto o MLP quanto o SOM foram aplicados a representações multidimensionais conforme apresentado na subseção 5.1.3. Todos os vídeos da base de dados foram submetidos aos algoritmos e as seguintes representações foram testadas:

- coordenadas espaciais (x, y, z) de todos os pontos de interesse;
- coordenadas espaciais (x, y) da mão direita;
- coordenadas espaciais (x, y, z) da mão direita;
- velocidade e aceleração vetorial e escalar obtida para todos os pontos de interesse;
- velocidade e aceleração vetorial obtida para todos os pontos de interesse;
- velocidade escalar da mão direita;
- velocidade escalar de todos os pontos de interesse;
- coordenadas espaciais (x, y, z) de todos os pontos de interesse combinadas com a angulação formada entre as mãos e o tronco e a angulação entre o ponto central das mãos e seus respectivos pulsos;
- angulação formada entre as mãos e o tronco e a angulação entre o ponto central das mãos e seus respectivos pulsos;
- coordenadas espaciais (x, y, z) de todos os pontos de interesse combinadas com a angulação formada entre as mãos e o tronco.

Cada dado do conjunto de dados utilizado é um *frame* e cada *frame* é representado por um subconjunto de características, que são os pontos de interesse do corpo que foram rastreados e combinados para a construção das representações multidimensionais apresentadas. Portanto, essas características compõem as dimensões de cada *dataset* utilizado nos testes e, dessa forma, se um *frame* for representado por duas características ele terá duas dimensões.

Para essa bateria de testes a mão direita foi escolhida em algumas configurações por ser a mão dominante dos gesticuladores da base de dados. Para todas as configurações anteriores foram feitos testes com e sem a aplicação do filtro *wavelet* e para as configurações envolvendo aspectos espaciais foram feitos testes com e sem normalização. Para os aspectos temporais o uso de normalização não se fez necessário uma vez que foram disponibilizados na base de dados utilizada por esse trabalho já normalizados.

Seguindo a definição de janela de dados apresentada na subseção 5.1.3, foram realizados testes com janelas de tamanho 1, 20 e 40. O *frame* central foi utilizado como o *frame* de interesse em todos os testes por ter gerado os melhores resultados nos testes realizados por Madeo, Peres e Lima (2016).

Para os experimentos conduzidos com aprendizado supervisionado, ou seja, considerando o problema de segmentação das fases do gesto como uma tarefa de classificação, cinco implementações do MLP foram usadas com variações da *feedforwardnet*.

A implementação *feedforwardnet* é caracterizada como o modo *default* do MLP implementado pelo MATLAB[®], e permite que os parâmetros sejam configurados conforme necessidade da experimentação. Uma das configurações utilizadas é escolha do algoritmo de otimização para o *backpropagation* e, nesse trabalho, foram usadas variações contemplando a *feedforwardnet* com:

- gradiente descendente;
- gradiente descendente com taxa de aprendizado adaptativa;
- gradiente descendente com momentum;
- gradiente descendente com taxa de aprendizado adaptativa e momentum;
- Levenberg-Marquardt.

A figura 12 apresenta o esquema de funcionamento do MLP implementado como uma *feedforwardnet*. Com uma tangente sigmoide usada como função de ativação dos neurônios da camada oculta e uma função linear para ativação dos neurônios da camada de saída.

Figura 12 – Esquema da implementação *feedforwardnet*.

Fonte: Documentação do MATLAB®, 2017

Em todas as redes o bias foi inicializado com 1. O treinamento foi realizado com o método de *backpropagation*, com ajuste de pesos em lote sobre o erro quadrático médio e taxa de aprendizado inicial configurada para variar entre 3 valores, 0,1, 0,01 e 0,001. Como condições de parada no treinamento foi utilizado o valor do gradiente, caso alcance valor menor do que $1e - 5$, e o erro quadrático médio, caso seu valor não reduza por seis iterações consecutivas.

Para os experimentos nessas implementações, foram utilizadas três configurações de número de neurônios para a camada oculta: o valor de 10 neurônios, configuração *default* das implementações e utilizada empiricamente, o definido pela equação 12 proposta por Fausett (1994), onde N é o número de dados de entradas equivalente ao número de neurônios da camada de entrada da MLP, e o mesmo valor utilizado como tamanho da janela de dados, definição apresentada na subseção 5.1.3, com mais duas variações: uma com dois pontos a menos e outra com dois pontos a mais.

$$h(n) = 2 \times \sqrt{N} \quad (12)$$

Para todas as configurações de número de neurônios na camada oculta foram excluídas aquelas cujo cálculo resultou em valores menores do que um neurônio.

Como estratégia para divisão dos dados de entrada do classificador entre treinamento, validação e testes, foram selecionados subconjuntos de dados formados por trechos consecutivos de vídeo seguindo as estratégias:

- *k-fold* com $K = 10$ e $K = 3$.
- *holdout* com os dados divididos entre treinamento, validação e testes nas proporções:
 - 70-15-15; 15-70-15; 15-15-70;
 - 60-20-20; 20-60-20; 20-20-60;
 - 40-30-30; 30-40-30; 30-30-40.

Ainda sobre o k -fold, segundo Haykin (2009), essa é uma estratégia para trabalhar os dados de treinamento e validação. Os dados de testes não fazem parte dos $folds$, e devem ser selecionados separadamente. Assim, duas estratégias foram seguidas para obter um conjunto de testes para cada configuração apresentada. Inicialmente, os dados foram divididos em fo partes iguais, com fo sendo igual ao número de $folds + 1$. Por exemplo, na configuração de k -fold com $k = 10$ os dados foram divididos em 11 partes. Em seguida, a parte fo_{11} foi selecionada para representar o conjunto de testes na primeira execução do modelo classificado e uma das duas estratégias foram aplicadas: manter os dados em fo_{11} na sua posição original e fazer com que os demais $folds$ revezem entre treinamento e validação a cada execução do modelo classificador; incluir os dados de treinamento no revezamento entre os $folds$ dedicados ao treinamento e validação.

A figura 13 apresenta a primeira estratégia, exemplificando uma configuração de k -fold com $K = 10$ e uma parte fixa para testes, de forma que os dados para treinamento e validação rodam a cada execução do k -fold, mas a última parte dos dados permaneça reservada para testes. Na figura 14 é apresentada a segunda estratégia com o funcionamento do k -fold com $K = 10$ e uma parte dos dados reservada para testes que roda junto com os dados utilizados para treinamento e validação.

Figura 13 – Estratégia fixa: k -fold com $k = 10$. TR - treino; VL - validação; TE - teste.

Iteração	K-fold										Testes
	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11
i1	TR	TR	TR	TR	TR	TR	TR	TR	TR	VL	TE
i2	VL	TR	TR	TR	TR	TR	TR	TR	TR	TR	TE
i3	TR	VL	TR	TR	TR	TR	TR	TR	TR	TR	TE
...											TE
i10	TR	TR	TR	TR	TR	TR	TR	TR	VL	TR	TE

Fonte: Ricardo Alves Feitosa, 2018

Figura 14 – Estratégia cíclica: k -fold com $k = 10$. TR - treino; VL - validação; TE - teste.

Iteração	K-fold + Testes										
	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11
i1	TR	TR	TR	TR	TR	TR	TR	TR	TR	VL	TE
i2	TE	TR	TR	TR	TR	TR	TR	TR	TR	TR	VL
i3	VL	TE	TR	TR	TR	TR	TR	TR	TR	TR	TR
...											
i10	TR	TR	TR	TR	TR	TR	TR	TR	VL	TE	TR

Fonte: Ricardo Alves Feitosa, 2018

As implementações fornecem saídas no domínio dos números reais, limitadas no intervalo $[-1,1]$. Assim, para que fosse possível obter um resultado referente à classificação dos *frames* do vídeo dentro do conjunto de classes finito ($\{D, P, S, H, R\}$), duas estratégias foram adotadas:

1. Utilizando faixas de corte variadas para converter o formato numérico do domínio dos números reais, $\mathbb{R}$, para o domínio $-1, 1$, sendo que os limiares para as faixas de corte utilizadas foram 0,2, 0,1, 0,0, -0,1, -0,2 para a *feedforwardnet*. Dessa forma, para cada dado apresentado ao classificador, e para cada neurônio disposto na camada de saída do MLP, o valor no domínio $\mathbb{R}$ é mapeado para o valor 1 se ele for maior ou igual ao limiar da faixa de corte, caso contrário, será mapeado para o valor -1 .
2. Mapeando para o valor 1 a saída de maior valor dentre os cinco valores de saída do MLP, e para -1 os valores dos demais neurônios da camada de saída.

Diante da necessidade de tratar o problema de segmentação das fases do gesto como um problema de classificação multiclasse, como definido na subseção 2.2.2, uma codificação para ser usada como saída do MLP é feita usando um formato de cadeia de bits de tamanho cinco, sendo que cada bit é referente a uma das classes (ou fases do gesto). Nos experimentos do MLP foi usada uma codificação bipolar.

Para exemplificação, considere a classe D (ou “período de descanso”). Desde que se trata da primeira das cinco fases segundo a definição do problema, a primeira posição na cadeia de bits deverá ser valorada com 1, e as demais com -1 na representação bipolar, resultando na cadeia $[1 -1 -1 -1 -1]$.

A figura 15² apresenta uma organização das estratégias usadas durante os testes realizados com o MLP sobre os aspectos derivados dos seis pontos monitorados: mão esquerda, mão direita, pulso esquerdo, pulso direito, cabeça e tronco. Esse esquema foi aplicado a cada um dos sete vídeos disponíveis na base de dados utilizada por esse trabalho. Na figura 15 cada representação está acompanhada da dimensão do *dataset* utilizado.

² No ambiente de hardware usado para esse trabalho não foi possível executar a configuração com janela de tamanho 40 e mais de 18 características que descrevem os aspectos do gesto.

Figura 15 – Organização de testes para o MLP.

Fonte: Ricardo Alves Feitosa, 2018

Por exemplo, um dos casos de testes foi utilizar uma representação construída com todas as coordenadas da mão direita, sobre a qual foi aplicado um pré-processamento com filtro *wavelet* e com normalização para então ser processada por um MLP configurado como uma rede *feedforward* com Levenberg-Marquardt, taxa de aprendizado de 0,01, 10 neurônios na camada oculta e a divisão dos dados entre treinamento, validação e testes definida com o uso de um k -fold com $k = 10$.

Para os experimentos conduzidos com aprendizado não supervisionado, ou seja, considerando o problema de segmentação das fases do gesto como uma tarefa de agrupamento, uma implementação da rede SOM disponibilizada na SOM Toolbox³ e executada dentro do ambiente do MATLAB[®] foi utilizada. Os parâmetros da rede SOM, assim como foi com o MLP, também sofreram variações nos experimentos. Foram utilizados mapas variando em 1x5, 1x10, 10x10, 20x20 e a configuração *default* com cálculo automático de tamanho de mapa. Essa configuração *default* calcula o número de unidades do mapa seguindo o proposto por Vesanto et al. (2000) na equação 13, onde d_{lem} é o número de dados de entrada. As dimensões do mapa são determinadas pelos dois maiores autovalores dos dados

³ Disponível em: <http://www.cis.hut.fi/projects/somtoolbox/>.

de treinamento, de forma que os comprimentos x e y do mapa sejam proporcionais a esses valores e o produto entre eles seja o mais próximo possível do $mUnits$.

$$mUnits = 20 \times dlem^{0,54321} \quad (13)$$

Com os mapas 1x5 o objetivo foi verificar o quão próximo da classificação especialista uma execução com SOM poderia chegar, verificando a quais conclusões um aprendizado não supervisionado chegaria quando limitado ao mesmo número de cinco saídas possíveis, sendo cinco o número de fases do gesto consideradas para a classificação. Nos demais mapas a expectativa foi verificar quais grupos foram formados pelo aprendizado não supervisionado e, a partir dos resultados, analisar que grupos poderiam ser vistos como sugestões para novas fases do gesto.

Também foram consideradas variações para a taxa de aprendizado e o número de épocas. Para a taxa de aprendizado as variações foram entre 0,01, 0,05 e 1 para as fases de ordenação e convergência e o *default* de 0,5 para a fase de ordenação com 0,05 para a fase de convergência. Para o número de épocas, a variação foi entre 100 e 1000 na fase de ordenação e o *default* de 4 vezes *mpd* para a fase de ordenação e 16 vezes *mpd* para a fase de convergência, sendo que $mpd = \text{número de unidades do mapa} / \text{número de dados usados para treinamento}$.

Para o treinamento, duas configurações de algoritmos foram aplicadas: uma em lote, que é a configuração *default*, e uma sequencial. Na configuração em lote, as atualizações de pesos são realizadas quando todos os dados de entrada já foram apresentados ao mapa. Na configuração sequencial, o neurônio vencedor para uma dada entrada tem seus pesos calculados no momento em que foi descoberto como vencedor. De acordo com Kohonen (2001), no método sequencial existe uma quantidade maior de cálculos e dependência sobre a ordem dos dados de entrada, enquanto o método em lote é, comparativamente, mais rápido para concluir os cálculos e não depende da ordem dos dados de entrada.

Os demais parâmetros foram determinados seguindo os valores *default* de três neurônios como raio de vizinhança inicial, topologia hexagonal do mapa, distância matricial Manhattan e distância vetorial Euclidiana. A estratégia de utilizar algumas configurações como *default* se deve ao fato de variações nesses parâmetros aumentarem os erros topológicos e de quantização da rede em testes não exaustivos que foram feitos com a base de dados utilizada por esse trabalho. Aqui salienta-se o fato de terem sido testes não exaustivos

uma vez que o objetivo desse trabalho é avaliar o quão discriminante são as diferentes representações utilizadas para as fases do gesto, usando os algoritmos MLP e SOM como ferramentas e não explorando suas possibilidades de otimização. Aqui já é possível adiantar a oportunidade de trabalhos futuros explorando esses algoritmos quanto ao seu potencial de melhorar os resultados com a otimização dos parâmetros.

Na figura 16 estão organizadas as estratégias de testes realizados com o SOM sobre os aspectos derivados dos mesmos seis pontos do corpo monitorados e utilizados pelo MLP: mão esquerda, mão direita, pulso esquerdo, pulso direito, cabeça e tronco. Esse esquema também foi aplicado para cada um dos sete vídeos disponíveis na base de dados utilizada por esse trabalho. Na figura 16 cada representação está acompanhada da dimensão do *dataset* utilizado.

Figura 16 – Organização de testes para o SOM.

Fonte: Ricardo Alves Feitosa, 2018

Por exemplo, um dos casos de testes foi a utilização de uma representação com todos os dados de velocidade e aceleração, sobre a qual foi aplicado um pré-processamento com filtro *wavelet*, para, em seguida, ser processada por um SOM configurado para corrigir o erro em lote, com tamanho do mapa definido para 10x10, 100 épocas para a fase de ordenação e convergência e taxa de aprendizado de 0,05.

Seguindo a execução dos testes apresentados nos esquemas das figuras 15 e 16, cada caso de teste representou um conjunto de dados submetido à uma configuração do MLP e do SOM em formato de matriz de características, nesse trabalho também chamado de representação para as fases do gesto.

5.2 Experimentos com MLP

Conforme apresentado na subseção 5.1.4, o MLP foi executado em rotinas dentro do ambiente MATLAB[®]. E, conforme descrito na seção 1.3, os resultados foram avaliados quanto às métricas de precisão, revocação, acurácia e F1-Score e com a análise das matrizes de confusão formada pela quantidade de dados presentes em cada classe predita pelo classificador contra a quantidade de dados presentes na classe esperada para cada configuração de execução do MLP. Com esses experimentos, foram verificados quais os modelos de representação do gesto e quais configurações do MLP produziram os resultados mais próximos do esperado na segmentação das fases do gesto.

Como o problema foi modelado como multiclasse, o cálculo das métricas de avaliação foi realizado por classe. Apenas a acurácia foi possível obter, também, como uma métrica para a configuração total, considerando todas as classes, com o uso da equação 14. Essa equação divide a soma dos dados da diagonal principal pela soma de todos os dados da matriz de confusão.

$$accu_{tot} = \frac{\sum_5^{i=1} x_{ii}}{\sum_5^{i=1} (\sum_5^{j=1} y_{ij})} \quad (14)$$

Cada conjunto de resultados referente à um caso de teste possui:

- as medidas de precisão, revocação, acurácia e F1-Score para cada classe (fase do gesto) resultante da classificação;
- a medida de acurácia total do MLP;
- a classe gerada pelo MLP para cada dado da representação;
- uma matriz de confusão formada pela quantidade de dados presentes em cada classe obtida contra a quantidade de dados em cada classe esperada.

Para os testes onde a divisão de dados entre treinamento, validação e testes foi realizada com o *k-fold*, cada *fold* foi analisado isoladamente. Nos resultados que serão

apresentados, os aspectos do gesto que foram utilizados para cada teste estarão no formato matricial relacionados na tabela 2.

Tabela 2 – Relação entre o formato matricial e os aspectos do gesto nele representados.

Aspectos espaciais - Coordenadas espaciais	
1x18	Todas as coordenadas espaciais
4x5	Só x e y da mão direita
4x6	Todas as coordenadas da mão direita
Aspectos espaciais - Coordenadas espaciais e ângulos	
1x22	Todos os dados
19x22	Só os ângulos
1x20	Todos os ângulos menos o formado entre o pulso e o ponto central da mão
Aspectos temporais	
1x32	Todos os dados de velocidade e aceleração
1x24	Só a velocidade vetorial e aceleração vetorial
26x26	Só a velocidade escalar da mão direita
25x32	Só a velocidade escalar e aceleração escalar

As próximas subseções apresentam os resultados obtidos e estão organizadas por aspectos do gesto, começando com uma subseção sobre os resultados obtidos com as características do gesto relacionadas à aspectos temporais, que foram as que apresentaram os melhores resultados quanto à acurácia total do classificador. Essa subseção começará apresentando os melhores resultados para cada vídeo. E, ainda nessa subseção, as análises seguem pelo vídeo a1 por ter alcançado os melhores resultados, com recortes sobre as características do gesto e a configuração utilizada pelo MLP. A subseção seguinte explora os aspectos espaciais que utilizam apenas as coordenadas dos membros monitorados na construção das representações e, também, explorando o impacto da adição da característica de angulação sobre o resultado obtido pelo classificador.

5.2.1 Aspectos temporais

Com a execução dos testes foram gerados 564.288 resultados para o MLP. Essa quantidade de modelos é proibitiva para uma análise individual de cada resultado, exigindo que os dados fossem tabulados e ordenados segundo algum critério. O critério utilizado foi ordenação decrescente da acurácia total. Como uma primeira seleção para as análises, foram separados os 5 resultados com melhor acurácia para cada vídeo, dispostos na tabela 3⁴.

⁴ As abreviaturas do cabeçalho das tabelas estão descritas na lista de abreviaturas e siglas.

Tabela 3 – Melhores resultados identificados quanto a acurácia total por vídeo.

#	Vídeo	Corte	Ângulo	Filtro	Norm	AspG	NeOc	DivDados	Acc_Total
Vídeo a1									
1-1	a1-va3	maior	-	<i>wavelet</i>	-	1x32	51	<i>k-fold</i> - $k = 10$	0,95
1-2	a1-va3	maior	-	-	-	1x24	44	<i>k-fold</i> - $k = 10$	0,94
1-3	a1-raw	maior	-	<i>wavelet</i>	-	1x18	38	<i>k-fold</i> - $k = 10$	0,93
1-4	a1-raw	maior	sim	<i>wavelet</i>	sim	1x22	42	<i>k-fold</i> - $k = 10$	0,93
1-5	a1-raw	maior	-	-	sim	1x18	38	<i>k-fold</i> - $k = 10$	0,93
Vídeo a2									
1-6	a2-va3	maior	-	<i>wavelet</i>	-	1x32	11	<i>k-fold</i> - $k = 10$	0,80
1-7	a2-va3	maior	-	<i>wavelet</i>	-	1x24	10	<i>k-fold</i> - $k = 10$	0,79
1-8	a2-va3	-0,1	-	<i>wavelet</i>	-	1x32	11	<i>k-fold</i> - $k = 10$	0,78
1-9	a2-va3	0	-	<i>wavelet</i>	-	1x32	11	<i>k-fold</i> - $k = 10$	0,78
1-10	a2-va3	-0,2	-	<i>wavelet</i>	-	1x32	11	<i>k-fold</i> - $k = 10$	0,78
Vídeo a3									
1-11	a3-raw	maior	-	<i>wavelet</i>	-	1x18	8	<i>holdout</i> - 602020	0,74
1-12	a3-raw	maior	-	<i>wavelet</i>	-	1x18	8	<i>k-fold</i> - $k = 10$	0,73
1-13	a3-raw	maior	sim	-	-	1x20	9	<i>holdout</i> - 701515	0,71
1-14	a3-raw	maior	sim	-	sim	1x22	9	<i>holdout</i> - 701515	0,71
1-15	a3-raw	maior	sim	-	-	1x20	9	<i>k-fold</i> - $k = 10$	0,70
Vídeo b1									
1-16	b1-raw	maior	sim	-	-	1x20	9	<i>k-fold</i> - $k = 10$	0,74
1-17	b1-raw	maior	sim	<i>wavelet</i>	sim	1x20	9	<i>k-fold</i> - $k = 10$	0,73
1-18	b1-raw	maior	-	-	-	1x18	8	<i>k-fold</i> - $k = 10$	0,71
1-19	b1-raw	maior	sim	-	-	1x22	9	<i>k-fold</i> - $k = 10$	0,71
1-20	b1-va3	maior	-	<i>wavelet</i>	-	1x24	10	<i>k-fold</i> - $k = 10$	0,70
Vídeo b2									
1-21	b3-raw	maior	sim	-	-	1x20	9	<i>k-fold</i> - $k = 10$	0,72
1-22	b3-raw	maior	sim	<i>wavelet</i>	-	1x22	9	<i>k-fold</i> - $k = 10$	0,72
1-23	b3-raw	maior	sim	-	-	1x22	9	<i>k-fold</i> - $k = 10$	0,71
1-24	b3-raw	maior	sim	-	sim	1x20	9	<i>k-fold</i> - $k = 10$	0,70
1-25	b3-raw	maior	sim	-	-	1x22	9	<i>k-fold</i> - $k = 10$	0,70
Vídeo c1									
1-26	c1-raw	maior	sim	<i>wavelet</i>	sim	1x22	9	<i>k-fold</i> - $k = 10$	0,78
1-27	c1-raw	maior	sim	<i>wavelet</i>	sim	1x20	9	<i>k-fold</i> - $k = 10$	0,76
1-28	c1-va3	maior	-	<i>wavelet</i>	-	1x32	11	<i>holdout</i> - 701515	0,76
1-29	c1-raw	maior	-	<i>wavelet</i>	-	1x18	8	<i>k-fold</i> - $k = 10$	0,76
1-30	c1-va3	maior	-	<i>wavelet</i>	-	1x32	11	<i>k-fold</i> - $k = 10$	0,75
Vídeo c3									
1-31	c3-raw	maior	sim	<i>wavelet</i>	-	1x22	9	<i>k-fold</i> - $k = 10$	0,77
1-32	c3-raw	maior	-	-	-	1x18	8	<i>holdout</i> - 701515	0,73
1-33	c3-raw	-0,1	sim	<i>wavelet</i>	-	1x22	9	<i>k-fold</i> - $k = 10$	0,73
1-34	c3-raw	maior	sim	-	sim	1x22	9	<i>holdout</i> - 701515	0,72
1-35	c3-raw	-0,2	sim	<i>wavelet</i>	-	1x22	9	<i>k-fold</i> - $k = 10$	0,72

Esses resultados foram todos obtidos com o uso de uma *feedforwardnet* com Levenberg-Marquardt e taxa de aprendizado de 0,01 e conjunto de testes cíclico para o *k-fold*. Nota-se uma ligeira distância entre as acurácias totais do vídeo a1 para o segundo colocado, o vídeo a2, de 0,12, uma diferença considerável ao salientar que 0,8 é a melhor acurácia do vídeo a2. Isso indica que as melhores características para o vídeo a1 não coincidem com as melhores características para os demais vídeos da base. O vídeo a1 alcançou os melhores resultados de acurácia com as características do gesto representadas por aspectos temporais, enquanto os melhores resultados dos vídeos de a3 até c3 foram obtidos com características representadas com aspectos espaciais em diferentes combinações.

Como salientando no início dessa seção, as próximas análises seguirão sobre o vídeo a1. A tabela 4 elenca os 20 melhores resultados obtidos dentre todos os experimentos que foram realizados para esse vídeo.

Tabela 4 – Melhores resultados identificados quanto a acurácia total para o vídeo a1.

#	Vídeo	Corte	Ângulo	Filtro	Norm	AspG	NeOc	Acc_Total
2-1	a1-va3	maior	-	<i>wavelet</i>	-	1x32	51	0,95
2-2	a1-va3	maior	-	-	-	1x24	44	0,94
2-3	a1-raw	maior	-	<i>wavelet</i>	-	1x18	38	0,93
2-4	a1-raw	maior	sim	<i>wavelet</i>	sim	1x22	42	0,93
2-5	a1-raw	maior	-	-	sim	1x18	38	0,93
2-6	a1-raw	maior	sim	-	-	1x22	42	0,93
2-7	a1-va3	0,2	-	<i>wavelet</i>	-	1x32	51	0,92
2-8	a1-raw	maior	sim	<i>wavelet</i>	sim	1x20	40	0,92
2-9	a1-va3	0,1	-	<i>wavelet</i>	-	1x32	51	0,92
2-10	a1-raw	0	sim	<i>wavelet</i>	sim	1x22	42	0,92
2-11	a1-va3	0	-	<i>wavelet</i>	-	1x32	51	0,92
2-12	a1-va3	-0,1	-	<i>wavelet</i>	-	1x32	51	0,92
2-13	a1-raw	0,1	sim	<i>wavelet</i>	sim	1x22	42	0,92
2-14	a1-raw	maior	sim	-	-	1x20	40	0,92
2-15	a1-raw	maior	-	-	-	1x18	38	0,92
2-16	a1-va3	-0,2	-	<i>wavelet</i>	-	1x32	51	0,92
2-17	a1-raw	maior	sim	-	sim	1x22	42	0,92
2-18	a1-va3	-0,1	-	-	-	1x24	44	0,92
2-19	a1-va3	0	-	-	-	1x24	44	0,92
2-20	a1-raw	-0,1	sim	<i>wavelet</i>	sim	1x22	42	0,91

Esses 20 melhores resultados foram todos obtidos com o uso de uma *feedforwardnet* com Levenberg-Marquardt, taxa de aprendizado de 0,01, janela de tamanho 20 e divisão dos dados entre treinamento, validação e testes com o uso de um *k-fold* com $k = 10$ e conjunto de testes cíclico. Como essa é a lista dos melhores resultados gerados, há representações que utilizam aspectos espaciais e temporais. Os aspectos espaciais serão explorados nas próximas seções.

Analizando apenas a acurácia, os melhores resultados foram todos obtidos sobre o vídeo a1. Como o treinamento, validação e testes foram realizados com dados de um mesmo vídeo, pode-se afirmar que o a1 tem uma configuração que torna a classificação mais fácil de ser realizada do que os demais vídeos.

Os melhores resultados foram obtidos com o uso de todas as características de velocidade e aceleração, independentemente de serem configurações que misturam características vetoriais e escalares, resultado #2-1. Quando se removeu as características de velocidade e aceleração escalar, resultado #2-2, a acurácia teve uma ligeira redução, mas ainda ficou à frente de execuções do MLP que utilizaram como faixa de corte um critério diferente de escolher o neurônio de saída com o maior valor para representar a classe.

É interessante observar também que o uso do filtro teve impacto positivo no resultado #2-1, quando todas as informações de velocidade e aceleração foram utilizadas,

mas não apresentou essa mesma melhoria quando os dados escalares foram dispensados no resultado #2-2. Assim, o uso de filtros mostrou eficiência em deixar as representações mais discriminantes quando são misturados diferentes tipos de aspectos na construção da representação.

Mas, além da acurácia, deve-se analisar também as demais métricas de avaliação. A métrica analisada na sequência foi a acurácia por fase. A tabela 5 expande os resultados de acurácia para as fases do gesto.

Tabela 5 – Melhores resultados identificados quanto a acurácia por fase para o vídeo a1.

#	Acc_D	Acc_P	Acc_S	Acc_H	Acc_R
2-1	0,99	0,91	0,91	1,00	0,93
2-2	1,00	0,85	0,94	0,82	0,88
2-3	0,97	0,86	0,95	0,49	0,90
2-4	0,99	0,88	0,90	1,00	0,88
2-5	0,98	0,84	0,90	1,00	0,90
2-6	0,97	0,86	0,92	0,82	0,88
2-7	0,97	0,91	0,88	1,00	0,88
2-8	0,94	0,87	0,93	1,00	0,86
2-9	0,97	0,91	0,88	1,00	0,88
2-10	0,97	0,88	0,89	1,00	0,88
2-11	0,98	0,91	0,87	1,00	0,89
2-12	0,97	0,90	0,87	1,00	0,92
2-13	0,97	0,88	0,88	1,00	0,88
2-14	0,98	0,85	0,91	0,64	0,86
2-15	0,93	0,90	0,93	0,69	0,89
2-16	0,97	0,90	0,86	1,00	0,92
2-17	0,95	0,88	0,89	1,00	0,88
2-18	0,99	0,82	0,90	0,79	0,83
2-19	0,97	0,83	0,91	0,77	0,84
2-20	0,96	0,88	0,88	1,00	0,88

É nessa análise que a diferença entre o uso de todas as características do aspecto temporal mostra seu benefício em relação às características apenas vetoriais. O resultado #2-1 conseguiu manter o equilíbrio em altos valores de acurácia para todas as fases do gesto, alcançando 100% para a fase de *hold*. Quando as características escalares foram removidas da representação, esse equilíbrio foi penalizado, reduzindo as acurácias das fases de preparação, *hold* e retração e aumentando levemente as acurácias das fases de descanso e *stroke*.

Não obstante as diferenças entre os melhores resultados, todos foram obtidos com o uso do método de cálculo de neurônios da camada oculta proposto por Fausett (1994) na equação 12, onde a quantidade de neurônios é diretamente proporcional à quantidade de dados de entrada.

Consequente à análise das acurácias, o F1-Score foi analisado para as 10 representações elencadas com as melhores acurácias. A tabela 6 apresenta os valores obtidos. É interessante observar que os valores também se mantiveram altos para essa métrica.

Tabela 6 – Melhores resultados identificados quanto ao F1-Score por fase seguindo a ordenação por acurácia total.

#	FSc_D	FSc_P	FSc_S	FSc_H	FSc_R
2-1	0,98	0,85	0,94	0,89	0,93
2-2	0,96	0,89	0,96	0,85	0,89
2-3	0,97	0,79	0,96	0,66	0,90
2-4	0,94	0,86	0,95	1,00	0,89
2-5	0,95	0,82	0,95	0,94	0,88
2-6	0,96	0,82	0,95	0,78	0,86
2-7	0,98	0,92	0,93	1,00	0,93
2-8	0,93	0,78	0,96	1,00	0,90
2-9	0,98	0,91	0,93	1,00	0,92
2-10	0,95	0,91	0,94	1,00	0,93

Para o F1-Score, os resultados foram semelhantes aos obtidos pela acurácia por fase, com exceção do equilíbrio entre as fases do resultado #2-1 que, no F1-Score, estão mais próximos da proporção apresentada no resultado #2-2, mostrando que a escolha entre utilizar as características de velocidade e aceleração escalares na construção da representação para reconhecer as fases do gesto alcançaram um resultado melhor do classificador. Porém esse ganho deve ser avaliado quanto ao custo versus benefício do emprego de mais recursos computacionais para seu processamento.

Em certas aplicações, pode ser desejável que o algoritmo apenas reconheça a diferença entre as fases de *stroke* e *hold* ou descanso, para identificar picos de movimento contra momentos em que não há movimento. A tabela 7 apresenta os melhores resultados para essas fases analisando apenas o F1-Score de cada fase, todos obtidos com o uso de uma *feedforwardnet* com Levenberg-Marquardt, taxa de aprendizado de 0,01, janela de tamanho 20 e divisão dos dados entre treinamento, validação e testes com o uso de um *k-fold* com $k = 10$ e conjunto de testes cíclico.

Tabela 7 – Melhores resultados identificados quanto ao F1-Score para as fases de *stroke*, descanso e *hold*.

#	Vídeo	Corte	Ângulo	Filtro	Norm	AspG	NeOc	FSc_D	FSc_P	FSc_S	FSc_H	FSc_R
Melhores resultados por F1-Score do <i>Stroke</i>												
3-1	al-va3	maior	-	-	-	1x24	44	0,96	0,89	0,96	0,85	0,89
3-2	al-raw	maior	sim	<i>wavelet</i>	sim	1x20	40	0,93	0,78	0,96	1,00	0,90
3-3	al-raw	maior	-	<i>wavelet</i>	-	1x18	38	0,97	0,79	0,96	0,66	0,90
3-4	al-raw	maior	sim	<i>wavelet</i>	-	1x22	42	0,89	0,69	0,95	0,93	0,82
3-5	al-raw	maior	sim	-	-	1x20	40	0,95	0,78	0,95	0,69	0,87
Melhores resultados por F1-Score do Descanso												
4-1	al-va3	0	-	<i>wavelet</i>	-	1x32	51	0,99	0,92	0,93	0,96	0,93
4-2	al-va3	0,1	-	<i>wavelet</i>	-	1x32	51	0,98	0,91	0,93	1,00	0,92
4-3	al-va3	-0,1	-	<i>wavelet</i>	-	1x32	51	0,98	0,92	0,93	0,91	0,93
4-4	al-va3	-0,2	-	<i>wavelet</i>	-	1x32	51	0,98	0,91	0,93	0,93	0,94
4-5	al-va3	maior	-	<i>wavelet</i>	-	1x32	51	0,98	0,85	0,94	0,89	0,93
Melhores resultados por F1-Score do <i>Hold</i>												
5-1	al-va3	0,1	-	<i>wavelet</i>	-	1x32	51	0,98	0,91	0,93	1,00	0,92
5-2	al-va3	0,2	-	<i>wavelet</i>	-	1x32	51	0,98	0,92	0,93	1,00	0,93
5-3	al-raw	0	sim	<i>wavelet</i>	sim	1x22	42	0,95	0,91	0,94	1,00	0,93
5-4	al-raw	maior	sim	<i>wavelet</i>	sim	1x22	42	0,94	0,86	0,95	1,00	0,89
5-5	al-raw	0,1	sim	<i>wavelet</i>	sim	1x22	42	0,94	0,92	0,94	1,00	0,93

Apesar de não haver nenhuma configuração que obteve 100% para essa métrica na fase de *stroke*, valores altos foram alcançados de até 0,96. O mesmo ocorreu para a fase de descanso, chegando a valores muito próximos do 100%, inclusive com um destaque para o resultado #4-2 que conseguiu 0,98 de F1-Score para a fase de descanso e 1 para a fase de *hold*, mantendo valores altos para as demais classes também. Explorando os resultados para a fase de *hold*, os 5 melhores resultados conseguiram 100% para o F1-Score.

Quanto ao tamanho de janela, nota-se que os dados com janelas de tamanho 20 e 40 ocupam as posições de maior acurácia dos testes, mas não se pode afirmar que os dados com janela de tamanho 1 possuam desempenho ruim em termos de acurácia, já que o melhor valor foi encontrado na posição #1-53 com uma acurácia total de 0,87, seguindo esse mesmo critério de ordenação. Porém, cabe salientar que as acurácias não foram equilibradas entre todas as fases do gesto, ficando com 0,66, 0,33 e 0,66 para a preparação, *hold* e retração, respectivamente, mas conseguindo 0,98 e 0,91 para as fases de descanso e *stroke*.

Assim, retoma-se a necessidade de análise do custo versus benefício da quantidade de dados apresentados ao MLP contra o resultado final da execução, já que as execuções do algoritmo sobre dados contendo janela de tamanho 20 e 40 consomem tempo computacional significativamente maior do que uma mesma configuração de aspectos com tamanho de janela igual a 1.

Os resultados apresentados mostram a eficiência do MLP na identificação das fases do gesto representadas utilizando mais de 18 aspectos do gesto, independente do uso de aspectos espaciais ou temporais. A tabela 8 elenca os cinco melhores resultados por acurácia total para modelos de representação com menos de 18 aspectos para o gesto.

Tabela 8 – Acurácias e F1-Scores dos melhores resultados identificados quanto a acurácia total para modelos de representação com menos de 18 aspectos para o gesto.

#	Vídeo	Corte	Ângulo	Filtro	Norm	Jal	AspG	Rede	TxA	NeOc	DivDados
6-1	a1-va3	maior	-	<i>wavelet</i>	-	20	25x32	Ffnet: L-M	0,01	25	<i>k-fold</i> - $k = 10$
6-2	a1-va3	maior	-	-	-	20	25x32	Ffnet: L-M	0,01	25	<i>k-fold</i> - $k = 10$
6-3	a1-raw	maior	sim	<i>wavelet</i>	-	20	19x22	Ffnet: L-M	0,01	18	<i>k-fold</i> - $k = 10$
6-4	a1-raw	maior	sim	<i>wavelet</i>	sim	20	19x22	Ffnet: L-M	0,01	18	<i>k-fold</i> - $k = 10$
6-5	a1-raw	maior	-	-	sim	20	4x6	Ffnet: L-M	0,01	15	<i>holdout</i> - 701515
#	Acc_Total	Acc_D	Acc_P	Acc_S	Acc_H	Acc_R	FSc_D	FSc_P	FSc_S	FSc_H	FSc_R
6-1	0,90	0,97	0,71	0,95	0,26	0,76	0,94	0,74	0,94	0,40	0,81
6-2	0,89	0,96	0,70	0,97	0,54	0,62	0,92	0,78	0,93	0,70	0,74
6-3	0,87	0,96	0,67	0,84	0,77	0,81	0,92	0,72	0,89	0,85	0,76
6-4	0,86	0,98	0,57	0,88	0,74	0,62	0,87	0,70	0,90	0,85	0,76
6-5	0,85	0,99	0,61	0,87	0,26	0,60	0,91	0,64	0,89	0,41	0,69

Para essas configurações, foram obtidos resultados com altos valores de acurácia e F1-Score, com os cinco melhores resultados tendo representantes nas 3 estratégias de utilização dos aspectos do gesto, os espaciais somente com coordenadas, os espaciais com angulação e os temporais. As representações que foram construídas com aspectos temporais mantiveram as primeiras colocações quanto a acurácia, assim como na tabela 4, seguidas pelas representações que usaram angulação e pelas que utilizaram apenas as coordenadas, mas com não mais de 0,01 de diferença entre um resultado e seu antecessor.

Porém, ainda comparando com os melhores resultados gerais da tabela 4, com menos de 18 aspectos o resultado para esses modelos não foi equilibrado entre todas as fases do gesto. Analisando tanto as acurácias quanto os F1-Scores notou-se que sempre há fases com resultados abaixo da média, como no próprio melhor resultado #6-1, com acurácia de apenas 0,26 para a fase de *hold* e F1-Score de 0,4.

Também é importante destacar os resultados que foram obtidos com apenas uma ou duas características para construção das representações do gesto. A tabela 9 enumera esses resultados.

Tabela 9 – Acurácias e F1-Scores dos melhores resultados identificados quanto a acurácia total para modelos de representação com um ou dois aspectos para o gesto.

#	Vídeo	Corte	Ângulo	Filtro	Norm	Jal	AspG	Rede	TxA	NeOc	DivDados
7-1	a1-raw	maior	-	-	sim	20	4x5	Ffnet: L-M	0,01	13	<i>k-fold</i> - <i>k</i> = 10
7-2	a1-va3	maior	-	<i>wavelet</i>	-	20	26x26	Ffnet: L-M	0,01	9	<i>k-fold</i> - <i>k</i> = 10
#	Acc.Total	Acc.D	Acc.P	Acc.S	Acc.H	Acc.R	FSc.D	FSc.P	FSc.S	FSc.H	FSc.R
7-1	0,83	0,98	0,49	0,86	0,00	0,66	0,91	0,55	0,87	0,00	0,69
7-2	0,76	0,87	0,47	0,93	0,31	0,15	0,83	0,54	0,83	0,44	0,24

Nesse conjunto de resultado, o modelo construído a partir das coordenadas *x* e *y* da mão direita obteve acurácia total maior do que a obtida pelo modelo que utilizou apenas a velocidade escalar da mão direita, resultados #7-1 e #7-2 respectivamente. Isso se for analisada apenas a acurácia total, de 0,83 contra 0,76. Analisando o resultado por fase, o obtido com o modelo do resultado #7-2 não anulou nenhuma fase do gesto, enquanto o #7-1 ficou com zero de acurácia e F1-Score para a fase de *hold*, ainda que para ambos os resultados a identificação do dueto descanso e *stroke* tenha sido realizado com boas taxas de acurácia e F1-Score.

Por conseguinte, é possível afirmar que uma representação utilizando apenas um ou dois aspectos do gesto é suficiente para a atividade de identificação do gesto e não-gesto. Esse é um cenário semelhante ao identificado na análise das tabelas 4 e 5, quando comparados os resultados obtidos com janelas de tamanho 1, 20 e 40, nos quais o resultado

com janelas de tamanho 1 também se mostraram promissores para a identificação das fases de descanso e *stroke*, com acurácias e F1-Score até melhores do que os apresentados na tabela 9.

5.2.2 Aspectos espaciais

Apesar do melhor resultado ter sido obtido com o uso dos aspectos temporais, em termos de acurácia eles não se distanciaram dos obtidos utilizando os aspectos espaciais. A diferença do resultado #2-1 para o #2-3, que utilizou todas as coordenadas espaciais, foi de apenas 0,02 de acurácia, a mesma diferença para o resultado #2-4 que utilizou a informação de angulação combinada com todas as coordenadas espaciais. A menor quantidade de características que produziu o melhor resultado dessa seleção foi o apresentado pelo #2-3, com a aplicação do filtro *wavelet*.

A utilização do filtro para uma mesma configuração não se mostrou benéfica para todos os resultados, de forma que ela deve ser analisada caso a caso. Analisando o resultado #2-3 e #2-5 por exemplo, a aplicação do filtro gerou o mesmo resultado na acurácia. O mesmo vale para a normalização, como entre os resultados #2-4 e #2-6 nos quais a combinação da normalização com o filtro não elevaram significativamente a acurácia.

Essa visão é interessante para uma análise de custo versus benefício, se for considerado que, quanto maior a dimensionalidade do espaço de representação do gesto, maior o custo computacional do classificador. Portanto, é interessante verificar que utilizando menos recursos computacionais em relação aos melhores resultados, é possível obter acurácias muito próximas dos primeiros colocados.

Analisando os resultados de acurácia por fases do gesto, o resultado #2-3 mostrou dificuldade para identificar a fase de *hold*, enquanto o seu par, resultado #2-5, conseguiu resultados mais equilibrados entre as fases quando não utilizou o filtro, chegando a 100% de acurácia para essa mesma fase. Complementando a análise da fase de *hold*, nos resultados ela foi a mais bem identificada nas execuções, alcançando o 100% de acurácia em muitos dos testes, mostrando-se uma fase cuja ausência de movimentos significa uma característica discriminante para os classificadores. A despeito de não ser possível afirmar que alguma fase do gesto teve grande dificuldade de ser reconhecida pelos 20 melhores resultados, mesmo com o apresentado pela fase de *hold* do caso #2-3, que se mostrou um resultado isolado.

Quanto ao F1-Score elencado na tabela 6, o resultado #2-3 reflete no F1-Score para identificação do *hold* a mesma dificuldade já percebida na acurácia. A despeito disso, vários outros resultados conseguiram 100% também de F1-Score para essa classe.

Separando na tabela 10 as métricas de precisão e revocação para os mesmos casos da tabela 7, é visível que existe certo equilíbrio entre as duas métricas.

Tabela 10 – Precisão e revocação dos melhores resultados identificados quanto ao F1-Score para as fases de *stroke*, descanso e *hold*.

#	Pre_D	Pre_P	Pre_S	Pre_H	Pre_R	Rec_D	Rec_P	Rec_S	Rec_H	Rec_R
Precisão e revocação dos melhores resultados por F1-Score do <i>Stroke</i>										
3-1	0,92	0,94	0,99	0,89	0,89	1,00	0,85	0,94	0,82	0,88
3-2	0,91	0,71	0,99	1,00	0,95	0,94	0,87	0,93	1,00	0,86
3-3	0,96	0,73	0,97	1,00	0,90	0,97	0,86	0,95	0,49	0,90
3-4	0,97	0,54	0,98	1,00	0,81	0,82	0,97	0,93	0,87	0,84
3-5	0,92	0,72	1,00	0,76	0,87	0,98	0,85	0,91	0,64	0,86
#	Pre_D	Pre_P	Pre_S	Pre_H	Pre_R	Rec_D	Rec_P	Rec_S	Rec_H	Rec_R
Precisão e revocação dos melhores resultados por F1-Score do Descanso										
4-1	0,99	0,93	0,99	0,93	0,97	0,98	0,91	0,87	1,00	0,89
4-2	0,99	0,91	0,99	1,00	0,97	0,97	0,91	0,88	1,00	0,88
4-3	0,99	0,94	1,00	0,83	0,95	0,97	0,90	0,87	1,00	0,92
4-4	0,99	0,92	1,00	0,87	0,96	0,97	0,90	0,86	1,00	0,92
4-5	0,97	0,79	0,98	0,80	0,94	0,99	0,91	0,91	1,00	0,93
#	Pre_D	Pre_P	Pre_S	Pre_H	Pre_R	Rec_D	Rec_P	Rec_S	Rec_H	Rec_R
Precisão e revocação dos melhores resultados por F1-Score do <i>Hold</i>										
5-1	0,99	0,91	0,99	1,00	0,97	0,97	0,91	0,88	1,00	0,88
5-2	0,99	0,93	0,99	1,00	0,98	0,97	0,91	0,88	1,00	0,88
5-3	0,92	0,94	1,00	1,00	0,98	0,97	0,88	0,89	1,00	0,88
5-4	0,91	0,85	1,00	1,00	0,89	0,99	0,88	0,90	1,00	0,88
5-5	0,92	0,96	1,00	1,00	0,98	0,97	0,88	0,88	1,00	0,88

Nesses exemplos, o MLP não teve grandes dificuldades em classificar corretamente as fases identificadas, nem em deixar de identificar uma fase do gesto dentro das existentes no conjunto de dados, precisão e revocação respectivamente, com uma ressalva para o resultado #5-3, que obteve apenas 0,49 de revocação para a fase de *hold*, e que foi o responsável por deixar o F1-Score desse resultado como 0,66 mesmo com uma precisão de 100%. Assim, analisando os melhores resultados do *stroke* e do descanso na tabela 10, não é possível afirmar que haja uma característica específica que esteja limitando um resultado de 100% para o F1-Score.

Analisando os melhores resultados gerais com menos de 18 aspectos da tabela 8, percebe-se que o melhor deles, resultado #6-1, não está distante do resultado #6-5, com os mesmos 0,26 de acurácia para o *hold* e F1-Score com diferença de apenas 0,01 para essa mesma fase. A fase de *hold* se distanciou da média, mas não foi a única. No resultado #6-4 a fase de preparação foi a que se distanciou das demais, com 0,57 de acurácia e 0,7 de F1-Score.

A tabela 11 lista os melhores resultados para modelos com apenas um ou dois aspectos do gesto na construção da representação e janelas de tamanho 1 para uma análise de quão simples uma representação do gesto pode ser e, ainda assim, apresentar bons resultados sob uma determinada ótica. Nessa tabela foram incluídos os melhores resultados de todas as variações possíveis em relação a filtro e normalização dos dados por ser mais fácil identificar a influência desses pré-processamentos sobre modelos que não utilizam a configuração de janela de dados.

Tabela 11 – Acurácias e F1-Scores dos melhores resultados identificados quanto a acurácia total para representações com uma ou duas características e janela de tamanho 1.

#	Vídeo	Corte	Ângulo	Filtro	Norm	Jal	AspG	Rede	TxA	NeOc	DivDados
8-1	a1-raw	maior	-	<i>wavelet</i>	sim	1	4x5	Ffnet: L-M	0,01	3	<i>holdout</i> - 403030
8-2	a1-raw	maior	-	<i>wavelet</i>	-	1	4x5	Ffnet: L-M	0,01	3	<i>k-fold</i> - $k = 10$
8-3	a1-va3	maior	-	<i>wavelet</i>	-	1	26x26	Ffnet: L-M	0,01	2	<i>k-fold</i> - $k = 10$
8-4	a2-raw	maior	-	-	-	1	4x5	Ffnet: L-M	0,01	3	<i>holdout</i> - 701515
8-5	a1-raw	maior	-	-	sim	1	4x5	Ffnet: L-M	0,01	3	<i>k-fold</i> - $k = 10$
8-6	a1-va3	maior	-	-	-	1	26x26	Ffnet: L-M	0,01	2	<i>k-fold</i> - $k = 10$
#	Acc.Total	Acc_D	Acc_P	Acc_S	Acc_H	Acc_R	FSc_D	FSc_P	FSc_S	FSc_H	FSc_R
8-1	0,71	0,97	0,00	0,85	0,03	0,00	0,84	0,00	0,75	0,05	0,00
8-2	0,70	0,98	0,00	0,80	0,44	0,00	0,81	0,00	0,77	0,47	0,00
8-3	0,70	0,90	0,00	0,91	0,00	0,00	0,83	0,00	0,76	0,00	0,00
8-4	0,70	0,97	0,09	0,90	0,00	0,00	0,81	0,16	0,79	0,00	0,00
8-5	0,68	0,97	0,01	0,79	0,00	0,00	0,83	0,01	0,71	0,00	0,00
8-6	0,61	0,74	0,00	0,84	0,00	0,00	0,69	0,00	0,68	0,00	0,00

Se for aceito algum nível de erro, todas essas configurações também são interessantes para a identificação de descansos e *strokes*. Apesar de praticamente anularem a identificação de todas as demais fases do gesto, para as duas fases principais que diferenciam o gesto do não-gesto as acurácias e F1-Scores se mantiveram acima dos 0,68. O uso de filtro e de normalização dos dados têm pouca influência sobre o resultado final da representação que utiliza apenas aspectos espaciais, mas o filtro melhora o resultado final da representação que utiliza os aspectos temporais. Ratificando uma constatação que pode ser feita com as análises sobre os melhores resultados gerais. Assim, confirma-se que com poucas características os resultados com o uso de aspectos espaciais são melhores do que com aspectos temporais.

A alteração da taxa de aprendizado entre as configurações do MLP não exerceu influência significativa sobre os resultados. Na tabela 12 os resultados #9-1 e #9-2 apresentam os dois melhores resultados com a variação dessa configuração. Há taxas altas de acurácia e de F1-Score para o reconhecimento de todas as fases do gesto, não se distanciando dos melhores resultados gerais obtidos com a taxa de aprendizado de 0,01.

O mesmo não pode ser dito para as variações sobre o tipo de rede utilizada para processar os modelos de representação, com os melhores resultados também listados na tabela 12. As demais redes apresentaram dificuldades para identificar certas fases, como a preparação, o *hold* e a retração. A configuração que utilizou a *feedforwardnet* com gradiente descendente apresentou dificuldade para reconhecer essas fases, e o resultado apenas piorou quando se adicionou a característica de taxa de aprendizado adaptativa e *momentum*, com praticamente apenas o descanso e o *stroke* sendo reconhecidos.

Tabela 12 – Acurácias e F1-Scores dos melhores resultados identificados quanto a acurácia total para as demais variações de configuração do MLP.

#	Vídeo	Corte	Ângulo	Filtro	Norm	Jal	AspG	Rede	TxA	NeOc	DivDados
Taxa de aprendizado											
9-1	a1-va3	maior	-	<i>wavelet</i>	-	20	1x24	Ffnet: L-M	0,1	409	<i>k-fold</i> - $k = 10$
9-2	a1-va3	maior	-	<i>wavelet</i>	-	20	1x24	Ffnet: L-M	0,001	409	<i>k-fold</i> - $k = 10$
Rede											
9-3	a1-va3	maior	-	<i>wavelet</i>	-	20	1x32	Ptnet	0,1	472	<i>k-fold</i> - $k = 10$
9-4	a1-raw	maior	-	<i>wavelet</i>	-	20	1x18	Ffnet: GD	0,01	355	<i>k-fold</i> - $k = 10$
9-5	a1-raw	maior	-	<i>wavelet</i>	-	20	1x18	Ffnet: GD Adp+Mom	0,01	10	<i>k-fold</i> - $k = 10$
9-6	a1-raw	maior	sim	<i>wavelet</i>	-	20	1x23	Ffnet: GD Adp	0,1	10	<i>k-fold</i> - $k = 10$
9-7	a1-raw	maior	sim	<i>wavelet</i>	-	20	1x21	Ffnet: GD Mom	0,01	10	<i>k-fold</i> - $k = 10$
Neurônios - Sendo: tamanho da janela, tamanho da janela + 2 e fixo em 10.											
9-8	a1-va3	maior	-	<i>wavelet</i>	-	20	1x18	Ffnet: L-M	0,1	10	<i>k-fold</i> - $k = 10$
9-9	a1-va3	maior	-	<i>wavelet</i>	-	20	1x18	Ffnet: L-M	0,01	3	<i>k-fold</i> - $k = 10$
9-10	a1-raw	-0,2	-	<i>wavelet</i>	-	20	1x18	Ffnet: L-M	0,001	1	<i>k-fold</i> - $k = 10$
Divisão											
9-11	a1-raw	maior	-	-	-	20	1x18	Ffnet: L-M	0,001	355	<i>k-fold</i> - $k = 3$
9-12	a1-va3	maior	-	<i>wavelet</i>	-	20	1x32	Ptnet	0,1	472	<i>holdout</i> - 303040
9-13	a1-raw	maior	sim	<i>wavelet</i>	sim	20	1x23	Ffnet: GD Adp	0,1	3	<i>holdout</i> - 151570
9-14	a1-va3	maior	-	<i>wavelet</i>	-	20	1x32	Ptnet	0,01	10	<i>holdout</i> - 202060
9-15	a1-raw	maior	sim	<i>wavelet</i>	sim	20	1x21	Ffnet: GD Adp	0,1	3	<i>holdout</i> - 304030
9-16	a1-raw	maior	-	<i>wavelet</i>	sim	20	1x18	Ffnet: GD Adp	0,1	10	<i>holdout</i> - 206020
9-17	a1-raw	maior	-	<i>wavelet</i>	-	20	1x18	Ffnet: GD Adp	0,01	10	<i>holdout</i> - 157015
#	Acc.Total	Acc.D	Acc.P	Acc.S	Acc.H	Acc.R	FSc.D	FSc.P	FSc.S	FSc.H	FSc.R
Taxa de aprendizado											
9-1	0,91	0,93	0,88	0,94	0,82	0,78	0,92	0,88	0,94	0,89	0,81
9-2	0,91	0,94	0,85	0,92	0,82	0,81	0,92	0,84	0,94	0,90	0,83
Rede											
9-3	0,85	0,96	0,36	0,91	0,67	0,69	0,94	0,45	0,87	0,67	0,75
9-4	0,84	0,94	0,34	0,97	0,41	0,52	0,90	0,37	0,93	0,53	0,67
9-5	0,80	0,98	0,04	0,97	0,31	0,30	0,90	0,07	0,86	0,44	0,44
9-6	0,79	0,98	0,00	0,99	0,26	0,22	0,90	0,00	0,85	0,40	0,35
9-7	0,77	0,98	0,00	0,98	0,00	0,10	0,88	0,00	0,85	0,00	0,17
Neurônios - Sendo: tamanho da janela, tamanho da janela + 2 e fixo em 10.											
9-8	0,89	0,96	0,56	0,93	0,56	0,81	0,95	0,67	0,91	0,68	0,80
9-9	0,88	0,96	0,50	0,95	0,00	0,82	0,96	0,57	0,91	0,00	0,77
9-10	0,77	1,00	0,00	1,00	0,00	0,00	0,88	0,00	0,88	0,00	0,00
Divisão											
9-11	0,84	0,95	0,50	0,88	0,77	0,60	0,90	0,56	0,89	0,66	0,64
9-12	0,78	0,94	0,41	0,86	0,15	0,35	0,88	0,46	0,81	0,25	0,48
9-13	0,76	0,98	0,00	0,95	0,00	0,07	0,86	0,00	0,84	0,00	0,13
9-14	0,73	0,83	0,25	0,85	0,18	0,50	0,84	0,34	0,77	0,25	0,50
9-15	0,75	0,96	0,00	0,98	0,00	0,00	0,88	0,00	0,81	0,00	0,00
9-16	0,74	0,95	0,00	0,96	0,00	0,00	0,85	0,00	0,82	0,00	0,00
9-17	0,73	0,98	0,00	0,90	0,00	0,00	0,84	0,00	0,81	0,00	0,00

O uso do método de cálculo de neurônios da camada oculta proposto por Fausett (1994) na equação 12, também foi o que obteve os melhores resultados. No geral, notou-se que quanto menor o número de neurônios, pior o resultado. A tabela 12 também

elencar os melhores resultados obtidos com as demais estratégias para obtenção do número de neurônios para a camada oculta. Nos resultados de #9-8 a #9-10, a acurácia e o F1-Score também foram melhores para as fases de descanso e *stroke*. Nesse recorte foi interessante observar que, quanto menor o número de neurônios, maior teve que ser a taxa de aprendizado para os resultados se manterem próximos em termos de acurácia e F1-Score.

Quanto ao método de divisão dos dados entre treinamento, validação e testes, o *k-fold* cíclico com $k = 10$ foi a melhor configuração. As demais estão listadas nos resultados de #9-11 até #9-17 da tabela 12. A configuração de *k-fold* com $k = 3$ também teve um bom resultado quando analisada sua capacidade em reconhecer todas as fases, com alguma dificuldade para as fases de preparação, *hold* e retração.

Os *holdouts* vêm logo em seguida, e o que se observou foi que os resultados vão piorando à medida que a quantidade de dados separados para treinamento é reduzida, e um conjunto de validação maior do que o conjunto de testes também reduziu a eficiência do MLP, mesmo que não tão significativamente.

O *k-fold* com $k = 10$ foi implementado tanto com o conjunto de testes fixo quanto com o conjunto de testes cíclico, conforme descrito na subseção 5.1.4. Todos os melhores resultados foram obtidos com o conjunto de testes rodando junto com os demais conjuntos de forma cíclica e posicionado entre os *folds* 3, 5 e 7. Essas posições para os *folds* de testes foram as que figuraram entre as melhores porque tomou-se o cuidado de considerar apenas as execuções que possuíam todas as fases do gesto representadas em todos os conjuntos de dados. O conjunto de testes em posições diferentes das citadas não possuíam todas as fases do gesto representadas e as execuções considerando esses conjuntos foram desprezadas. Por esse motivo as execuções com o último *k-fold* sendo usado de forma fixa para os testes também foram desprezadas.

As representações do gesto apresentadas têm potencial para reconhecimento das fases do gesto com o uso do MLP, tanto para o reconhecimento de todas as fases quanto para identificar situações de gesto e não-gesto focadas no descanso e *stroke*. Para a primeira situação, o uso de mais de 18 aspectos do gesto para construção da representação, seja com aspectos espaciais ou temporais, devem ser priorizados, e o tamanho da janela deve ser configurado de acordo com as necessidades de acurácia e F1-Score, salientando o fato de que, quanto maior a janela, melhores os resultados, mas aumentam o consumo de recursos computacionais. Para essa aplicação, o MLP deve ser configurado preferencialmente como

uma *feedforwardnet* com Levenberg-Marquardt e o número de neurônios sendo definido pela equação 12. Essa foi a configuração com maior influência positiva sobre os resultados.

Se o foco for apenas o reconhecimento de fases específicas, como o descanso e o *stroke*, a modelagem dos gestos e a configuração do MLP pode ser simplificada. O uso de poucas características para modelagem da representação, tamanho reduzido de janela de dados, bem como o uso de poucos neurônios na camada escondida, já pode ser considerado.

5.3 Experimentos com SOM

O SOM foi avaliado quanto ao erro de quantização e erro topológico obtido sobre cada execução realizada sobre os modelos de representação dos gestos. Esses erros foram analisados isoladamente, e usados como critério de priorização para então selecionar as imagens de matriz-U para serem verificadas.

O conjunto de arquivos resultante de cada execução contém o erro de quantização obtido para o modelo, o erro topológico obtido para o modelo, a matriz-U gerada, a matriz-U com a identificação do rótulo⁵ referentes às classes existentes no conjunto de dados e associado aos dados representados pelos neurônios do mapa, e a matriz-U com a identificação de volume de dados associados aos neurônios do mapa. Com esses dados foi possível elaborar as análises pretendidas para um algoritmo de aprendizado não supervisionado, verificando quais grupos foram formados e quais as fases do gesto o especialista havia atribuído para os dados desses grupos a fim de entender as diferenças e semelhanças encontradas.

As próximas subseções estão organizadas por aspectos do gesto e apresentam os resultados obtidos, começando com uma subseção sobre os resultados obtidos com as características do gesto relacionadas à aspectos temporais, por terem obtido os melhores resultados de agrupamento quanto aos erros de quantização e topológicos. Essa subseção apresentará os resultados para diferentes características de aspectos temporais, seguindo com análises sobre os resultados das representações que utilizaram apenas a velocidade escalar da mão direita, por ter alcançado os melhores resultados, verificando outras características do gesto e a configuração utilizada pelo SOM. A subseção seguinte analisa

⁵ No SOM os rótulos dos dados foram utilizados apenas para fins de análise e não para treinamento da rede neural.

os resultados obtidos com representações construídas a partir dos aspectos espaciais com e sem a adição da característica de angulação.

5.3.1 Aspectos temporais

Com a execução dos testes do SOM foram obtidos 40.992 resultados, que foram tabulados e ordenados crescentemente pelo erro de quantização em uma das tabulações e pelo erro topológico em outra das tabulações. Para iniciar as análises, foram separados os melhores resultados gerais que, conforme exposto na tabela 13, foram obtidos com as representações utilizando apenas a velocidade escalar da mão direita. As demais características também serão analisadas em seguida.

Tabela 13 – Melhores resultados identificados quanto ao erro de quantização por representação temporal.

Vídeo	Ângulo	Filtro	Norm	Jal	AspG	TipoSOM	TamMapa	Épocas	TxA	Qt_Erro	Tp_Erro
Apenas a velocidade escalar da mão direita											
b1-va3	-	<i>wavelet</i>	-	1	26x26	lote	1x169	<i>default</i>	<i>default</i>	3,73E-05	0,00
a3-va3	-	-	-	1	26x26	seq	20x20	100	0,5	4,16E-05	0,73
Apenas a velocidade escalar e aceleração escalar											
b1-va3	-	<i>wavelet</i>	-	1	25x32	lote	20x20	<i>default</i>	<i>default</i>	8,35E-04	0,07
b1-va3	-	<i>wavelet</i>	-	1	25x32	lote	20x20	100	<i>default</i>	8,35E-04	0,07
Apenas a velocidade vetorial e aceleração vetorial											
a2-va3	-	<i>wavelet</i>	-	1	1x24	lote	20x20	<i>default</i>	<i>default</i>	2,03E-03	0,05
a2-va3	-	<i>wavelet</i>	-	1	1x24	lote	20x20	100	<i>default</i>	2,03E-03	0,05
Todos os dados de velocidade e aceleração											
b1-va3	-	<i>wavelet</i>	-	1	1x32	lote	20x20	<i>default</i>	<i>default</i>	3,47E-03	0,05
b1-va3	-	<i>wavelet</i>	-	1	1x32	lote	20x20	100	<i>default</i>	3,47E-03	0,05

A tabela 14 lista os dez melhores resultados em termos de erro de quantização e os dez melhores resultados em termos de erro topológico obtidos com o uso de janela de tamanho 1 e apenas a velocidade escalar da mão direita.

Analisando apenas os erros, a primeira conclusão é que quanto menos características do gesto são utilizadas, menores são os erros produzidos, sem notável influência do uso ou não de filtro, normalização, tipo do SOM, tamanho do mapa, épocas e taxas de aprendizado. Nesse sentido, o modelo com menos aspectos produzidos possui apenas a velocidade escalar da mão direita. O resultado #1-1 foi o melhor obtido nos dois critérios, sendo o mesmo apresentado no resultado #1-11. Por ter sido um resultado gerado sobre um mapa de uma dimensão matricial e com muitos neurônios, sua análise visual não trouxe muitas informações. Assim, as análises foram feitas sobre os BMUs resultantes e consolidado na tabela 15.

Tabela 14 – Melhores resultados identificados quanto aos erros de quantização e topológicos isoladamente.

#	Vídeo	Ângulo	Filtro	Norm	TipoSOM	TamMapa	Épocas	TxA	Qt_Erro	Tp_Erro
Menor erro de quantização										
1-1	b1-va3	-	<i>wavelet</i>	-	lote	1x169	<i>default</i>	<i>default</i>	3,73E-05	0,00
1-2	a3-va3	-	-	-	seq	20x20	100	0,5	4,16E-05	0,73
1-3	b1-va3	-	-	-	seq	20x20	<i>default</i>	0,1	4,19E-05	0,74
1-4	a3-va3	-	-	-	seq	20x20	100	0,1	4,23E-05	0,71
1-5	b1-va3	-	-	-	seq	20x20	<i>default</i>	<i>default</i>	4,27E-05	0,77
1-6	a3-va3	-	-	-	seq	20x20	<i>default</i>	1	4,29E-05	0,74
1-7	b1-va3	-	-	-	seq	20x20	100	1	4,32E-05	0,75
1-8	b1-va3	-	<i>wavelet</i>	-	lote	20x20	<i>default</i>	<i>default</i>	4,35E-05	0,89
1-9	b1-va3	-	<i>wavelet</i>	-	lote	20x20	100	<i>default</i>	4,35E-05	0,89
1-10	b1-va3	-	<i>wavelet</i>	-	lote	20x20	1000	<i>default</i>	4,35E-05	0,89
Menor erro topológico										
1-11	b1-va3	-	<i>wavelet</i>	-	lote	1x169	<i>default</i>	<i>default</i>	3,73E-05	0
1-12	a3-va3	-	<i>wavelet</i>	-	lote	1x210	<i>default</i>	<i>default</i>	4,77E-05	0
1-13	b1-va3	-	-	-	lote	1x169	<i>default</i>	<i>default</i>	5,32E-05	0
1-14	c1-va3	-	<i>wavelet</i>	-	lote	1x169	<i>default</i>	<i>default</i>	5,44E-05	0
1-15	a3-va3	-	-	-	lote	1x210	<i>default</i>	<i>default</i>	5,53E-05	0
1-16	a2-va3	-	<i>wavelet</i>	-	lote	1x182	<i>default</i>	<i>default</i>	6,30E-05	0
1-17	a1-va3	-	<i>wavelet</i>	-	lote	1x210	<i>default</i>	<i>default</i>	6,46E-05	0
1-18	c1-va3	-	-	-	lote	1x169	<i>default</i>	<i>default</i>	6,78E-05	0
1-19	c3-va3	-	<i>wavelet</i>	-	lote	1x196	<i>default</i>	<i>default</i>	6,81E-05	0
1-20	c3-va3	-	-	-	lote	1x196	<i>default</i>	<i>default</i>	6,99E-05	0

Tabela 15 – Matriz de confusão entre as fases de cada dado definida pelo especialista e a fase mais representativa associada aos BMUs para os resultados #1-1 e #1-11.

Fase do dado	Fase associada ao BMU				
	D	P	S	H	R
D	36	14	2	18	0
P	2	313	53	39	5
S	4	105	150	23	5
H	22	85	17	90	3
R	0	43	20	6	14

A associação não foi direta entre a realizada pelo especialista e o resultado do BMU. Dados de diferentes fases foram associadas à um mesmo BMU e a fase com o maior número de representantes associados ao BMU foi escolhida como a fase do BMU.

O resultado #1-2 foi obtido com um mapa de 20x20, o que já possibilita uma análise visual. A figura 17 apresenta a matriz-U obtida por esse resultado sobre a qual duas visualizações são mostradas: a da esquerda com a adição do *label* da fase do gesto representado pelas diferentes regiões do mapa, e a da direita com a identificação do volume de dados em cada região do mapa destacado em preto.

Figura 17 – Matriz-U do resultado #1-2.

Fonte: Ricardo Alves Feitosa, 2018

Na figura 17, as áreas com o tom mais azulado são as que possuem a maior concentração de neurônios e as áreas com o tom mais amarelado são as que possuem menor concentração de neurônios. No canto superior esquerdo nota-se uma grande concentração de dados definidos pelo especialista como pertencentes à fase de descanso, enquanto no canto inferior esquerdo está uma concentração de dados das fases de *stroke*, preparação e retração. Indo mais para a direita, há uma mistura das fases e com poucos dados, sem uma clara formação de agrupamentos.

Assim, é possível concluir que o SOM agrupou os dados classificados como descanso em um mesmo grupo, mas não conseguiu separar de forma clara os dados das demais fases do gesto, salientando que nessa representação o SOM foi aplicado apenas sobre uma característica do gesto, a velocidade escalar da mão direita, e sem o uso de filtro. Esse foi o mesmo efeito produzido pelos resultados de #1-3 a #1-7.

Com a aplicação do filtro *wavelet* o resultado mudou. A figura 18 apresenta a matriz-U do resultado #1-8, sobre a qual é possível identificar alguns agrupamentos destacados em vermelho.

Figura 18 – Matriz-U do resultado #1-8.

Fonte: Ricardo Alves Feitosa, 2018

O grupo identificado como *A* possui descansos, *holds* e preparações em sua maioria, com poucos casos de *strokes* e retrações. Esse grupo é um bom candidato a representar fases do gesto com características de pouca movimentação. O grupo *B*, por sua vez, se concentra nas fases de preparação e alguns *strokes*, mostrando-se um grupo com boa capacidade para identificar as fases de transição para o *stroke*. É interessante observar também o grupo *C*, também concentrado na fase de preparação, mas com alguns exemplares de *hold*, podendo ser visto como um grupo de transição para o *hold*. Por fim, há o grupo *D*, com a maioria dos dados concentrados na fase de *stroke*, assim com potencial para identificação dessa característica.

O uso de filtro para os aspectos temporais resultou em grupos mais bem definidos. Na tabela 16 estão elencados os resultados com melhor erro de quantização para mais de uma característica dos aspectos temporais do gesto. Com a adição de mais características para o gesto, o mapa de 20 x 20 também foi o que apresentou os melhores resultados.

Tabela 16 – Melhores resultados identificados quanto aos erros de quantização para mais de um aspecto temporal.

#	Vídeo	Ângulo	Filtro	Norm	AspG	Épocas	TxA	Qt_Erro	Tp_Erro
2-1	b1-va3	-	wavelet	-	25x32	default	default	8,35E-04	6,74E-02
2-2	b1-va3	-	wavelet	-	25x32	100	default	8,35E-04	6,74E-02
2-3	a3-va3	-	-	-	25x32	default	default	1,81E-03	8,31E-02
2-4	a2-va3	-	wavelet	-	1x24	default	default	2,03E-03	4,60E-02
2-5	a2-va3	-	wavelet	-	1x24	100	default	2,03E-03	4,60E-02
2-6	b1-va3	-	wavelet	-	1x32	default	default	3,47E-03	4,77E-02
2-7	b1-va3	-	wavelet	-	1x32	100	default	3,47E-03	4,77E-02
2-8	a3-va3	-	-	-	1x24	default	default	4,27E-03	4,54E-02
2-9	a3-va3	-	-	-	1x32	default	default	5,56E-03	5,03E-02

A figura 19 apresenta a matriz-U do resultado #2-1, onde é possível separar grupos como o *A* para identificar dados semelhantes ao *hold* e o grupo *B*, que tem uma boa capacidade para identificar dados relacionados à fase de preparação e *stroke*. Os grupos *C* e *D* se mostraram muito bons para identificar *strokes*, principalmente o grupo *C* com picos bem definidos sobre as demais regiões do mapa. Dois grupos com alta capacidade para identificar *strokes* podem significar que os *strokes* do grupo *C*, são diferentes dos do grupo *D*, uma vez que os dados desse último estão em um mesmo grupo de dados relacionados ao *hold* e à preparação. Pode-se dizer até que o grupo *C* tem a capacidade de identificar os dados com picos de gesto, enquanto o *D* identifica dados de *frames* relativos à uma fase de transição com características mais específicas e voltadas ao início do *stroke* do que o grupo *B*, por exemplo.

Figura 19 – Matriz-U do resultado #2-1.

Fonte: Ricardo Alves Feitosa, 2018

Quando a informação de filtro é retirada dessa configuração, tem-se o resultado #2-3, cuja matriz-U está na figura 20. Vê-se um comportamento semelhante ao apresentado anteriormente, com uma grande concentração de dados voltados para a fase de *hold*, destacados pelo grupo *A*, que vão se distânciando visualmente entre as demais fases do gesto. A adição de mais características do gesto relacionados à velocidade escalar definiram um agrupamento de *strokes* caracterizado por neurônios distantes entre si, identificados como *B*.

Figura 20 – Matriz-U do resultado #2-3.

Fonte: Ricardo Alves Feitosa, 2018

Porém, a adição de aspectos deve ser feita com cuidado: quando foi utilizada apenas a informação de velocidade vetorial, apresentada na figura 21 referente ao resultado #2-4, foi notado um comportamento semelhante à não adição de filtro, com um grande grupo composto por dados da fase de descanso, e grupos contendo dados mais distantes visualmente entre si com dados de toda as fases do gesto.

Figura 21 – Matriz-U do resultado #2-4.

Fonte: Ricardo Alves Feitosa, 2018

A característica de velocidade escalar, portanto, foi a que trouxe melhores informações para tornar a representação discriminante. Com base no que foi visto, quando se combinou a informação de velocidade escalar com a velocidade vetorial, os resultados apresentaram ligeira melhora para análise visual. A figura 22 apresenta o resultado #2-6 com grupos levemente mais bem definidos do que o resultado #2-4. São visíveis concen-

trações de dados como no grupo *A*, com *holds* e descansos, e nos grupos *B* e *C*, voltados para identificação de fases de transição e *strokes*, respectivamente.

Figura 22 – Matriz-U do resultado #2-6.

Fonte: Ricardo Alves Feitosa, 2018

Isto posto, os melhores resultados foram obtidos considerando apenas a informação de velocidade escalar com a aplicação de filtro. A variação no tipo do algoritmo SOM, épocas e taxa de aprendizado não exerceram influência significativa sobre os resultados, mantendo-os iguais aos apresentados para cada caso.

A tabela 17 lista os melhores resultados quanto ao erro de quantização para variações nas configurações de tamanho de janela e tamanho do mapa. Todos os resultados foram obtidos com o uso do filtro *wavelet*.

Tabela 17 – Melhores resultados identificados quanto aos erros de quantização para tamanho de mapa diferente de 20x20 e resultados com janelas de 20 e 40 mesmo em mapas 20x20.

#	Vídeo	Jal	TipoSOM	TamMapa	Épocas	TxA	Qt_Erro	Tp_Erro
3-1	b1-va3	1	seq	10x10	default	0,1	1,42E-04	6,98E-01
3-2	a3-va3	1	lote	21x10 (default)	default	default	1,17E-03	3,33E-02
3-3	b1-va3	1	lote	1x10	default	default	6,21E-04	0
3-4	b1-va3	1	seq	1x5	1000	1	1,20E-03	0
3-5	b1-va3	20	lote	20x20	default	default	3,31E-03	6,19E-02
3-6	b1-va3	40	lote	20x20	default	default	6,99E-03	3,59E-02

O uso dos mapas 10x10 está plotado na figura 23. Para o resultado #3-1 há os grupos *A* e *B* com concentração de muitos dados relacionados à *holds* e preparação, respectivamente. Os agrupamentos formados pelo resultado #3-2 estão concentrados nos grupos *C* e *D* para descanso e *stroke*. Quanto maior o mapa, maior a capacidade da rede em encontrar os grupos.

Figura 23 – Matriz-U dos resultados #3-1 e #3-2.

Fonte: Ricardo Alves Feitosa, 2018

O uso de mapas de tamanho 1x5 e 1x10 têm o objetivo de verificar o quão próximo da classificação do especialista o SOM consegue chegar. Como são mapas de uma dimensão matricial, a análise foi realizada sobre as matrizes de confusão entre os dados e seus BMUs. A tabela 18 lista esses resultados.

Tabela 18 – Matriz de confusão entre as fases de cada dado definida pelo especialista e a fase mais representativa associada aos BMUs para os resultados #3-3 e #3-4.

Resultado #3-3	Fase associada ao BMU					Resultado #3-4	Fase associada ao BMU				
Fase do dado	D	P	S	H	R	Fase do dado	D	P	S	H	R
D	0	15	0	55	0	D	0	10	0	60	0
P	0	327	37	48	0	P	0	310	0	102	0
S	0	188	75	24	0	S	0	208	23	56	0
H	0	103	2	112	0	H	0	84	0	133	0
R	0	58	15	10	0	R	0	72	1	10	0

Os agrupamentos formados pelo SOM não foram iguais às classificações geradas pelo especialista. Em ambas as execuções, mais BMUs foram associados às fases de preparação e *hold* do que à fase de *stroke*, e as fases de descanso e retração não tiveram BMUs associados. Os dados classificados como essas fases foram associados à BMUs das demais fases. É

interessante observar que os dados classificados como descanso sempre foram associados à BMUs de grupos com dados predominantes das fases de preparação e *hold*, e que a retração se dividindo entre preparação, *stroke* e *hold*. As demais execuções envolvendo mapas de tamanho 1x10 e 1x5 seguiram os mesmos desfechos em relação a seus resultados.

Quanto às variações nos tamanhos da janela, aumentar seu valor, apesar de aumentar o erro de quantização e topológico, aumentou a capacidade de segregar os grupos no mapa quando comparado o resultado #1-2 com o #3-5 e #3-6. As janelas de tamanho 20 deixaram em evidência os grupos *A*, *B* e *C* da figura 24: o grupo *A*, formado por dados de *hold*, preparação e descanso como em resultados anteriores, os grupos *B* e *C*, bem definidos para as fases de preparação e *stroke*, respectivamente.

Figura 24 – Matriz-U dos resultados #3-5 e #3-6.

Fonte: Ricardo Alves Feitosa, 2018

A execução com janela de tamanho 40 evidenciou os grupos *D*, *E*, *F* e *G*, um número maior de grupos em comparação à execução utilizando janela de tamanho 20, com o grupo *D* formado por dados da fase de *hold*, os grupos *F* e *G* bem definidos para

identificação de preparação e *stroke*, nessa ordem, e o grupo *E* com uma mistura de dados para a fase de preparação e *stroke*.

5.3.2 Aspectos espaciais

Os melhores resultados obtidos quanto ao erro de quantização e erro topológico foram orientados a aspectos temporais, mas bons resultados também foram obtidos com o uso de aspectos espaciais. Nessa subseção serão listados os resultados obtidos tanto com o uso apenas das coordenadas espaciais, quanto com sua combinação com a informação de angulação. A tabela 19 elenca os melhores resultados para diferentes combinações de características para representar aspectos espaciais dos gestos, todas obtidas com correção de erro em lote, mapa de 20x20 e cálculo *default* para o número e épocas e para a taxa de aprendizado.

Tabela 19 – Melhores resultados identificados para cada caso de teste quanto aos erros de quantização para modelos formados com aspectos espaciais.

#	Vídeo	Ângulo	Filtro	Norm	Jal	AspG	Qt_Erro	Tp_Erro
4-1	b3-raw	-	-	-	1	4x5	0,01	0,04
4-2	b3-raw	-	<i>wavelet</i>	-	1	4x5	0,01	0,08
4-3	b3-raw	sim	<i>wavelet</i>	-	1	19x22	0,07	0,06
4-4	b3-raw	sim	<i>wavelet</i>	sim	1	19x22	0,07	0,06
4-5	b3-raw	-	<i>wavelet</i>	-	20	4x5	0,13	0,04
4-6	b3-raw	sim	<i>wavelet</i>	-	1	1x22	0,13	0,04
4-7	b3-raw	-	<i>wavelet</i>	-	40	4x5	0,25	0,06
4-8	b3-raw	-	<i>wavelet</i>	-	20	1x18	0,46	0,05
4-9	b3-raw	sim	<i>wavelet</i>	-	20	19x22	0,71	0,03
4-10	a1-raw	-	<i>wavelet</i>	-	20	4x6	0,83	0,04
4-11	b3-raw	sim	<i>wavelet</i>	-	20	1x20	0,94	0,04
4-12	a1-raw	-	<i>wavelet</i>	-	20	1x18	1,76	0,03
4-13	b1-raw	-	-	sim	1	1x18	2,06	0,12
4-14	a2-raw	sim	<i>wavelet</i>	-	20	1x20	4,10	0,03
4-15	b1-raw	-	<i>wavelet</i>	sim	20	4x5	54,11	0,04
4-16	b1-raw	sim	<i>wavelet</i>	sim	40	1x22	257,24	0,03

O resultado da adição de filtro sobre as características dos aspectos espaciais não foi tão significativo quanto para as características dos aspectos temporais, como pode ser visto na figura 25 com a plotagem da matriz-U dos resultados #4-1 e #4-2, com e sem filtro, respectivamente.

Figura 25 – Matriz-U dos resultados #4-1 e #4-2.

Fonte: Ricardo Alves Feitosa, 2018

Sem a utilização do filtro, foram identificados os grupos *A*, *B*, *C* e *D*, para identificação de *stroke*, *hold*, preparação e descanso, nessa ordem. Já com a utilização de filtro, foram identificados os grupos *F* e *I* bem definidos para identificação de preparação e descanso, e os grupos *E*, *G* e *H*, com sobreposições de dados das fases de *stroke*, *hold* e preparação, respectivamente. Os resultados #4-2, #4-5 e #4-7 obtiveram a mesma configuração de agrupamento, mostrando que a adição da informação de janela também não tem efeitos significativos sobre o resultado final do agrupamento.

Os resultados #4-1, #4-2, #4-5 e #4-7 foram obtidos usando apenas as coordenadas *x* e *y* da mão direita. Quando a coordenada *z* foi adicionada, o resultado #4-10 foi um dos melhores obtidos, mesmo se tratando do vídeo a1, e sua matriz-U resultante está plotada na figura 26.

Figura 26 – Matriz-U do resultado #4-10.

Fonte: Ricardo Alves Feitosa, 2018

Os grupos *A* e *B* representam bem a fase de descanso, sendo que esses dois grupos provavelmente são o mesmo grupo formado por uma dobra no mapa já que se encontram posicionados nos cantos superiores. Os grupos *C*, *D* e *E* identificando a fase de *stroke*, também podem ser um único grupo formado por dobras no mapa. A adição da coordenada z aumentou a capacidade do SOM de manter concentrados os BMUs associados à fase de *stroke* e descanso, mostrando-se uma importante característica para esse fim.

Quando todas as coordenadas espaciais foram consideradas, os melhores resultados obtidos foram os identificados como #4-8 e #4-12, que têm suas matrizes-U plotadas na figura 27.

Muitos grupos foram visualizados, e notou-se que estão bem distantes entre si quando comparados com execuções com menos características do gesto. É provável que muitas dobras tenham ocorrido no mapa. Os grupos *A* e *B* provavelmente são o mesmo grupo, identificando dados das fases *hold*, *stroke* e preparação, assim como os grupos *G* e *H* para as fases de descanso e preparação. O grupo *D* parece bem definido para reconhecer *strokes*, e os grupos *C*, *E* e *F* reconheceram semelhanças entre dados das fases de preparação, *stroke* e *hold*, podendo, também, ser o mesmo grupo do formado por *A* e *B* em decorrência de uma dobra no mapa.

Figura 27 – Matriz-U dos resultados #4-8 e #4-12.

Fonte: Ricardo Alves Feitosa, 2018

Os grupos *I*, *J*, *N* e *O* também agruparam concentrações de dados que representaram o *stroke*, podendo, inclusive, ser o mesmo grupo formado por uma dobra no mapa, assim como os grupos *J* e *K*, mas voltados para a identificação do descanso. O grupo *M* foi um dos poucos dentre todas as execuções que conseguiu reunir uma grande quantidade de dados representantes da retração. Assim, utilizar todas as coordenadas espaciais parece uma estratégia interessante para se tentar recuperar a informação de retração, mas não é tão boa para as demais fases como o obtido por configurações mais simples como a utilização de aspectos temporais de aceleração e velocidades escalares ou o uso de aspectos espaciais com apenas as coordenadas da mão direita.

Quando os dados foram normalizados, houve um aumento no erro de quantização e topológico e o resultado visual também ficou com os agrupamentos menos discriminantes como pode ser visto na figura 28. Há uma grande concentração dos dados sem que seja possível separar os grupos. Esse comportamento ocorreu mesmo com a adição do filtro *wavelet* que, conforme já mencionado, exerceu pouca influência sobre os resultados obtidos com modelos gerados a partir das coordenadas espaciais.

Figura 28 – Matriz-U do resultado #4-13.

Fonte: Ricardo Alves Feitosa, 2018

A tabela 19 também elencou os melhores resultados para diferentes combinações de aspectos espaciais dos gestos utilizando a informação de ângulo. Um dos melhores resultados está identificado como #4-3 e representado pela matriz-U da figura 29. Quando apenas a informação de ângulo foi utilizada, foi possível identificar agrupamentos de fases do gesto como os identificados por *A*, *B*, *C* e *D*, nos quais há a representatividade das fases de descanso, *hold*, preparação e *stroke*, mas sempre misturadas com informações previamente classificadas como pertencentes a outra fase.

Figura 29 – Matriz-U do resultado #4-3.

Fonte: Ricardo Alves Feitosa, 2018

Por outro lado a informação de angulação melhorou os resultados de forma mais significativa quando combinada com outros aspectos, como os demais aspectos espaciais, como pode ser visto na figura 30 relativa aos resultados #4-11 e #4-6. Os BMUs ficaram

dispersos, semelhante ao que ocorreu com o resultado #4-8 no qual as mesmas coordenadas espaciais foram utilizadas sem a angulação.

No resultado com angulação formada por todos os ângulos menos o formado entre o pulso e o ponto central da mão, os grupos se mostraram mais consistentes, com mais concentrações de dados de uma mesma fase do gesto, como nos agrupamentos identificados como *B* e *C*, formados quase que exclusivamente por preparações e *holds*. O grupo *A* também possui uma boa representatividade da fase de descanso, misturando-se em alguns momentos com dados classificados pelo especialista como pertencentes às fases de preparação e retração, o que pode ser considerado como aceitável quando essas fases são visualizadas como gestos com pouca variação em relação as coordenadas espaciais e abertura de ângulos, ponderando sobre inícios de preparações e finais de retrações. O mesmo pode ser dito dos grupos *D* e *E*, nos quais *strokes* são colocados no mesmo grupo de preparações e retrações, fases também semelhantes ao observar o final das preparações e início das retrações.

Figura 30 – Matriz-U dos resultados #4-11 e #4-6.

Fonte: Ricardo Alves Feitosa, 2018

Quando a informação da angulação formada entre o pulso e o ponto central da mão foi adicionada à representação, o resultado foi semelhante, mas com grupos não tão bem definidos como os do resultado sem essa informação. Esse cenário pode ser visto no resultado #4-6 que também está plotado na figura 30, no qual os grupos H e I estão bem relacionados às fases de preparação e descanso, e os grupos A e B com o *stroke* misturado a dados pertencentes às fases de preparação e *hold*, de forma que os grupos F e G são provavelmente o mesmo grupo formado por uma dobra no mapa.

O uso do filtro relacionado à angulação também não gerou bons resultados. O próprio erro de quantização e baixo erro topológico deixam claro o quão complexa é a organização dos dados em relação ao mapa. Esse cenário pode ser visualizado no resultado #4-16, com a matriz-U da figura 31.

Figura 31 – Matriz-U do resultado #4-16.

Fonte: Ricardo Alves Feitosa, 2018

Esse resultado segue um comportamento semelhante ao apresentado pelo resultado #4-13, com uma grande concentração do BMUs e a dificuldade em identificar agrupamentos formados por essa configuração.

6 Conclusões

A representação do gesto deve ser adequada o suficiente para que a segmentação das suas fases de forma automática possa ser executada. Como representação adequada, se enquadra aquela discriminante o suficiente para que os algoritmos de aprendizado de máquina consigam executar a tarefa de segmentação e, conseqüentemente, descobrir conhecimento sobre os padrões de gestos.

Neste trabalho foram apresentadas abordagens para a construção das representações, considerando aspectos espaciais e temporais. Sobre esses aspectos foram feitas combinações de características descritivas dos gestos, construindo as representações multidimensionais sobre as quais os algoritmos de reconhecimento de padrões, MLP e SOM, foram executados. Com o aprendizado supervisionado realizado pelo MLP, os resultados foram avaliados de forma a identificar o quão assertivas foram as classificações obtidas pelos modelos aprendidos em relação às classificações que eram esperadas de acordo com as rotulações dos especialistas, concluindo sobre quais as representações do gesto foram as mais discriminantes na obtenção dos melhores resultados. Com o aprendizado não supervisionado, a expectativa também foi identificar as fases do gesto propostas por Kendon (1980), McNeill (1992) e McNeill (2015), porém sem o viés do especialista, bem como verificar se os agrupamentos gerados poderiam sugerir fases do gesto diferentes das propostas.

Com os experimentos realizados sobre as diferentes características usadas para descrever os aspectos do gesto é possível concluir que as representações dos gestos que produzem bons resultados para o MLP são diferentes das que produzem bons resultado com o SOM.

Para o MLP o uso de características dos aspectos temporais conseguiu resultados melhores do que os espaciais, por mais que a diferença tenha sido pequena. As representações mais discriminantes utilizaram todos os dados de velocidade e aceleração, seguidos pelos modelos que utilizaram apenas a velocidade vetorial e aceleração vetorial, mostrando capacidade para identificar todas as fases do gesto com altos níveis de acurácia e F1-Score. Os modelos de representação formados apenas pela velocidade escalar e aceleração escalar e os formados apenas pela velocidade escalar da mão direita apresentaram bons resultados para identificação das fases de descanso e *stroke*.

As características de aspectos espaciais que formaram as representações mais discriminantes foram aqueles que utilizaram todas as coordenadas espaciais, as coordenadas espaciais junto com todos ângulos e as coordenadas espaciais junto com o ângulo formado entre o ponto central da mão e o tronco do corpo. Essas foram as configurações que obtiveram altos valores de acurácia e F1-Score para todas as fases do gesto. Para identificação apenas do descanso e do *stroke*, as representações formadas pelo x e y da mão direita, por todas as coordenadas da mão direita e apenas pelos ângulos apresentaram bons resultados. A informação de ângulo adicionada à representação melhora a capacidade discriminativa da representação construída sobre os aspectos espaciais.

O uso do filtro *wavelet* e o maior tamanho de janela de dados influenciaram positivamente nos resultados que utilizaram características dos aspectos temporais no MLP. Com o tamanho da janela melhorando os resultados também para as representações que utilizaram características dos aspectos espaciais. Porém, cabe salientar o elevado custo computacional consequente ao aumento do tamanho das janelas de dados. Já a aplicação da normalização sobre os dados não exerceu influência significativa sobre os resultados.

Para o SOM, as representações mais discriminantes também foram as obtidas com características dos aspectos temporais. Porém, quanto mais aspectos do gesto foram adicionados à representação, piores foram os resultados quanto aos erros de quantização e topológicos, mas não necessariamente produziram piores resultados para a análise visual. As melhores representações que permitiram a descoberta de conhecimento sobre padrões foram as que utilizaram apenas a velocidade escalar da mão direita e a velocidade escalar e aceleração escalar de todos os membros. As representações formadas pela velocidade vetorial e aceleração vetorial e por todos os dados de velocidade e aceleração foram suficientes apenas para identificação de fases isoladas, com exceção da fase de retração, que não foi bem reconhecida com o uso dessas representações.

Representações do gesto que utilizaram as características dos aspectos espaciais envolvendo apenas x e y da mão direita e todas as coordenadas da mão direita também alcançaram bons resultados, com capacidade para formar grupos para todas as fases do gesto. A adição da coordenada z aumentou a capacidade do SOM de formar grupos mais bem definidos com dados das fases de *stroke* e descanso. Utilizar todas as coordenadas espaciais só foi capaz de reconhecer fases isoladas e pouco definidas, porém foi a única representação que conseguiu delinear um grupo capaz de mapear os dados referentes à retração.

Utilizar todos os ângulos isoladamente também obteve bons resultados para identificar as fases do gesto, mas com grupos pouco definidos se comparados com as representações anteriores. Combinar essa informação com as demais características de aspectos espaciais melhorou a capacidade do SOM em mapear as fases do gesto.

A aplicação do filtro *wavelet* e o tamanho da janela também influenciaram positivamente os resultados, porém em menor grau sobre os aspectos espaciais do que sobre os aspectos temporais. Já a normalização dos dados prejudicou a capacidade do SOM em gerar grupos bem definidos.

O SOM e as representações das fases do gesto podem, portanto, serem aplicadas com o objetivo de segmentar as fases do gesto e obter bons resultados sem o viés introduzido na análise supervisionada. Sobre a capacidade do SOM em gerar conhecimento que permita a proposta de fases do gesto diferentes da apresentada por Kendon (1980), McNeill (1992) e McNeill (2015), os testes não levam a entender que uma nova fase poderia ser criada. Nas análises sobre os mapas de tamanho 1x5 e 1x10, e até sobre os demais, os grupos formados direcionam que os dados classificados previamente por especialistas, como *stroke*, preparação e retração, poderiam ser colocados em um mesmo grupo, assim como as fases de transição entre *strokes* e *holds* com sua fase adjacente.

Como os testes com as diferentes representações exigiram certo grau de configuração das ferramentas MLP e SOM, também é possível fazer afirmações sobre quais configurações impactam significativamente os resultados. Para o MLP, os melhores resultados foram os obtidos utilizando a rede configurada como *feedforward* com Levenberg-Marquardt e o número de neurônios sendo definido pela equação 12. A divisão dos dados entre treinamento, validação e testes deve ser realizado garantindo que haja representantes de todas as fases que se pretende classificar em cada um desses conjuntos de dados. Para o SOM, o tamanho do mapa em 20x20, o número de épocas entre 100 e o definido como *default* na subseção 5.1.4 e a taxa de aprendizado de 0,5 para ordenação e 0,05 para convergência conseguem mapear grupos mais bem definidos.

6.1 Trabalhos publicados

Três publicações foram produzidas com conteúdos parciais derivados desse projeto, sendo que os dois primeiros foram apresentados e publicados nos anais do Simpósio

Brasileiro de Sistemas de Informação – SBSI 2016 e o terceiro foi apresentado e publicado nos anais da *10th International Conference on Agents and Artificial Intelligence – ICAART 2018*.

O primeiro trata-se de uma revisão sistemática submetido e publicado na trilha principal do evento. Está intitulado como "Representação de Gestos em Análise Sintática: uma Revisão Sistemática da Literatura" (FEITOSA; PERES; LIMA, 2016b). Nesse mesmo evento foi apresentado um resumo do projeto de pesquisa prévio a esse trabalho, publicado e apresentado no IX Workshop de Teses e Dissertações em Sistemas de Informação – WTDSI 2016. Esse, por sua vez, está intitulado como "Estudo da Representação de Dados para Segmentação das Fases do Gesto usando aprendizado de máquina" (FEITOSA; PERES; LIMA, 2016a).

Como terceira publicação, há um resumo estendido com um recorte dos primeiros resultados obtidos com o MLP sobre o problema apresentado nesse trabalho, publicado e apresentados sob o título: "*Multidimensional Representations for the Gesture Phase Segmentation Problem: An Exploratory Study using Multilayer Perceptrons*" (FEITOSA et al., 2018).

6.2 Limitações e trabalhos futuros

O foco desse trabalho foi o estudo de representações de dados. Os algoritmos de aprendizado de máquina, neste contexto, foram ferramentas que propiciaram a realização do estudo e, dessa forma, não foi escopo desse trabalho explorá-los como objeto de pesquisa. Por esse motivo, não foram exploradas técnicas de processamento das representações, como o HMM e outras mencionadas ao longo desse trabalho.

Os resultados apresentados nesse trabalho evidenciaram que o uso de representações para as fases do gesto não é genérico a qualquer gesticulador e contexto. As representações que apresentaram os melhores resultados, apesar de se manterem como as melhores quanto ao seu poder discriminativo para descoberta de conhecimento sobre os gestos independente do usuário e vídeo sobre o qual foram aplicados, não apresentaram o mesmo desempenho entre todos eles. Os resultados do MLP, por exemplo, foram melhores sobre o vídeo a1, por mais que as mesmas representações que obtiveram bons resultados para esse vídeo, também o obtiveram para os demais, mesmo que em menor grau. No SOM, os melhores

resultados foram obtidos para o vídeo b1, também mantendo a consistência de ranking de resultados para os demais vídeos, mas com resultados de menor qualidade.

Assim, um dos trabalhos futuros que podem ser derivados desse é o estudo da adição de um aspecto que identifique o contexto do vídeo ou como ele pode ser tratado para compor a representação dos gestos que será submetida às ferramentas de reconhecimento de padrões. É sabido que esse estudo vem sendo realizado, inclusive, pelo grupo de pesquisa no qual o presente trabalho se insere.

Como já mencionado, as análises sobre os resultados do SOM não levaram à conclusão de que é possível sugerir uma nova fase para representar o gesto além das já apresentadas por Kendon (1980), McNeill (1992) e McNeill (2015). Porém, o volume de resultados produzidos com as execuções do SOM ainda tem potencial para revelar conhecimento a partir de um aprofundamento das análises, com outros critérios, métricas e examinando o conteúdo de cada grupo de forma diferente do realizado por esse trabalho.

Foi notória a capacidade das representações do gesto apresentadas em conseguirem discriminar informação suficiente para que as ferramentas de aprendizado de máquina conseguissem aprender padrões sobre os gestos. Porém, como já mencionado, essas ferramentas foram utilizadas para estudo dos modelos de representação e não foram estudadas quando ao seu potencial de otimização com variações de parâmetros e outras configurações. As conclusões referentes às diferentes representações das fases do gesto são válidas por todas terem sido executadas sobre um contexto controlado, com os mesmos dados e com as mesmas ferramentas. Mas cabe, como trabalhos futuros, explorar as ferramentas de aprendizado de máquina de forma minuciosa e mais exaustiva.

Outro ponto importante a ser salientado é o desbalanceamento dos dados do *dataset* utilizado nesse trabalho. Os resultados foram obtidos certificando-se que todas as fases do gesto estavam representadas nas etapas de treinamento, validação e testes dos algoritmos, porém o balanceamento não foi garantido. Durante uma gesticulação é natural imaginar que haja uma distância entre a quantidade de *frames* relacionados à cada fase do gesto, como pode ser visto na tabela 1. No vídeo a1, por exemplo, há consideravelmente mais *frames* representando o *stroke* do que o *hold*, enquanto no vídeo b1 há pouco representantes do descanso e da retração. No problema explorado por Souza et al. (2016) fica claro que algoritmos de aprendizado de máquina assumem que os dados utilizados para treinamento e testes apresentam a mesma distribuição probabilística de forma que classes desbalanceadas podem afetar negativamente seu desempenho. Assim, como trabalho futuro, cabe também

explorar e tratar o efeito do desbalanceamento sobre os resultados da segmentação das fases do gesto.

Não há a garantia de que as representações geradas a partir das combinações dos aspectos dos gestos propostas pelo presente trabalho sejam as mais discriminantes dentre todas as combinações possíveis para a tarefa de segmentação das fases do gesto. Porém, essa não era uma pretensão direta desse estudo. O objetivo foi encontrar a representação que se mostra suficientemente discriminante para a tarefa em questão. Assim, explorar outros aspectos não mostrados nesse trabalho quanto ao seu poder discriminativo também se mostra uma tarefa promissora para contribuir com a expansão do conhecimento na área de análise de gestos.

Referências¹

- ABID, M.; PETRIU, E.; AMJADIAN, E. Dynamic sign language recognition for smart home interactive application using stochastic linear formal grammar. *IEEE Trans. on Inst. and Measurement*, v. 64, n. 3, p. 596–605, mar. 2015. ISSN 0018-9456. Citado na página 36.
- BAILADOR, G.; TRIVIÑO, G. Pattern recognition using temporal fuzzy automata. *Fuzzy Sets and Systems*, Elsevier, v. 161, n. 1, p. 37–55, 2010. Citado 2 vezes nas páginas 36 e 37.
- BARALDI, S.; BIMBO, A. D.; LANDUCCI, L. Natural interaction on tabletops. *Multimedia Tools and Applications*, v. 38, n. 3, p. 385–405, 2008. ISSN 1573-7721. Disponível em: <http://dx.doi.org/10.1007/s11042-007-0195-7>. Citado na página 16.
- BIOLCHINI, J. et al. *Systematic rev. in Soft. Eng.* System Engineering and Computer Science Department COPPE/UFRJ, Technical Report ES-679-05, 2005. 45 p. Citado na página 24.
- BRYLL, R.; QUEK, F.; ESPOSITO, A. Automatic hand hold detection in natural conversation. In: *IEEE Workshop on Cues in Communication*. [S.l.: s.n.], 2001. Citado na página 32.
- CARAMIAUX, B.; WANDERLEY, M. M.; BEVILACQUA, F. Segmenting and parsing instrumentalists' gestures. *J. of New Music Research*, v. 41, n. 1, p. 13–29, 2012. Citado na página 36.
- CARIDAKIS, G.; ASTERIADIS, S.; KARPOUZIS, K. Non-manual cues in automatic sign language recognition. *Personal and Ubiquitous Comput.*, v. 18, n. 1, p. 37–46, 2014. Citado na página 35.
- DARDAS, N. H.; GEORGANAS, N. D. Real-time hand gesture detection and recognition using bag-of-features and support vector machine techniques. *IEEE Trans. Inst. and Measurement*, v. 60, n. 11, p. 3592–3607, 2011. Citado na página 35.
- DARDAS, N. H.; SILVA, J. M.; EL-SADDIK, A. Target-shooting exergame with a hand gesture control. *Multimedia Tools Appl.*, v. 70, n. 3, p. 2211–2233, 2014. Citado na página 35.
- DIETTERICH, T. G. Machine learning for sequential data: A review". In: *Proc. of Joint Structural, Syntactic, and Statistical Pattern Recogn. International Workshops*. [S.l.]: Springer, 2002. p. 15–30. Citado 2 vezes nas páginas 19 e 34.
- DROSOU, A. et al. Spatiotemporal analysis of human activities for biometric authentication. *Computer Vision and Image Understanding*, Elsevier, v. 116, n. 3, p. 411–421, 2012. Citado 2 vezes nas páginas 16 e 19.
- FAUSETT, L. (Ed.). *Fundamentals of Neural Networks: Architectures, Algorithms, and Applications*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1994. ISBN 0-13-334186-0. Citado 6 vezes nas páginas 38, 39, 42, 57, 67 e 74.

¹ De acordo com a Associação Brasileira de Normas Técnicas. NBR 6023.

FEITOSA, R. A.; PERES, S. M.; LIMA, C. A. Estudo da representação de dados para segmentação das fases do gesto usando aprendizado de máquina. In: SIMPÓSIO BRASILEIRO DE SISTEMAS DE INFORMAÇÃO - SBSI. *Anais do IX Workshop de Teses e Dissertações em Sistemas de Informação*. [S.l.], 2016. p. 27–29. Citado na página 96.

FEITOSA, R. A.; PERES, S. M.; LIMA, C. A. Representação de gestos em análise sintática: uma revisão sistemática da literatura. In: SIMPÓSIO BRASILEIRO DE SISTEMAS DE INFORMAÇÃO - SBSI. *Proceedings of the XII Brazilian Symposium on Information Systems*. [S.l.], 2016. p. 502–509. Citado na página 96.

FEITOSA, R. A. et al. Multidimensional representations for the gesture phase segmentation problem - an exploratory study using multilayer perceptrons. In: INSTICC. *Proceedings of the 10th International Conference on Agents and Artificial Intelligence - Volume 2: ICAART*. [S.l.]: SciTePress, 2018. p. 347–354. ISBN 978-989-758-275-2. Citado na página 96.

FREITAS, F. et al. Grammatical facial expressions recognition with machine learning. In: AAAI. *Proceedings of 27th Florida artificial intelligence research society conference*. [S.l.], 2014. p. 180–185. Citado na página 19.

GAGLIO, S. et al. Gesture recognition for improved user experience in a smart environment. In: SPRINGER. *Congress of the Italian Association for Artificial Intelligence*. [S.l.], 2013. p. 493–504. Citado na página 37.

GEBRE, B. G.; WITTENBURG, P.; LENKIEWICZ, P. Towards automatic gesture stroke detection. In: EUROPEAN LANGUAGE RESOURCES ASSOCIATION. *LREC 2012: 8th International Conference on Language Resources and Evaluation*. [S.l.], 2012. p. 231–235. Citado 2 vezes nas páginas 18 e 32.

GIBBON, D. Gesture theory is linguistics: On modelling multimodality as prosody. In: *PACLIC*. [S.l.: s.n.], 2009. p. 9–18. Citado na página 27.

HACHAJ, T.; OGIELA, M. R. Rule-based approach to recognizing human body poses and gestures in real time. *Multimedia Syst.*, v. 20, n. 1, p. 81–99, 2014. Citado 2 vezes nas páginas 36 e 37.

HAYKIN, S. S. *Neural networks and learning machines*. [S.l.]: Pearson Upper Saddle River, NJ, USA, 2009. v. 3. Citado 9 vezes nas páginas 38, 39, 40, 41, 42, 43, 44, 45 e 58.

JACOB, M. G.; WACHS, J. P. Context-based hand gesture recognition for the operating room. *Pattern Recognition Letters*, v. 36, p. 196 – 203, 2014. ISSN 0167-8655. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0167865513002225>. Citado na página 20.

KANTZ, H.; SCHREIBER, T. *Nonlinear time series analysis*. [S.l.]: Cambridge university press, 2004. v. 7. Citado na página 53.

KENDON, A. Gesticulation and speech: Two aspects of the process of utterance. *The Relationship of verbal and nonverbal communication*, Mouton Publishers, p. 207–227, 1980. Citado 10 vezes nas páginas 16, 22, 27, 30, 31, 32, 33, 93, 95 e 97.

KHAN, S.; BAILEY, D.; GUPTA, G. S. Detecting pauses in continuous sign language. In: IEEE. *Mechatronics and Machine Vision in Practice (M2VIP)*, 2012 19th International Conference. [S.l.], 2012. p. 11–15. Citado 2 vezes nas páginas 36 e 37.

KIM, D.; SONG, J.; KIM, D. Simultaneous gesture segmentation and recognition based on forward spotting accumulative {HMMs}. *Pattern Recognition*, v. 40, n. 11, p. 3012 – 3026, 2007. ISSN 0031-3203. Disponível em: <http://www.sciencedirect.com/science/article/pii/S003132030700088X>. Citado na página 16.

KITA, S.; GIJN, I.; HULST, H. Movement phases in signs and co-speech gestures, and their transcription by human coders. In: *Proc. of Int. Gesture Workshop Bielefeld*. [S.l.]: Springer, 1998. p. 23–35. Citado 4 vezes nas páginas 17, 23, 28 e 31.

KOHONEN, T. *Self-Organizing Maps*. [S.l.]: Springer Berlin Heidelberg, 2001. (Physics and astronomy online library). ISBN 9783540679219. Citado 2 vezes nas páginas 43 e 61.

KOJIMA, A.; TAMURA, T.; FUKUNAGA, K. Natural language description of human activities from video images based on concept hierarchy of actions. *International Journal of Computer Vision*, Springer, v. 50, n. 2, p. 171–184, 2002. Citado 2 vezes nas páginas 36 e 37.

KRIPPENDORFF, K. *Content Analysis: An Introduction to Its Methodology*. [S.l.]: SAGE Publications, 2012. ISBN 9781483307039. Citado na página 48.

KYAN, M. et al. An approach to ballet dance training through ms kinect and visualization in a cave virtual reality environment. *ACM Trans. Intell. Syst. Technol.*, ACM, New York, NY, USA, v. 6, n. 2, p. 23:1–23:37, mar. 2015. ISSN 2157-6904. Citado na página 35.

KYRIAZIS, N. et al. A generative approach to tracking hands and their interaction with objects. In: *Man–Machine Interactions 4*. [S.l.]: Springer, 2016. p. 19–28. Citado na página 22.

LEE, G. C.; YEH, F.-H.; HSIAO, Y.-H. Kinect-based taiwanese sign-language recognition system. *Multimedia Tools and Applications*, v. 75, n. 1, p. 261–279, 2016. ISSN 1573-7721. Disponível em: <http://dx.doi.org/10.1007/s11042-014-2290-x>. Citado na página 19.

LIANG, H.; YUAN, J.; THALMANN, D. Parsing the hand in depth images. *IEEE Trans. on Multimedia*, v. 16, n. 5, p. 1241–1253, aug. 2014. ISSN 1520-9210. Citado na página 35.

LIMA, C. A. M.; COELHO, A. L. V. Kernel machines for epilepsy diagnosis via eeg signal classification: A comparative study. *Artif. Intell. Med.*, Elsevier Science Publishers Ltd., Essex, UK, v. 53, n. 2, p. 83–95, out. 2011. ISSN 0933-3657. Disponível em: <http://dx.doi.org/10.1016/j.artmed.2011.07.003>. Citado na página 51.

LIU, J. et al. Non-manual grammatical marker recognition based on multi-scale, spatio-temporal analysis of head pose and facial expressions. *Image Vision Comput.*, v. 32, n. 10, p. 671–681, 2014. Citado na página 35.

LIU, S. et al. Fashion parsing with weak color-category labels. *IEEE Trans. on Multimedia*, v. 16, n. 1, p. 253–265, jan. 2014. ISSN 1520-9210. Citado 2 vezes nas páginas 34 e 35.

LIU, S. et al. Fashion parsing with video context. *IEEE Trans. on Multimedia*, v. 17, n. 8, p. 1347–1358, aug. 2015. Citado 2 vezes nas páginas 34 e 35.

LÜCKING, A. et al. Data-based analysis of speech and gesture: the Bielefeld Speech and Gesture Alignment corpus (SaGA) and its applications. *Journal on Multimodal User Interfaces*, Springer Science + Business Media, v. 7, n. 1-2, 2013. Citado na página 36.

MADEO, R. C. B. et al. A committee machine implementing the pattern recognition module for fingerspelling applications. In: *Proceedings of the 2010 ACM Symposium on Applied Computing*. New York, NY, USA: ACM, 2010. (SAC '10), p. 954–958. ISBN 978-1-60558-639-7. Disponível em: <http://doi.acm.org/10.1145/1774088.1774287>. Citado na página 19.

MADEO, R. C. B.; PERES, S. M.; LIMA, C. A. Gesture phase segmentation using support vector machines. *Expert Systems with Applications*, v. 56, p. 100 – 115, 2016. ISSN 0957-4174. In press. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0957417416300525>. Citado 6 vezes nas páginas 18, 19, 20, 21, 48 e 56.

MADEO, R. C. B.; WAGNER, P. K.; PERES, S. M. A review of temporal aspects of hand gesture analysis applied to discourse analysis and natural conversation. *Int. J. of C. Sci. & Inf. Tech.*, v. 5, n. 4, dec. 2013. Citado 2 vezes nas páginas 19 e 36.

MARTELL, C.; KROLL, J. Corpus-based gesture analysis: an extension of the form dataset for the automatic detection of phases in a gesture. *International Journal of Semantic Computing*, World Scientific, v. 1, n. 04, p. 521–536, 2007. Citado na página 32.

MCNEILL, D. Hand and mind: What the hands reveal about thought. University of Chicago Press, 1992. Citado 10 vezes nas páginas 16, 22, 27, 30, 31, 32, 33, 93, 95 e 97.

MCNEILL, D. *Why We Gesture: The Surprising Role of Hand Movements in Communication*. [S.l.]: Cambridge University Press, 2015. ISBN 9781107137189. Citado 10 vezes nas páginas 16, 22, 27, 30, 31, 32, 33, 93, 95 e 97.

NGUYEN, T. D.; RANGANATH, S. Facial expressions in american sign language: Tracking and recognition. *Pattern Recogn.*, v. 45, n. 5, p. 1877–1891, 2012. ISSN 0031-3203. Citado na página 35.

ONG, S. C.; RANGANATH, S. Automatic sign language analysis: A survey and the future beyond lexical meaning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 27, n. 6, p. 873–891, Jun 2005. ISSN 0162-8828. Citado 2 vezes nas páginas 16 e 19.

OZTURK, O. et al. Boosting real-time recognition of hand posture and gesture for virtual mouse operations with segmentation. *Applied Intelligence*, v. 43, n. 4, p. 786–801, 2015. ISSN 1573-7497. Disponível em: <http://dx.doi.org/10.1007/s10489-015-0680-z>. Citado na página 22.

POPA, D. et al. Real time trajectory based hand gesture recognition. *WSEAS Transactions on Information Science and Applications*, Citeseer, v. 5, n. 4, p. 532–546, 2008. Citado na página 32.

QUEK, F. et al. Multimodal human discourse: gesture and speech. *ACM Transactions on Computer-Human Interaction (TOCHI)*, ACM, v. 9, n. 3, p. 171–193, 2002. Citado na página 31.

RAMAKRISHNAN, A. S.; NEFF, M. Segmentation of hand gestures using motion capture data. In: INTERNATIONAL FOUNDATION FOR AUTONOMOUS AGENTS AND MULTIAGENT SYSTEMS. *Proceedings of the 2013 international conference on Autonomous agents and multi-agent systems*. [S.l.], 2013. p. 1249–1250. Citado 2 vezes nas páginas 18 e 32.

ROSANI, A.; CONCI, N.; NATALE, F. G. B. D. Human behavior recognition using a context-free grammar. *J. Electronic Imaging*, v. 23, n. 3, p. 033016, 2014. Citado na página 36.

SALEM, M. et al. Generation and evaluation of communicative robot gesture. *International Journal of Social Robotics*, v. 4, n. 2, p. 201–217, 2012. ISSN 1875-4805. Disponível em: <http://dx.doi.org/10.1007/s12369-011-0124-9>. Citado na página 20.

SEMMLOW, J.; GRIFFEL, B. *Biosignal and Medical Image Processing, Third Edition*. [S.l.]: Taylor & Francis, 2014. ISBN 9781466567368. Citado 2 vezes nas páginas 50 e 51.

SILVA, L. da; PERES, S.; BOSCARIOLI, C. *Introdução à Mineração de Dados: Com Aplicações em R*. [S.l.]: Elsevier Brasil, 2017. ISBN 9788535284478. Citado 2 vezes nas páginas 39 e 43.

SMITH, N. A. *Linguistic Structure Prediction*. [S.l.]: Morgan & Claypool, 2011. 247 p. Citado 2 vezes nas páginas 19 e 34.

SOUZA, J. C. S. de et al. Recent advances for handling imbalance and uncertainty in labelling in medicinal chemistry data analysis. In: IEEE. *SAI Computing Conference (SAI), 2016*. [S.l.], 2016. p. 217–222. Citado na página 97.

SPANO, L. D.; CISTERNINO, A.; PATERNÒ, F. A compositional model for gesture definition. In: SPRINGER. *International Conference on Human-Centred Software Engineering*. 2012. p. 34–52. ISBN 978-3-642-34347-6. Disponível em: http://dx.doi.org/10.1007/978-3-642-34347-6_3. Citado na página 19.

SUAREZ, J.; MURPHY, R. R. Hand gesture recognition with depth images: A review. In: *2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication*. [S.l.: s.n.], 2012. p. 411–417. ISSN 1944-9445. Citado na página 22.

SUBASI, A. Eeg signal classification using wavelet feature extraction and a mixture of expert model. *Expert Syst. Appl.*, Pergamon Press, Inc., Tarrytown, NY, USA, v. 32, n. 4, p. 1084–1093, maio 2007. ISSN 0957-4174. Disponível em: <http://dx.doi.org/10.1016/j.eswa.2006.02.005>. Citado na página 51.

VESANTO, J. et al. Som toolbox for matlab 5. *Helsinki University of Technology, Finland*, 2000. Citado na página 60.

WAGNER, P. et al. Gesture unit segmentation using spatial-temporal information and machine learning. In: . The AAAI Press, 2014. p. 101–106. Disponível em: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-84923853806&partnerID=40&md5=ba63a06ec66edcca1bc2349860ca3085>. Citado na página 18.

WAGNER, P. K. et al. Gesture unit segmentation using spatial-temporal information and machine learning. In: *Proceedings of the 27th Florida Artificial Intelligence Research Society Conference (FLAIRS)*. [S.l.]: AAAI Press, 2014. p. 101–106. Citado na página 48.

XU, W.; LEE, E.-J. Hand gesture recognition using improved hidden markov models. *J. of Korea Multimedia Soc.*, v. 14, n. 7, p. 866–871, 2011. Citado na página 35.

YIN, Y.; DAVIS, R. Real-time continuous gesture recognition for natural human-computer interaction. In: *2014 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*. [S.l.: s.n.], 2014. p. 113–120. ISSN 1943-6092. Citado 2 vezes nas páginas 18 e 32.

ZHANG, X. et al. Human pose estimation and tracking via parsing a tree structure based human model. *IEEE Trans. Syst., Man, and Cyber.: Systems*, v. 44, n. 5, p. 580–592, 2014. Citado na página 34.

ZHAO, L. et al. Tracking human pose using max-margin markov models. *IEEE Trans. on Image Proces.*, v. 24, n. 12, p. 5274–5287, 2015. Citado na página 34.

ZHU, L. et al. Max margin learning of hierarchical configural deformable templates (hcdts) for efficient object parsing and pose estimation. *Int. J. of Comp. Vision*, v. 93, n. 1, p. 1–21, 2011. Citado na página 35.