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

JALLYSSON MIRANDA ROCHA

**A influência do contexto de discurso na segmentação automática das fases do
gesto com aprendizado de máquina supervisionado**

São Paulo

2018

JALLYSSON MIRANDA ROCHA

**A influência do contexto de discurso na segmentação automática das fases do
gesto com aprendizado de máquina supervisionado**

Versão original

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

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

Rocha, Jallysson Miranda

A influência do contexto de discurso na segmentação automática das fases do gesto com aprendizado de máquina supervisionado / Jallysson Miranda Rocha ; orientadora, Sarajane Marques Peres ; coorientador, Clodoaldo Aparecido de Moraes Lima. – 2018.
97 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 original

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 Jallysson Miranda Rocha, sob o título “**A influência do contexto de discurso na segmentação automática das fases do gesto com aprendizado de máquina supervisionado**”, 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 _____ de _____ de _____ pela comissão julgadora constituída pelos doutores:

Prof. Dr. _____

Instituição: _____

Presidente

Prof. Dr. _____

Instituição: _____

Prof. Dr. _____

Instituição: _____

Prof. Dr. _____

Instituição: _____

Resumo

ROCHA, Jallysson Miranda. **A influência do contexto de discurso na segmentação automática das fases do gesto com aprendizado de máquina supervisionado**. 2018. 97 f. Dissertação (Mestrado em Ciências) – Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, São Paulo, 2018.

Gestos são ações que fazem parte da comunicação humana. Frequentemente, eles ocorrem junto com a fala e podem se manifestar por uma ação proposital, como o uso das mãos para explicar o formato de um objeto, ou como um padrão de comportamento, como coçar a cabeça ou ajustar os óculos. Os gestos ajudam o locutor a construir sua fala e também ajudam o ouvinte a compreender a mensagem que está sendo transmitida. Pesquisadores de diversas áreas são interessados em entender como se dá a relação dos gestos com outros elementos do sistema linguístico, seja para suportar estudos das áreas da Linguística e da Psicolinguística, seja para melhorar a interação homem-máquina. Há diferentes linhas de estudo que exploram essa temática e entre elas está aquela que analisa os gestos a partir de fases: preparação, *pré-stroke hold*, *stroke*, *pós-stroke hold*, *hold* e retração. Assim, faz-se útil o desenvolvimento de sistemas capazes de automatizar a segmentação de um gesto em suas fases. Técnicas de aprendizado de máquina supervisionado já foram aplicadas a este problema e resultados promissores foram obtidos. Contudo, há uma dificuldade inerente à análise das fases do gesto, a qual se manifesta na alteração do contexto em que os gestos são executados. Embora existam algumas premissas básicas para definição do padrão de manifestação de cada fase do gesto, em contextos diferentes tais premissas podem sofrer variações que levariam a análise automática para um nível de alta complexidade. Este é o problema abordado neste trabalho, a qual estudou a variabilidade do padrão inerente à cada uma das fases do gesto, com apoio de aprendizado de máquina, quando a manifestação delas se dá a partir de um mesmo indivíduo, porém em diferentes contextos de produção do discurso. Os contextos de discurso considerados neste estudo são: contação de história, improvisação, descrição de cenas, entrevistas e aulas expositivas.

Palavras-chaves: Análise de Gesto. Fases do Gesto. Contexto de Discurso. Segmentação Automática. Aprendizado de Máquina. Máquina de Vetores Suporte.

Abstract

ROCHA, Jallysson Miranda. **The influence of the speech context on the automatic segmentation of the phases of the gesture with supervised machine learning.** 2018. 97 p. Dissertation (Master of Science) – School of Arts, Sciences and Humanities, University of São Paulo, São Paulo, 2018.

Gestures are actions that make part of human communication. Commonly, gestures occur at the same time as the speech and they can manifest either through an intentional act, as using the hands to explain the format of an object, or as a pattern of behavior, as scratching the head or adjusting the glasses. Gestures help the speaker to build their speech and also help the audience to understand the message being communicated. Researchers from several areas are interested in understanding what the relationship of gestures with other elements of the linguistic system is like, whether in supporting studies in Linguistics or Psycho linguistics, or in improving the human-machine interaction. There are different lines of study that explore such a subject, and among them is the line that analyzes gestures according to their phases: preparation, pre-stroke hold, stroke, post-stroke hold, hold and retraction. Thus, the development of systems capable of automating the segmentation of gestures into their phases can be useful. Techniques that implement supervised machine learning have already been applied in this problem and promising results have been achieved. However, there is an inherent difficulty to the analysis of phases of gesture that is revealed when the context (in which the gestures are performed) changes. Although there are some elementary premises to set the pattern of expression of each gesture phase, such premises may vary and lead the automatic analysis to high levels of complexity. Such an issue is addressed in the work herein, whose purpose was to study the variability of the inherent pattern of each gesture phase, using machine learning techniques, when their execution is made by the same person, but in different contexts. The contexts of discourse considered in this study are: storytelling, improvisation, description of scenes, interviews and lectures.

Keywords: Analysis of Gesture. Gesture Phases. Discourse Context. Automatic Segmentation. Machine Learning. Support Vector Machine.

Lista de figuras

Figura 1 – Quadros amostrados de um vídeo: fases do gesto	13
Figura 2 – Quadros referentes à transição de uma fase <i>stroke</i> para uma fase de retração	14
Figura 3 – Aplicabilidade da segmentação das fases do gesto	14
Figura 4 – Classificação de dados linearmente separáveis com SVM	24
Figura 5 – Representação de um conjunto de dados não linear	24
Figura 6 – Segmentação das fases do gesto	30
Figura 7 – Erros de segmentação no problema de segmentação das fases do gesto .	32
Figura 8 – Intervalo de qualidade dos valores de K do coeficiente <i>Krippendorff's Alpha</i>	33
Figura 9 – Esquema das atividades de um gesticulador: durante uma sessão, um gesticulador participa da captação de vídeos para um ou mais contextos. Nesse exemplo, <i>A</i> , <i>B</i> , <i>C</i> , <i>D</i> dizem respeito a quatro contextos diferentes	37
Figura 10 – Pontos de articulação no corpo do gesticulador: mãos, pulsos, cotovelos, tronco e cabeça	38
Figura 11 – Sensor <i>Microsoft Kinect</i> e seus componentes	39
Figura 12 – Exemplos para análise de rotulação considerando ambas as mãos e gerando um rótulo, ou considerando cada uma das mãos, gerando dois rótulos	41
Figura 13 – Amostragem de quadros rotulados para a análise discordância para um vídeo: (a) com todos os quadros; (b) para um vídeo com redução de períodos de descansos	45
Figura 14 – Destaque para as discordâncias de rotulação	46
Figura 15 – Efeitos da normalização aplicada às coordenadas espaciais <i>x</i> e <i>y</i> em um trecho de vídeo	48
Figura 16 – Filtro <i>wavelet</i> aplicado na coordenada tridimensional <i>x</i> da mão direita.	49
Figura 17 – Sinal correspondente à velocidade (quadro superior) e aceleração (quadro inferior) escalares da mão direita	51
Figura 18 – Uma janela retangular para construir uma representação vetorial do tempo em relação ao quadro de interesse	52

Figura 19 – Quantidade (figura a esquerda) e porcentagem (figura a direita) de quadros por fase do gesto no <i>corpus</i> , considerando os seis vídeos rotulados pelos especialistas E1 e E2	53
Figura 20 – Padrão de gesticulação do gesticulador G1 em sessões diferentes para os contextos: contação de histórias (azul), aulas expositivas (amarelo) .	54
Figura 21 – Padrão de gesticulação do gesticulador G1 em contextos diferentes: improvisação (azul), contação de histórias (laranja), aula expositiva (cinza)	54
Figura 22 – Composição de conjuntos de treinamento usando n vídeos: $n - 1$ vídeos para treinamento e um vídeo para teste. Cada vídeo de m quadros é usado em sua totalidade	58

Lista de tabelas

Tabela 1 – Exemplo de matriz de confusão multiclasse	23
Tabela 2 – Trabalhos correlatos x aplicação	34
Tabela 3 – Distribuição dos vídeos que compõem o <i>corpus</i> por gesticulador G_i e por contexto C_j	40
Tabela 4 – Ilustração da organização das rotulações junto aos quadros de vídeo. .	42
Tabela 5 – Rotulações (Rot) e Coeficientes de Concordância (<i>Krippendorf's Alpha</i>) para oito vídeos	42
Tabela 6 – Quantidade de quadros por fases do gesto segundo rotulação realizada observando as duas mãos, por cada especialista, para oito vídeos	43
Tabela 7 – Coeficientes de Concordância (<i>Krippendorf's Alpha</i>) para vídeos rotula- dos por dois especialistas após supressão de quadros de vídeo de fácil rotulação	44
Tabela 8 – Matriz numérica que representa um vídeo	46
Tabela 9 – Espaço de busca relacionado à representação do gesto	59
Tabela 10 – Espaço de busca referente aos parâmetros da SVM	59
Tabela 11 – Melhores parâmetros para a tarefa de segmentação das fase do gesto em estudo neste trabalho	60
Tabela 12 – Melhores resultados entre os classificadores gerados no experimento #1.1	63
Tabela 13 – Matriz de confusão para o melhor resultado obtido no experimento #1.1	63
Tabela 14 – Melhor resultado entre os modelos classificadores gerados no experi- mento #2 para o gesticulador A com treinamento na história 1 e teste na história 2	64
Tabela 15 – Matriz de confusão para o melhor resultado obtido no experimento #2	65
Tabela 16 – Resultado do modelo treinado com a história 1 e testado com a história 3	65
Tabela 17 – Melhor resultado entre os modelos classificadores gerados no experi- mento #1.3 para o treinamento na história 2 e 3 e teste na história 1	66
Tabela 18 – Matriz de confusão para o melhor resultado obtido no experimento #1.3	66
Tabela 19 – Desempenho dos classificadores para cada fase do gesto, em termos de F -score máximo e mínimo.	67

Tabela 20 – Análise qualitativa dos melhores modelos obtidos sob cada rotulação .	68
Tabela 21 – Detalhamento dos erros de substituição por fase do gesto	69
Tabela 22 – Resultados gerais do processo de seleção dos melhores modelos classifi- cadores	71
Tabela 23 – Valores máximos e mínimos de F -score obtidos nos melhores modelos de cada característica em cada fase do gesto e no período de descanso .	72
Tabela 24 – Parâmetros usados nos melhores modelos de SVM obtidos com repre- sentação espacial sem uso de filtro	72
Tabela 25 – Parâmetros usados nos melhores modelos de SVM obtidos com repre- sentação por velocidade escalar sem uso de filtro	73
Tabela 26 – Valores máximos e mínimos de F -score de cada vídeo entre os melhores modelos nos dados espaciais e velocidade escalar em cada fase do gesto e período de descanso	74
Tabela 27 – Valores máximos e mínimos de F -score do treinamento dos modelos em V1.11.1 e teste nos outros vídeos nos dados espaciais em cada fase do gesto e período de descanso	75
Tabela 28 – Valores máximos e mínimos de F -score do treinamento dos modelos em V1.11.1 e teste nos outros vídeos nos dados representados por velocidade escalar em cada fase do gesto e período de descanso	76
Tabela 29 – Valores máximos e mínimos de F -score do treinamento dos modelos em coordenadas espacial e teste nos vídeos com a linha em destaque de cinza para cada parte da tabela	77
Tabela 30 – Valores máximos e mínimos de F -score do treinamento dos modelos em velocidade escalar e teste nos vídeos com a linha em destaque de cinza para cada parte da tabela	78
Tabela 31 – Valores máximos e mínimos de F -score do treinamento dos modelos em coordenadas espaciais e teste nos vídeos com a linha em destaque de cinza para cada parte da tabela	80
Tabela 32 – Valores máximos e mínimos de F -score do treinamento dos modelos em velocidade escalar e teste nos vídeos com a linha em destaque de cinza para cada parte da tabela	81

Sumário

1	Introdução	12
1.1	<i>Aplicabilidade da segmentação das fases do gesto</i>	14
1.2	<i>Definição do problema de segmentação das fases do gesto</i>	15
1.3	<i>Hipótese</i>	16
1.4	<i>Objetivos</i>	16
1.5	<i>Método</i>	17
1.6	<i>Organização da dissertação</i>	18
2	Aprendizado de máquina	20
2.1	<i>Avaliação de classificadores</i>	21
2.2	<i>Máquinas de vetores suporte</i>	23
2.3	<i>Considerações finais</i>	25
3	Análise automatizada das fases do gesto	26
3.1	<i>Teoria de gestos</i>	26
3.1.1	<i>Fases do gesto</i>	27
3.1.2	<i>Tipos de gestos</i>	28
3.2	<i>Segmentação das fases do gesto</i>	29
3.2.1	<i>Avaliação da segmentação das fases</i>	30
3.2.2	<i>Avaliação da concordância de rotulação</i>	32
3.3	<i>Trabalhos correlatos</i>	33
3.4	<i>Considerações finais</i>	34
4	Corpus de vídeos com gesticulação natural	36
4.1	<i>Projeto no comitê de ética em pesquisa</i>	36
4.2	<i>Execução da captação de dados</i>	38
4.3	<i>Rotulação dos vídeos</i>	40
4.4	<i>Pre-processamento e extração de características</i>	45
4.5	<i>Corpus versus conjunto de dados</i>	52
5	Experimentos, resultados e discussão	56
5.1	<i>Organização dos experimentos</i>	56

5.1.1	Estratégia de treino e teste de classificadores	57
5.1.2	Parametrização	58
5.2	<i>Resultados e discussões</i>	62
5.2.1	Resultados e análises: experimento do tipo #1	62
5.2.2	Resultados e análises: experimento do tipo #2	66
5.2.3	Resultados e análises: experimento do tipo #3	68
5.3	<i>Considerações finais</i>	81
6	Conclusões	82
6.1	<i>Contribuições</i>	83
6.2	<i>Trabalhos futuros</i>	84
	Referências ¹	86
	Apêndice A – Termo de consentimento livre e esclarecido . .	91
	Apêndice B – Roteiro de atividades dos participantes	94
	Apêndice C – Roteiro de perguntas do contexto de entrevista	95
	Apêndice D – Esquema gráfico de combinações de vídeos . . .	96
	Apêndice E – Filtro <i>Wavelet</i>	97

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

1 Introdução

Recentemente, a área de pesquisa em análise de gestos, manuais ou não manuais¹ têm recebido uma forte atenção de pesquisadores na computação. Tais pesquisas tem focado no desenvolvimento de novos métodos para interação humano computador, suportando a criação de aplicações das mais diversas naturezas. Jogos (LIU; KAVAKLI, 2010), interface de software, controle de equipamentos via gestos (PICKERING, 2005) e softwares de monitoramento e segurança são alguns exemplos de áreas de aplicação.

Em uma linha de aplicação mais acadêmica, destacam-se os pesquisadores da área de Linguística e/ou Psicolinguística interessados em interpretar gestos, fazendo uso de ferramentas computacionais (ALLWOOD et al., 2007; BRUGMAN; RUSSEL, 2004; GIBBON et al., 2003; KIPP, 2012; MARTELL, 2005; MARICCHIOLO; GNISCI; BONAIUTO, 2012) para suportar o seu trabalho de análise de discurso baseado em gestos, ou baseado em gestos concorrentes com a fala. Ferramentas que reproduzem vídeos capturados de pessoas falando e gesticulando, fornecem recursos de transcrição de componentes dos vídeos, como a fala, tom de voz e gestos observados pelos especialistas. Geralmente, esses estudos exigem a transcrição da fala e anotação de um ou vários elementos envolvidos no discurso tais como prosódia, gestos ou a direção do olhar.

A anotação de gestos, interesse do presente trabalho, está baseada em estudos de gestos que, entre outros aspectos, presumem que eles são compostos por unidades gestuais que correspondem ao movimento das mãos de uma posição de descanso há um local de expressividade do gestos, execução do gestos e então retorno à mesma ou à uma nova posição de descanso. As unidades gestuais, por sua vez, são compostas por fases gestuais Kendon (1980). As fases gestuais são segmentos que compõem a unidade gestual caracterizando o significado do gesto. Uma definição das fases gestuais (ilustradas na 1), foi proposta por Kendon (1980) e McNeill (1992), e correspondem a:

- preparação, fase em que as mãos se movem em direção ao local em que ocorrerá o movimento do gesto;
- *pré-stroke hold*, uma breve pausa que ocorre ao final da fase de preparação mantendo a mesma configuração das mãos;

¹ Expressões faciais são consideradas como gestos não manuais no contexto de comunicação em língua de sinais. Na gesticulação natural, movimentos do tronco, expressões faciais, deslocamentos do corpo ou movimentos da perna são todos considerados gestos não manuais.

- *stroke*, fase que ocorre obrigatoriamente no gesto, constituindo sua fase expressiva, ou seja, expressando o seu significado;
- *pós-stroke hold*, uma breve pausa ao final da fase de *stroke* mantendo a mesma configuração das mãos;
- *retração*, fase em que as mãos retornam à posição de descanso.
- Kita, Gijn e Hulst (1998) também definem a fase de *hold*, que caracteriza-se por ser uma fase que não apresenta movimento, mas que expressa significado.

Figura 1 – Quadros amostrados de um vídeo: fases do gesto

Fonte: Jallysson Miranda Rocha, 2018.

Encontrar as fases do gesto (segmentar as fases), consiste em encontrar as fases de preparação, *pré-stroke hold*, *stroke*, *pós-stroke hold*, *hold* e *retração*, considerando uma instância de gesticulação (um trecho de um discurso ou da conversação natural, por exemplo). Tradicionalmente, essa tarefa é resolvida de forma manual por especialistas em Linguística, Psicolinguísticas ou áreas correlatas. Sob um ponto de vista mais operacional, a segmentação das fases do gestos é uma ação na qual dada uma sequência de quadros de um vídeo, um especialista analisa cada quadro e os classifica como pertencente a uma das fases que compõem a unidade gestual. Trata-se portanto de um trabalho que exige tempo e atenção. Um vídeo com cinco minutos de gesticulação pode requerer cerca de oito horas de trabalho de um especialista, considerando a anotação de um ou dois aspectos do discurso. Esse é um trabalho de tomada de decisão com alta subjetividade, principalmente em quadros de transições entre fases. A figura 2 ilustra um período em que ocorre uma mudança de fase. Geralmente, essa mudança é difícil de ser determinada, ou seja, é complexo decidir em qual quadro ocorre a troca de fase do gesto.

Figura 2 – Quadros referentes à transição de uma fase *stroke* para uma fase de retração. Em destaque o quadro que o especialista julgou ser o último da fase de *stroke*.

Fonte: Jallysson Miranda Rocha, 2018.

1.1 Aplicabilidade da segmentação das fases do gesto

Na figura 3 é apresentado um esquema gráfico referente à aplicabilidade da segmentação das fases do gesto. Essa segmentação pode ser aplicada de duas maneiras: como um recurso que gera informação que pode ser usada diretamente na resolução de problemas associados à estudos sobre a prosódia em um discurso, da semântica de um discurso ou de padrões de produção da linguagem; como um recurso de pré-processamento de dados para a implementação de outros tipos de análise do gesto (teorias sobre tipos de gestos e funções de gestos).

Figura 3 – Aplicabilidade da segmentação das fases do gesto

Fonte: Sarajane Marques Peres - notas de aula, 2017

Ainda na figura, alguns exemplos de tarefas de análise de discurso que toma como base a segmentação das fases do gesto são detalhados. No primeiro exemplo, a identificação da fase *stroke* ajuda na identificação de um pico prosódico em um discurso falado. Geralmente, quando há gesticulação acompanhando a fala, a período em que uma ideia importante está sendo comunicada, é acompanhado de um gesto. O locutor assim o faz, de forma consciente ou inconsciente, para enfatizar uma ideia e tentar melhorar a compreensão do ouvinte sobre aquilo que ele quer comunicar. Assim, automação da análise da fala pode ser melhorada se o gesto que acompanha a fala também for analisado.

No segundo exemplo, é apresentada a situação em que a execução de um gesto complementa a expressividade de uma mensagem comunicada por fala. No exemplo, o locutor expressa oralmente a mensagem sobre ter aberto uma porta, e complementa a informação gesticulando de maneira a mostrar, espacialmente e temporalmente, o tipo de porta que ele abriu. Nesse caso, a segmentação correta da fase *stroke*, ou da fase *hold*, separam uma sequência de quadros de vídeo que deverão ser analisados, diante de uma nova tarefa de reconhecimento de padrões, para complementar a informação expressa no discurso.

Por fim, no terceiro exemplo, padrões de fala podem ser identificados a partir da análise de fases dos gestos. Neste caso, a organização das fases em relação a frequência e duração, aspectos facilmente identificados após o trabalho de segmentação das fases, pode ser combinada a análises mais específicas de determinadas fases, como a simetria das mãos em *strokes*, ou a associação de *holds* à trechos (transcritos) da fala.

1.2 Definição do problema de segmentação das fases do gesto

Neste projeto, a segmentação automática das fases do gesto é tratada como uma tarefa de classificação, resolvida a partir da aplicação de Máquinas de vetores Suporte (SVM, do inglês *Support Vector Machines*), uma técnica de aprendizado de máquina supervisionado.

Assim, a segmentação automática das fases do gesto é obtida por meio da análise de uma sequência de quadros $\{f_1 \cdots f_n\}$, consecutivos, correspondente a um vídeo. Esta sequência contém as imagens de uma pessoa em um contexto de discurso (descrevendo uma figura, narrando uma história ou participando de uma entrevista, por exemplo).

O problema de segmentação das fases do gestos, então, é modelado como uma tarefa de classificação na qual, dada uma representação para um quadro f_i de uma sequência de quadros s , o modelo classificador deve classificar esse quadro em uma das classes $c_i = \{D, P, H, S, R\}$, que correspondem respectivamente ao período de descanso e às fases preparação, *hold*, *stroke* e retração.

Contudo, para adequar o problema multiclasse para o uso técnica de SVM clássica, específica para problemas de classificação binária, os modelos de classificação foram obtidos a partir da estratégia “um contra todos”, a qual cria um classificador binário para cada classe, diferenciando-a das demais. Posteriormente, todos classificadores deve ser combinados em uma estrutura que permita considerar a resposta de cada um deles e gerar uma resposta final para a tarefa de classificação.

1.3 Hipótese

O padrão de gesticulação natural, embora analisado em relação a um mesmo executor, é dependente do contexto no qual ele é executado. Variáveis como nível de estresse, nível de atenção, objetivo de comunicação e público alvo, interferem na forma como uma pessoa constrói seu discurso (sua fala e seus gestos). Modelos classificadores têm um grau de sensibilidade à variabilidade de padrões, portanto podem fornecer mecanismos para observá-la. A variabilidade advinda da mudança de contexto será percebida pelos modelos classificadores e será medida a partir das alterações ocorridas no desempenho dos classificadores.

1.4 Objetivos

O objetivo deste projeto é verificar se e como o contexto de execução da gesticulação natural – em termos de gestos manuais - interfere nos padrões que caracterizam a manifestação de cada uma das fases do gesto; essa verificação está baseada na indução de modelos classificadores para segmentação das fases e, em havendo dependência de contexto, como essa dependência influencia no desempenho dos classificadores. Decorrente do objetivo acima delineado, os objetivos específicos são:

- desenvolver o ambiente de experimentação (um *corpus* de vídeos) no qual seja possível induzir e avaliar um conjunto de modelos classificadores, capazes de segmentar as fases do gesto com um nível de qualidade que permita a realização da análise sobre a influência do contexto na manifestação das fases do gesto;
- definir um protocolo de avaliação do desempenho quantitativo dos modelos classificadores, tanto no que diz respeito à avaliação de suas capacidades de segmentação das fases do gestos quanto no que diz respeito à comparação de seus desempenhos para avaliação da variabilidade dos padrões que caracterizam as fases do gesto em contextos de execução diferentes.

1.5 Método

Essa seção descreve, em linhas gerais, o método utilizado nesse trabalho para atingir os objetivos propostos, e analisar a hipótese definida. Este projeto se constitui como uma pesquisa experimental, que envolveu pesquisa bibliográfica, definição de protocolos para coleta de dados e para avaliação de classificadores, coleta de dados, experimentação computacional e tratamento e análise de dados.

Para a pesquisa bibliográfica foram realizados estudos exploratórios que propiciaram a obtenção de informações sobre pesquisas correlatas ao tema deste projeto. Protocolos de revisão sistemática (KITCHENHAM, 2004) foram seguidos para a realização do levantamento de trabalhos correlatos – trabalhos que visam analisar gestos mediante variações do contexto no qual eles são executados.

A definição de um protocolo de coleta de dados se fez necessária para maximizar as chances de obtenção de um ambiente de experimentação que propicie a análise das diferenças nos padrões gestuais de um mesmo gesticulador em diferentes contextos de comunicação. Esse protocolo estabelece as seguintes condições para obtenção de dados: os contextos de comunicação e gesticulação; os sensores usados para captação de dados; a restrição do espaço (3D) de movimentação do usuário; a afinidade do usuário com o contexto de comunicação; os períodos de captação de dados; o tempo de captação de dados. A coleta de dados foi realizada com o suporte do sensor *Kinect for Windows v1* da *Microsoft*, o qual possui câmera RGB para obtenção de imagens (RGB), câmera de infravermelho para obtenção da informação de profundidade, gravador para obtenção de

informação sonora, e está preparado para rastrear o corpo humano posicionado no seu campo de sensoriamento. A execução da captação de dados foi regulamentada em um projeto aprovado por um comitê de ética e registrado na Plataforma Brasil sob o número 53366516.9.0000.5390.

Após a coleta, os dados foram organizados como um *corpus* e documentados em termos de *meta-dados*, da estatística descritiva e de formas de visualização gráfica. Além disso, visto que a análise dos dados aqui proposta se baseia em uma técnica de aprendizado supervisionado, os dados foram rotulados por dois especialistas, e análise de concordância (ARTSTEIN; POESIO, 2008; ALVARES; ROMAN, 2013) foi aplicada sobre essas rotulações. A partir da existência do *corpus* rotulado, os dados foram pré-processados e representados em termos de uma representação vetorial, de forma adequada para apresentação à técnica de Máquinas de Vetores Suporte.

Um conjunto de experimentos com diferentes combinações de representação de dados, uso de janelas deslizantes e projeto de SVM (com variação de parâmetros de *kernel*) foram executados para a produção de classificadores. Na sequência, os resultados dos classificadores foram avaliados, seguindo um protocolo de avaliação de desempenho quantitativo, baseados em medidas de acurácia, revocação, precisar e *F-score*, e qualitativos, baseados em análise de erros de classificação ocorridos nas transições das fases e dentro das fases, e baseados em erros de inserção, exclusão e substituição.

1.6 Organização da dissertação

Esta dissertação está dividida em seis capítulos, considerando esta introdução, e quatro apêndices, com o seguinte conteúdo:

- No capítulo 2 são apresentadas breves explicações teóricas sobre aprendizado de máquina, máquinas de vetores suporte e avaliação de modelos classificadores em termos de acurácia e medidas derivadas da matriz de confusão;
- O capítulo 3 é construído com informações sobre a teoria dos gestos, mais especificamente sobre as fases do gesto, e informações sobre análise de concordância de rotulação e avaliação qualitativa da segmentação das fases do gesto. A análise de concordância de rotulação acompanhada da avaliação qualitativa atendem ao segundo objetivo específico estabelecido para esse trabalho.

- O capítulo 4 versa sobre o *corpus* de vídeos com gesticulação natural criado neste trabalho, atendendo ao primeiro objetivo específico delineado. O capítulo é construído com informações sobre a coleta de dados, sobre o processo de rotulação dos vídeos, e os procedimentos de pré-processamento aplicados aos vídeos para extração de características e construção de um conjunto de dados com representação vetorial para os vídeos.
- No capítulo 5 são apresentados os experimentos executados neste trabalho, aplicando SVM sobre os conjuntos de dados gerados, e aplicando as medidas quantitativas e qualitativas de avaliação da segmentação das fases do gesto implementa sob a forma da tarefa de classificação. Este capítulo conta com a discussão referente à relação do desempenho dos classificadores e os contextos de discurso.
- As considerações finais sobre o trabalho desenvolvido com um resumo das contribuições obtidas e com a sugestão de trabalhos futuros são apresentadas no capítulo 6.
- Por fim, nos apêndices são apresentados: o termo de consentimento livre e esclarecido exigido pelo comitê de ética, o roteiro de atividades dos participantes (gesticuladores) do procedimento de coleta de dados, o roteiro de perguntas usados para captação de vídeos no contexto de entrevistas, um esquema de figuras de contextos disponíveis e formas de combinação dos vídeos, e um pequeno resumo sobre o cálculo do filtro *Wavelet*, usado nos experimentos deste trabalho.

2 Aprendizado de máquina

Segundo Mitchell (1997), é pertinente definir “aprendizado de máquina” (AM), de maneira ampla e genérica, dizendo que AM inclui qualquer programa de computador que seja capaz de melhorar seu desempenho na resolução de alguma tarefa por meio do uso da experiência. Ainda o mesmo autor apresenta uma definição formal para AM: *Um programa de computador **aprende** a partir de experiências E , em relação a alguma classe de tarefas T e uma medida de desempenho P , se seu desempenho nas tarefas T , medidos por P , **melhora** com as experiências E* (MITCHELL, 1997). Nesta amplitude de definição, algoritmos tipicamente estudados e desenvolvidos nos mais diversos ramos da área de Inteligência Artificial e de áreas afins, se caracterizam como recursos para dotar uma “máquina” da capacidade de “aprender”. Insere-se neste contexto algoritmos de aprendizado sob diversos paradigmas - indutivo, dedutivo, evolucionário, por analogia ou por instrução.

Considerando o ponto de vista do trabalho proposto no presente projeto e, portanto, mais específico, o AM é aqui abordado segundo os preceitos do paradigma de aprendizado indutivo, no qual os argumentos de raciocínio são baseados na extrapolação de premissas, ou seja, conclusões gerais sobre o objeto de estudo são obtidas a partir de um conjunto de premissas que representam um conjunto particular de exemplos. Para Michalski (1983), o aprendizado indutivo é a habilidade de produzir generalizações acuradas a partir de poucos fatos ou de descobrir padrões em conjuntos de observações, aparentemente caóticas. Em suas observações publicadas nos anos 80, Michalski já defendia que o entendimento sobre esta habilidade assumiria um papel importante em termos práticos e constituir-se como chave de melhoria para métodos computacionais capazes de adquirir conhecimento. Formalmente, o AM indutivo assume um subconjunto $\mathbf{X}$, de um universo de discurso χ , formado por m vetores de dados n -dimensionais $\mathbf{x}_i$ ($i = 1, \dots, m$), associados a um conjunto $\mathbf{C}$ de k rótulos c_j ($j = 1, \dots, k$) de acordo com uma função F desconhecida. Considerando m como uma amostra representativa do universo de discurso χ onde o conjunto $\mathbf{X}$ ocorre, é possível realizar uma busca em um espaço de funções hipóteses $\mathbf{H}$ e encontrar uma função $\mathbf{F}'$ tal que $\mathbf{F}'(\mathbf{X}) \cong \mathbf{F}(\mathbf{X})$ e $\mathbf{F}'(\chi) \cong \mathbf{F}(\chi)$.

No presente trabalho, também há o interesse de estudo do comportamento de modelos classificadores no desempenho de uma tarefa (segmentação de gestos) considerando

alguns aspectos de execução dos gestos que podem, potencialmente, dificultar este tipo de estratégia de resolução de problemas. Sendo assim, assumindo o paradigma como objeto de estudo, e não necessariamente a técnica usada para implementar o paradigma, é mais relevante discutir como avaliar os resultados de um classificador do que discutir qual técnica de implementação do classificador está sendo usada. Devido a essa motivação, esse capítulo está dedicado primeiramente a descrever, de forma resumida, como pode se dar a avaliação dos resultados obtidos por um classificador dentro da resolução do problema de análises das fases do gesto; e em um segundo momento, apresentar brevemente a técnica de SVM, escolhida para implementar os classificadores que serão estudados.

2.1 Avaliação de classificadores

Tradicionalmente, uma medida de avaliação aplicada a classificadores é a acurácia. A acurácia de um classificador deve ser obtida em termos de seu erro de generalização:

$$Acuracia = 1 - \varepsilon_{generalizacao}$$

$$\varepsilon_{generalizacao}(f(x), D) = \sum_{\langle x, y \rangle \in x} D(x, y) \times \wp(f(x), y)$$

em que D é a distribuição de probabilidade que gera o conjunto de dados original e x é um exemplar do conjunto de dados usado no treinamento de uma função de perda binária

$$(\wp(f(x), y) = \begin{cases} 0 & \text{if } y = f(x) \\ 1 & \text{if } y \neq f(x) \end{cases}$$

Entretanto, para o caso de classificadores binários, é também necessário analisar a matriz de confusão e as medidas de avaliação geradas a partir dela (FAWCETT, 2006; POWERS, 2011), com destaque para a revocação, a precisão e taxa de falsos positivos, das quais derivam a medida *F-score*.

Na matriz de confusão, uma matriz quadrada de dimensões determinadas pela quantidade de classes estudada no problema de classificação, a quantidade de objetos classificados como o esperado é organizada na diagonal principal e as confusões cometidas pelo classificador – classificação de objetos na classe não esperada – são organizadas nas

posições fora da diagonal principal, de acordo com o erro cometido. As matrizes de confusão podem ser binárias (envolvendo duas classes) ou multiclases.

No caso deste trabalho, a construção da matriz de confusão precisa ser modificada visto que o problema tratado envolve múltiplas classes mas é tratado por um classificador binário dentro de uma estratégia de combinação de resultados chamada “um contra todos”. Nessa estratégia, um classificador binário é criada para resolver um subproblema binário, formado por cada uma das classes de interesse no problema original. Assim, no caso do presente trabalho, cinco classificadores são criados e suas respostas são combinadas para gerar a resposta final do problema. Por exemplo, o classificador com ativação mais forte determinada a classe escolhida para o objeto sob análise pelos cinco classificadores.

Um modelo de matriz de confusão multiclasse é apresentado na figura 1. Neste exemplo a matriz é composta pelas cinco classes estudadas no presente trabalho. A matriz é preenchida de acordo com as respostas do modelo classificador. Sobre essa matriz, a acurácia, precisão e revocação do classificador podem ser derivadas. O destaque na cor verde da matriz de confusão na tabela 1 identifica o total de acertos em cada uma das classes (TP), e sua soma sobre o total geral de amostras de objetos analisados corresponde à acurácia do classificador. A medida de precisão avalia a fração objetos classificados naquela classe que de fato é relevante; a medida de revocação avalia a fração relevante de objetos classificados para aquela classe. As medidas de precisão e revocação são ajustadas para considerar apenas falsos positivos e verdadeiros positivos diretamente relacionados a cada uma das classes.

A precisão é obtida por meio dos verdadeiros positivos e falsos positivos ($\frac{TP}{TP+FP}$), já a revocação é obtida por meio dos verdadeiros positivos e falsos negativos ($\frac{TP}{TP+FN}$). Na tabela 1, as células na cor vermelho representam as células que devem ser somadas, incluindo a célula na cor verde (total de acertos na fase de *hold*), para obter o valor de precisão para a fase de *hold*. Para a revocação, é preciso considerar as células em azul de toda a linha. Por fim, com as medidas de precisão e revocação é calculado a medida de *F-score* por meio de: $(2 * (precisão * revocação)) / (precisão + revocação)$, combinando as duas medidas por meio de uma média harmônica.

Tabela 1 – Exemplo de matriz de confusão multiclasse

		Classificado como				
		<i>D</i>	<i>P</i>	<i>H</i>	<i>S</i>	<i>R</i>
Classe real	<i>D</i>	?	?	?	?	?
	<i>P</i>	?	?	?	?	?
	<i>H</i>	?	?	23	?	?
	<i>S</i>	?	?	?	?	?
	<i>R</i>	?	?	?	?	?

Fonte: Jallysson Miranda Rocha, 2018.

2.2 Máquinas de vetores suporte

Máquinas de Vetores Suporte constituem uma classe de algoritmos de aprendizado com forte embasamento teórico pautado nos princípios da Teoria de Aprendizado Estatístico (VAPNIK, 1998). Foi inicialmente desenvolvida, no início dos anos 90, como uma generalização do algoritmo *Generalized Portrait* desenvolvido na Rússia por Vapnik, Lerner e Chervonenkis (VAPNIK; LERNER, 1963). Desde então, diferentes tipos de SVM foram desenvolvidas, usando diferentes estratégias de otimização, e para diferentes tarefas (classificação, regressão, agrupamento), como discutido por Suykens et al. (2002), Kwok (1998) e Schölkopf e Smola (2002). Em princípio, SVM tem como base o objetivo de maximizar a capacidade de generalização e evitar a ocorrência do fenômeno de sobre-ajuste na aproximação de uma superfície de decisão para um problema de aprendizado indutivo. Para isso, em contraposição a outras abordagens como as Redes Neurais Artificiais, SVM trabalha minimizando o **risco estrutural** ao invés de minimizar apenas o risco empírico. Assim, SVM trata o problema de encontrar um modelo que minimize o risco empírico e que pertença a uma classe de funções com baixa dimensão Vapnik-Chervonenkis (VC) (LORENA; CARVALHO, 2007).

As SVMs apresentam resultados eficazes em classificar dados linearmente separáveis ou próximos de serem separáveis linearmente (LORENA; CARVALHO, 2007). Na figura 4 é separado duas classes diferentes, com o melhor hiperplano de separação entre eles, maximizando a margem de separação entre as duas classes.

Mas, na maioria dos casos não é possível encontrar um hiperplano que divida satisfatoriamente os dados linearmente (LORENA; CARVALHO, 2007). Um exemplo é apresentado na figura 5, em que o uso de uma fronteira curva seria mais adequada na separação das classes (LORENA; CARVALHO, 2007). Então, neste caso, SVM é

Figura 4 – Classificação de dados linearmente separáveis com SVM

Fonte: Traduzido de Meyer e Wien (2001)

baseado na execução de um mapeamento não-linear de vetores de entrada do espaço de características original para um espaço de características de alta dimensão, e na otimização de um hiperplano capaz de separar dados no espaço de características de alta dimensão (VAPNIK, 1998).

Figura 5 – Representação de um conjunto de dados não linear. Legenda: (a) - conjunto de dados não linear; (b) - fronteira não linear no espaço de entradas; (c) fronteira linear no espaço de características

Fonte: Lorena e Carvalho (2007)

Considerando um conjunto de treinamento com N exemplos, definidos por $\{\mathbf{x}_i, y_i\}_{i=1}^n$ onde $\mathbf{x}_i$ é uma entrada, y_i é uma saída, e $y_i \in \{-1, +1\}$ a meta do SVM é encontrar um hiperplano de separação ótimo, o qual é dado por $h(\mathbf{x}_i) = \mathbf{w}^T \varphi(\mathbf{x}_i) + b$, onde $\mathbf{w}$ é o conjunto ótimo de pesos, b é o bias ótimo, e φ é o mapeamento não linear aplicado nas

entradas. SVM otimiza o hiperplano maximizando as distâncias entre este hiperplano e os dados mais próximos ($\mathbf{x}_i$), que corresponde a minimizar $\mathbf{w}$ usando, por exemplo¹:

$$\min \phi(\mathbf{w}, \mathbf{b}, \boldsymbol{\xi}) = \frac{1}{2} \mathbf{w}^T \mathbf{w} + C \sum_{i=1}^N \xi_i, \quad (1)$$

onde C é um fator de regularização, e ξ é um fator de erro, sujeito a

$$y_i(\mathbf{w}^T \phi(\mathbf{x}_i) + b) \geq 1 - \xi, \quad i = 1, \dots, N;$$

$$\xi_i \geq 0, \quad i = 1, \dots, N.$$

Aplicando o método Lagrangeano na equação 1, obtém-se

$$\max (\mathbf{L}_1(\boldsymbol{\alpha})) = \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i,j=1}^N \alpha_i \alpha_j y_i y_j \langle \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j) \rangle, \quad (2)$$

sujeito a

$$\sum_{i=1}^N \alpha_i y_i = 0,$$

$$C \geq \alpha_i \geq 0, \quad i = 1, 2, \dots, N,$$

onde $\boldsymbol{\alpha}$ são os multiplicadores de Lagrange.

Resolvendo o problema da equação 2 é possível resolver o problema na equação 1, desde que $\mathbf{w}$ possa ser definido em termos de $\boldsymbol{\alpha}$ (HAYKIN, 1999). Na equação 2, uma função *kernel* $K(\mathbf{x}_i, \mathbf{x}_j)$ pode ser usada para representar o produto interno $\langle \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j) \rangle$, executando um mapeamento não linear implícito.

2.3 Considerações finais

Este capítulo apresentou: uma breve introdução sobre conceitos de aprendizado de máquina e do paradigma indutivo, foco desse projeto; uma apresentação das avaliações aplicadas a modelos classificadores, com uso de matriz de confusão, com destaque as medidas derivadas a partir dela; uma apresentação da técnica de Máquinas de Vetores Suporte que é aplicada a este trabalho.

¹ Considerando uma otimização de margem *soft* usando uma 1-norma.

3 Análise automatizada das fases do gesto

Este capítulo apresenta resumidamente a área de pesquisa em Análise de Gestos, abordando a teoria geral de gestos com destaque para os tipos de gestos e fases dos gesto, e abordando aspectos do segmentação das fases do gesto manual e automatizada. O capítulo é finalizado com um resumo de iniciativas correlatas à pesquisa apresentada neste trabalho.

3.1 Teoria de gestos

A área de aplicação do presente projeto está diretamente relacionada à Análise de Comportamento Humano. Do ponto de vista da Ciência da Computação, esta área de aplicação é principalmente estudada, como discutido por Pantic et al. (2007), como uma abordagem que se preocupa com métodos de interação humano computador, baseados no comportamento do ser humano. Segundo o mesmo autor, ao modelar sistemas com tais formas de interação, é necessário ter em mente o que é comunicado (qual é o tipo de mensagem utilizada: mensagem linguística, sinal não linguístico, emoção, atitude), como a informação é passada (expressão facial, gesto, movimentos de cabeça) e por que a informação é passada (o contexto da ação: onde o usuário está, o que está fazendo, se há outras pessoas envolvidas).

Há também o interesse em estudar características referentes à comunicação via análise de comportamento humano que está localizada em outras áreas de pesquisa: na Linguística e na Psicolinguística. Nessas áreas é desenvolvida uma maneira singular de estudo que é principalmente pautado na análise de padrões espaço-temporais gerados pela realização de “movimentos” e “gestos”¹ e pela inserção e análise da significação destas variáveis no contexto comportamental (MCCLEARY; VIOTTI, 2009).

Diferentes estudos que revisam o estado da arte em análise de gestos trazem as diferentes facetas e desafios da área. Como forma de acessar um espectro bastante abrangente sobre esta área, é interessante considerar algumas revisões que têm sido publicadas na área. Do ano de 2005 até o ano de 2017, foram publicadas algumas revisões de literatura em análise de gestos. Juntas, essas revisões cobrem um conjunto de estudos

¹ Movimentos (corporais), embora possam fazer parte de um gesto, não podem ser considerados gestos em si. Porém, na linha de pesquisa que está sendo seguida neste projeto, um gesto sempre envolve um movimento. Dessa forma, no presente texto, o conceito de gestos será usado de maneira a englobar o conceito de movimento e, daqui para frente, seguir-se-á falando em análise de gestos apenas.

que ou focam em aplicações específicas ou na proposição de soluções genéricas para problemas de análise de gestos. Mitra e Acharya (2007) cobrem estudos em reconhecimento de gestos de mão, de braços e de cabeça, focando nos métodos e ferramentas usadas; enquanto Pickering (2005) faz uma análise mais específica sobre reconhecimento de gestos para interação entre humanos e veículos; Moni e Ali (2009) apresentam uma revisão cobrindo o uso de Modelos Escondidos de *Markov* no problema de reconhecimento de gestos manuais. Outro ponto de vista é considerado por Sowa (2008) e Liu e Kavakli (2010). O primeiro cobre estudos referentes a tecnologias de sensoriamento, segmentação de gestos e reconhecimento de gestos, e seus relacionamentos com reconhecimento de fala. O segundo apresenta uma revisão sobre interação multimodal em jogos de computador. Madeo, Wagner e Peres (2013) apresentam uma revisão de literatura sistemática sobre aspectos temporais e análise de gestos manuais, com um foco específico em análise da conversação natural. Esses autores organizam as informações dos estudos incluídos na revisão considerando: tipo de análise, métodos (com destaque para métodos de Aprendizado de Máquina) e aplicações. Mais recentemente, Madeo, Lima e Peres (2017) publicaram uma revisão sobre análise automatizada de gestos manuais com foco na análise de iniciativas de reconhecimento e segmentação de fases do gesto e de tipos funcionais de gestos.

3.1.1 Fases do gesto

O contexto do presente projeto é a área de análise linguística e psicolinguística, e, portanto, é necessário discutir, ainda que brevemente, alguns tópicos da teoria desta área de estudo. De acordo com Garnham (1994), Psicolinguística é o estudo de mecanismos mentais que dão embasamento à produção e compreensão da linguagem. No estudo da linguagem, para McNeill (1992), assim como para Kendon (1980), é também necessário considerar os gestos, já que eles são parte integrante da linguagem. Além disso, esses autores argumentam que gestos e fala são parte de um mesmo sistema linguístico, oriundos da mesma fonte semântica. Também Quek et al. (2002) reforçam a ideia de que estas modalidades não são redundantes, mas sim coexpressivas: elas compartilham uma fonte semântica mas expressam informação diferente. O mesmo autor apresenta um exemplo: enquanto a fala expressa que alguém abriu uma porta, um gesto de duas mãos, não simétrico, mostra que tipo de porta foi aberta. Algumas linhas de pesquisa dentro do contexto de estudo da

linguagem têm como objetivo estudar a análise da gesticulação (gestos que acompanham a fala) reconhecendo tipos especiais de gestos considerando sua estrutura temporal (WILSON; BOBICK; CASSELL, 1996) ou considerando seu conteúdo semântico (KETTEBEKOV et al., 2002; KETTEBEKOV, 2004; KETTEBEKOV; YEASIN; SHARMA, 2005). Alguns estudos, como os apresentados por Kendon (1980) e por Kita, Gijn e Hulst (1998) são baseados no estudo da divisão dos gestos em fases.

As fases do gesto estão diretamente ligadas ao que Kendon (2005) chamou de “movement excursions”. Segundo o mesmo autor, uma pessoa faz uma ou várias “movement excursions” durante um discurso. Estas excursões se referem a movimentos das mãos de alguma “posição de descanso” para uma região no espaço onde um movimento “importante” ocorre, e depois, de volta para a posição de descanso. Toda a excursão (ou trajetória) do movimento é chamada de “unidade gestual”, enquanto as posições que as mãos assumem entre elas são chamadas de “posições de descanso”. Uma unidade gestual pode ser segmentada em fases do gesto, que podem ser: preparação, na qual a mão se move para a posição onde o conteúdo do gesto será expressado; *pré-stroke hold*, a qual é uma breve pausa no fim da fase de preparação; *stroke*, que contém o pico de esforço do gesto e expressa seu conteúdo semântico; *pos-stroke hold*, que é uma breve pausa no fim de um *stroke*; retração, na qual a mão retorna para a posição de descanso. Em Kita, Gijn e Hulst (1998), é também sugerido que as fases *pré-stroke*, *stroke* e *pos stroke hold* compõem uma fase expressiva do gesto, a qual também pode ser composta por um único *hold* independente, isto é, uma pausa que expressa o conteúdo semântico do gesto. Finalmente, tanto Kita, Gijn e Hulst (1998) quanto Kendon (2005) consideram o conceito de frase gestual. Contudo, desde que não existe um consenso sobre este conceito², ele não é usualmente aplicado nas tarefas de segmentação de gestos.

3.1.2 Tipos de gestos

Embora os tipos de gestos não sejam objeto de estudo deste projeto, é interessante explicá-los de forma que se possa dar mais subsídios para o entendimento das fases do gesto. Isso por que é comum esperar que uma análise de gestos se preocupe com aspectos

² Kita, Gijn e Hulst (1998) defende que a frase gestual corresponde a todas as fases que permeiam uma única fase expressiva (o que incluiria a retração), enquanto Kendon (2005) define a frase gestual como um segmento contendo preparação e fase expressiva apenas.

referentes à forma ou ritmo, mas, como visto aqui, esse tipo de aspecto é relacionado a parte da teoria de gestos referente apenas aos tipos de gestos. McNeill (2008) e Gibbon (2009) apresentam, cada um, uma taxonomia para o estudo dos gestos. O primeiro, define cinco principais categorias para os gestos:

- icônicos: associados a objetos ou eventos concretos, e sua forma é bem similar a uma imagem relacionado com o conceito que eles representam;
- metafóricos: similar aos gestos icônicos, mas representam ideias abstratas;
- *beats*: marcam o ritmo da fala, não carregam conteúdo semântico e seu valor semiótico é revelar o que é importante na narrativa do discurso. *Beats* também são importantes porque possuem sempre duas fases de movimentos básicas – para dentro/para fora; para cima/para baixo; para esquerda/para direita; etc;
- coesivos: unem duas partes do discurso que são tematicamente relacionadas, mas temporalmente separadas. São usados para retomar algum tópico depois de um comentário, por exemplo;
- dêiticos: também são conhecidos como “gestos para apontar” e são usados para indicar objetos ou eventos concretos, ou ainda para localizar referências para conceitos abstratos no espaço gestual.

Gibbon (2009), defini as seguintes categorias que podem complementar a lista de McNeill (2008):

- emblemáticos: gestos que são vistos como convenções, como o gesto que indica despedida, ou que são específicos de uma linguagem ou cultura;
- afetivos: gestos que revelam estados e eventos emocionais.

3.2 Segmentação das fases do gesto

A segmentação das fases do gesto consiste em encontrar as fases do gesto **preparação**, **stroke**, **hold** e **retração**) considerando uma instância de gesticulação (um trecho de um discurso ou da conversação natural, por exemplo). Tradicionalmente, essa tarefa é resolvida de forma manual por especialistas em Linguística. Sob um ponto de vista mais operacional, a segmentação das fases do gestos é uma ação na qual dada uma sequência de quadros de um vídeo, um especialista analisa cada quadro e os classifica

como constituinte de uma das fases do gesto, que compõem a unidade gestual (cf. exemplo na figura 6). Trata-se portanto de um trabalho que demanda muito tempo. Um simples trecho de um vídeo com cinco minutos de gesticulação pode requerer cerca de oito horas de trabalho de um especialista, pois necessita de atenção especialmente nos quadros de transição, momento em que ocorre a mudança de fase. Geralmente essa mudança é difícil de ser determinar em qual quadro começa, e muitas vezes torna-se um processo subjetivo determinar essa transição.

Figura 6 – Segmentação das fases do gesto

Fonte: Jallysson Miranda Rocha, 2018

Para automatizar essa tarefa, pesquisadores como Ramakrishnan (2011), Ramakrishnan e Neff (2013), Madeo (2013) têm aplicado algoritmos de Aprendizado de Máquina, modelando o problema de segmentação das fases do gesto como um problema de classificação. Assim, resolver esse problema consiste em construir um modelo classificador que recebe a representação vetorial do gesto, a qual consiste em uma série de medições obtidas de um ou mais quadros de um vídeo (quando uma abordagem baseada em representação por janelas deslizantes é aplicada), e o classifica como pertencente a uma classe dentre as cinco classes³ esperadas (as fases do gesto).

3.2.1 Avaliação da segmentação das fases

Neste projeto, a segmentação automática das fases do gesto é obtida por meio da análise de uma sequência de quadros $\{f_1 \cdots f_n\}$, consecutivos, correspondente de um vídeo. Esta sequência contém as imagens de uma pessoa em um contexto de discurso (descrevendo uma figura, narrando uma história ou participando de uma entrevista, por exemplo). O problema de segmentação das fases dos gestos, então, pode ser modelado como uma tarefa de classificação na qual, dada a entrada de um quadro f_i de uma sequência de quadros s , o

³ Alguns pesquisadores, como Madeo (2013), consideram no momento da rotulação a classe de descanso como uma fase do gesto, na segmentação automática das fases do gesto

modelo classificador deve classificar esse quadro em uma das classes $c_i = \{D, P, H, S, R\}$, que correspondem respectivamente ao período de descanso e às fases de preparação, *hold*, *stroke* e retração, como definido no capítulo 1.

No caso específico da tarefa de análise de fases de gesto, é interessante avaliar o desempenho do classificador sob o ponto de vista de um especialista que usará o resultado produzido pelo algoritmo. A tarefa de segmentação das fases do gesto com base na informação de um vídeo, ou de seus quadros constituintes, torna a análise do gesto uma atividade de alta precisão. É fato que diferentes segmentadores humanos, ao segmentar um mesmo vídeo, gerarão resultados levemente diferentes, sendo que a maior parte das diferenças se dá nas transições entre fases. Assim, como a abordagem aqui adotada está pautada na análise do vídeo quadro a quadro e se trata de uma abordagem baseada em aprendizado supervisionado, medidas para análise de desempenho de classificadores deveriam considerar análise de qualidade de segmentação, que representa uma avaliação mais próxima sob o ponto de vista de um especialista (e.g. um linguísta). Para a avaliação final dos modelos, é considerado a resolução do problema de classificação multiclasse, visto no capítulo 2, que combina os resultados obtidos em cada subproblema de classificação binária.

Porém, para a avaliação da qualidade de segmentação, diferentes critérios de avaliação de erro podem ser seguidos. O principal erro a ser considerado é o *erro interno* (**EI**), ou seja, erros de classificação cometidos dentro de uma fase do gesto ou dentro de um período de descanso (cada quadro classificado erroneamente está dentro de uma das fases). Em contraposição a esse critério está a análise dos *erros de transição* (**ET**), que são erros cometidos na classificação de quadros que estão em regiões de transição entre fase. Esses erros são similares à discordância de rotulação entre dois ou mais especialistas (veja discussão na próxima seção), e possuem uma criticidade menor do que os outros erros. Os demais critérios usados neste trabalho para avaliação da segmentação são:

- *Erros de inserção* (**INS**), ocorrem quando a resposta final dos classificadores classifica uma sequência de quadros de um período de descanso como quadros pertencentes a uma das fases do gesto;
- *Erros de exclusão* (**EXC**), semelhante aos erros de inserção, porém ocorrem quando uma fase do gesto não é reconhecida pelo classificador, ou seja uma sequência de

quadros é classificada como um período de descanso incorretamente, quando todos os quadros pertencem a alguma fase do gesto;

- *Erros de substituição (SUB)*, os classificadores classificam incorretamente as fases do gesto, toda a sequência de quadros de uma fase do gesto é substituída por outra fase.

Na figura 7 é mostrado um exemplo de uma sequência de quadros com a identificação de cada um desses erros.

Figura 7 – Erros de segmentação no problema de segmentação das fases do gesto

Fonte: Jallysson Miranda Rocha, 2018.

É interessante destacar que todos os erros são erros internos, e um segmento de erros internos representa outros erros. Na figura 7, dentro do erro de transição (**ET**) não ocorrem erros internos, e sim um erro de transição.

3.2.2 Avaliação da concordância de rotulação

Além de avaliar os classificadores por meio das medidas obtidas da matriz de confusão e por meio dos critérios de qualidade de segmentação, também é necessário minimizar o impacto da subjetividade na rotulação realizada por diferentes especialistas sobre o conjunto de dados a ser usado para treinamento e teste dos classificadores. O critério de avaliação para essa questão foi a aplicação do coeficiente *Krippendorff's Alpha* (KRIPPENDORFF, 2007). Esse coeficiente permite medir o grau de confiabilidade na concordância de rotulação de todos os especialistas e é obtido por:

$$\alpha = 1 - \frac{D_o}{D_e},$$

em que D_o é a quantidade de desacordos observados, e D_e é a quantidade de acordos esperados. Em detalhes, $D_o : \frac{1}{n} \sum_c \sum_k O_{ck} \delta_{ck}^2$ e $D_e : \frac{1}{n(n-1)} \sum_c \sum_k n_c * n_k \delta_{ck}^2$, em que O_{ck} , n_c , n_k e n são as frequências de valores em uma matriz de coincidências; δ é uma função

que implementa uma métrica adequada para diferentes tipos de dados (KRIPPENDORFF, 2007). As métricas que podem ser implementadas em δ , são:

- nominal, $\delta_{ck}^2 = \begin{cases} 0 & \text{if } c=k, \\ 1 & \text{if } c \neq k, \end{cases}$
- ordinal, $\delta_{ck}^2 = (\sum_{g=c}^{g=k} n_g - \frac{n_c+n_k}{2})^2$;
- interval, $\delta_{ck}^2 = (c - k)^2$;
- ratio, $\delta_{ck}^2 = (\frac{c-k}{c+k})^2$

A saída final do coeficiente pode possuir valores diferentes a depender da métrica utilizada, que é relacionada ao tipo dos dados em uso. A saída é chamada de K tem seus valores variando de menor ou igual a 0,0 até 1,0, conforme a figura 8. Nessa figura também estão mostrados os intervalos em que K é dividido para fins de qualificação de um conjunto de rotulações. As qualificações vão de *pobre*, quando o valor de K é menor ou igual a 0,0 a *perfeito* quando K está no melhor intervalo, variando de 0,8 a 1,0. Conhecer esses intervalos é importante para definir a qualidade da rotulação entre dois ou mais especialistas, medindo o nível de concordância entre eles.

Figura 8 – Intervalo de qualidade dos valores de K do coeficiente *Krippendorff's Alpha*

Fonte: Traduzido de Artstein e Poesio (2008)

3.3 Trabalhos correlatos

Além dos estudos de Ramakrishnan (2011), Ramakrishnan e Neff (2013), Madeo (2013), que tratam diretamente da área de segmentação das fases do gesto, de forma automática com uso de algoritmos de aprendizado de máquina, foram encontrados outros estudos que apresentam iniciativas relacionados ao estudo do contexto de gesticulação. As formas como cada um desses estudos usam informação de contexto na análise do gesto varia e alguns exemplos são:

- uso de pistas contextuais que podem ajudar a reconhecer o gesto em execução em uma cena;

- análise da interação do gesticulador com objetos em cena (ex. segurar um caderno ou segurar um controle remoto);
- observação da interação de um locutor (que está gesticulando) e os ouvintes do discurso, observando principalmente ações do gesticulador tais como virar a cabeça na direção do ouvinte ou olhar diretamente para o ouvinte).

Na tabela 2 estão organizadas algumas informações sobre os trabalhos correlatos. Em negrito estão aqueles que mais se aproximam do assunto tratado no presente trabalho. Trata-se de estudos que usaram dados provenientes da tarefa de comunicação entre pessoas, desenvolveram métodos para classificação de fases do gesto e usaram os contextos de discurso de conversação e entrevista.

Tabela 2 – Trabalhos correlatos x aplicação

Referência	Aplicação
(KUROKI et al., 2015)	Apoio a surdos-mudos
(LU et al., 2014)	Interação com dispositivos móveis
(JACOB; WACHS, 2014)	Operação cirúrgica
(BHAME; SREEMATHY; DHUMAL, 2014)	IHC
(SHEN et al., 2014)	IHC
(DARDAS; SILVA; SADDIK, 2014)	Jogos de esforço
(OKADA et al., 2013)	Comunicação
(MATSUBARA; MORIMOTO, 2013)	Robótica
(LIANG; MINGZHI; YING, 2013)	IHC
(HUSSAIN; RASHID, 2012)	IHC
(ZHANG et al., 2011)	IHC
(REFICE et al., 2011)	Comunicação
(MOLINA et al., 2013)	Controle de aplicações
(ZHU; SHENG, 2011)	Apoio a idosos e deficientes
(DENG et al., 2011)	IHC
(PAULSON; CUMMINGS; HAMMOND, 2011)	Interação com objetos
(MARASOVIĆ; PAPIĆ, 2014)	Interação com dispositivos móveis

Fonte: Jallysson Miranda Rocha, 2018

3.4 Considerações finais

Neste capítulo o problema de segmentação das fases do gesto foi brevemente discutido dentro do contexto da área de Teoria dos Gestos. Critérios qualitativos de avaliação da segmentação sob o ponto de vista do especialista da área de análise de gestos, assim como a sugestão do uso de um coeficiente de concordância para avaliação do trabalho de rotulação de vídeos a ser usado como conjunto de dados de um classificador, foram

apresentados. A aplicação dos critérios qualitativos é proposta neste trabalho como ações de um protocolo de avaliação de desempenho dos modelos classificadores em complemento ao uso das medidas derivadas da matriz de confusão (comumente usada na área de aprendizado de máquina).

4 *Corpus* de vídeos com gesticulação natural

Este capítulo descreve o processo de construção do *corpus* de vídeos com gesticulação natural e o *corpus* em si. A construção desse *corpus* atende ao primeiro objetivo específico definido nesta dissertação de mestrado e seguiu o projeto aprovado pelo Comitê de Ética em Pesquisa em Seres Humanos (CEP) da Escola de Artes, Ciências e Humanidades da Universidade de São Paulo (EACH/USP), com Certificado de Apresentação para Apreciação Ética (CAAE) número 53366516.9.0000.5390.

O conteúdo deste capítulo está organizado da seguinte maneira: na seção 4.1, um resumo das principais características do projeto desenvolvido para a coleta dos dados e análise pelo CEP é apresentado; na seção 4.2, é apresentado o ambiente de captura dos dados incluindo informações sobre os gesticuladores e sobre os contextos; a seção 4.3 apresenta o processo de rotulação realizado sobre os vídeos; na seção 4.4 são apresentadas a codificação da informação gestual capturada pelo sensor, e as ações de pré-processamento para extração de características e adequação dos dados ao uso em classificadores que trabalham em um espaço vetorial; e por fim, na seção 4.5 é apresentado um resumo sobre o conteúdo organizado no *corpus* de vídeo e o que foi preparado como conjunto de dados, derivado do *corpus* de vídeo, para uso nos experimentos no capítulo 5.

4.1 *Projeto no comitê de ética em pesquisa*

Para a coleta dos dados referentes à gesticulação natural foi necessário apresentar um projeto ao CEP da EACH/USP pois se tratava da obtenção de imagens (vídeos) e áudio de seres humanos. Neste projeto, foram estabelecidos protocolos que possibilitaram criar, organizar, documentar e codificar o *corpus* de vídeo com gesticulação natural destinado a suportar estudos em análise de gesto.

De forma resumida, no protocolo de criação do *corpus* foram estabelecidas as seguintes condições:

- os contextos de comunicação e gesticulação - do tipo monólogo (contação de histórias, improvisação e descrição), do tipo diálogo (entrevista e diálogo) e do tipo exposição oral (aula);

- os sensores usados para captação dos vídeos e seus respectivos posicionamentos – câmera RGB, câmera de infravermelho e gravador de áudio;
- a restrição do espaço (3D) de movimentação do usuário e seu posicionamento – o usuário pode estar em pé ou sentado, com restrições de movimentação de acordo com o campo de rastreamento das câmeras;
- a afinidade do usuário com o contexto de comunicação – o gesticulador pode escolher o contexto de participação, se possível possibilitando a captação de dados em pelo menos dois contextos diferentes;
- os períodos e sessões de captação de dados – o gesticulador deve possibilitar a captação em pelo menos duas sessões por contexto;
- o tempo de captação de vídeos – a captação de cada vídeo (uma sessão em um contexto) deve durar de um a cinco minutos.

Para organização e documentação do *corpus* foi estabelecido um protocolo que determinou os meta-dados (data e hora de captura, sensores utilizados, tempo de captura, quantidade de quadros no vídeo, contexto associado, sessão associada, e idade, gênero e grau de instrução do gesticulador) que acompanham os vídeos e o uso de uma codificação numérica para descrição do conteúdo de cada vídeo. Na figura 9, é mostrado um esquema das atividades que um gesticulador pode realizar em uma sessão de captação de vídeo. Apenas um gesticulador participa de uma sessão e cada sessão pode abranger vários contextos. Os contextos podem se repetir em uma mesma sessão ou em sessões diferentes. O protocolo estabelece também que pequenos trechos com quadros (imagens) do vídeo podem ser usados para fins de documentação desde que o rosto do usuário não seja exposto.

Figura 9 – Esquema das atividades de um gesticulador: durante uma sessão, um gesticulador participa da captação de vídeos para um ou mais contextos. Nesse exemplo, *A*, *B*, *C*, *D* dizem respeito a quatro contextos diferentes

Fonte: Jallysson Miranda Rocha, 2018.

Por fim, um terceiro protocolo foi estabelecido para determinar a codificação da informação de interesse em um vídeo. Foi estabelecido que os gestos devem ser codificados por meio das coordenadas espaciais (x , y e z) de pontos de articulação do corpo do gesticulador, como ilustrado na figura 10. Sendo assim, o conteúdo dos vídeos são mapeados para matrizes numéricas nas quais cada linha corresponde a um quadro de vídeo e as colunas correspondem às coordenadas espaciais dos pontos do corpo do gesticulador. Cada linha da matriz conta ainda com um identificador e com rótulos associados manualmente (cf. seção 4.3).

Figura 10 – Pontos de articulação no corpo do gesticulador: mãos, pulsos, cotovelos, tronco e cabeça

Fonte: Jallysson Miranda Rocha, 2018.

O objetivo das regras estabelecidas nos protocolos de organização e de codificação é manter total segurança e sigilo em relação à identidade dos gesticuladores quando da publicação do *corpus*.

4.2 Execução da captação de dados

Inicialmente, foi criado um roteiro de atividades (conforme o apêndice C) para guiar a participação de cada gesticulador. A cada sessão, a ordem de trabalho com os contextos variou de forma a aumentar a variabilidade de contextos captados com usuários diferentes¹. As ações realizadas durante a captura foram, nesta ordem:

1. a apresentação das informações básicas sobre a coleta de dados para o participante (gesticulador), incluindo o objetivo geral da captação dos vídeos (análise de gesticulação) e o sensor de captação (*Microsoft Kinect* – figura 11);

¹ O contexto a ser trabalhado foi apresentado ao gesticulador como uma sugestão. A partir de sua anuência, procedeu-se com a captação.

Figura 11 – Sensor *Microsoft Kinect* e seus componentes

Fonte: Traduzido de *Microsoft Developer Network*

2. autorização da captação dos vídeos por meio da assinatura da autorização para captura de um Termo de Consentimento Livre e Esclarecido (TCLE), apresentado no Apêndice A;
3. orientação sobre a posição do gesticulador em frente ao sensor e sobre informações referentes ao contexto de gesticulação;
4. realização das atividades de gesticulação conforme definido no roteiro de atividades.

A captação dos vídeos foi feita por dois pesquisadores, sendo o principal o autor deste trabalho. O segundo pesquisador, a orientadora deste trabalho, forneceu auxílio durante as primeiras sessões de captação. As sessões de captação ocorreram entre 18/10/2016 a 03/12/2016 e contou com a participação de sete gesticuladores. Durante esse período de tempo foi possível captar vídeos para os seguintes contextos de gesticulação:

- C1** Entrevista – composta por três blocos de perguntas organizados em dois momentos durante a captura. No primeiro momento, a entrevista diz respeito aos blocos de perguntas 1 e 2; no segundo momento, a entrevista diz respeito aos blocos de perguntas 2 e 3. As perguntas foram definidas no protocolo do projeto analisado pelo CEP e podem ser analisadas no apêndice C.
- C2** Improvisação – contexto no qual um de quatro temas é sugerido ao gesticulador para que ele improvise um discurso. Os temas são: brincadeira de criança, situação inusitada em uma viagem, cena de um filme e situação inusitada no ambiente de trabalho.
- C3** Contação de história – contexto no qual uma de três histórias em quadrinhos (tirinhas) é apresentada ao gesticulador para que ele se baseie e conte uma história.

C4 Descrição – contexto no qual uma de três imagens contendo uma cena de um filme, uma cena de um jogo de *vídeo game* ou uma pintura (objeto de arte) é apresentada ao gesticulador para que ele descreva a imagem apresentada.

C5 Aula – captura realizada durante a aula de professores universitários.

O *corpus* de vídeo final conta com 89 vídeos e estão distribuídos em termos de contextos, usuários e sessões conforme listado na tabela 3. Em média foram captados 21 vídeos por contexto, com exceção do contexto C5, para o qual apenas 4 vídeos foram obtidos. Todas as captações foram feitas com um sensor *Microsoft Kinect* configurado para extrair as coordenadas espaciais dos pontos de articulação. Essas informações são então organizadas em arquivos (persistência das matrizes numéricas).

Tabela 3 – Distribuição dos vídeos que compõem o *corpus* por gesticulador G_i e por contexto C_j

Gesticulador	Idade	Gênero	Grau de instrução	C1	C2	C3	C4	C5	# vídeos
G1	42	F	Pós-graduação	2	4	4	2	2	14
G2	34	M	Pós-graduação	4	2	4	4	0	14
G3	39	M	Pós-graduação	4	3	2	5	2	16
G4	24	M	Graduação	2	4	4	2	0	12
G5	30	M	Pós-graduação	2	1	2	2	0	7
G6	39	M	Pós-graduação	4	4	2	4	0	14
G7	29	F	Pós-graduação	2	4	4	2	0	12
# vídeos				20	22	22	21	4	89

Fonte: Jallysson Miranda Rocha, 2018.

4.3 Rotulação dos vídeos

Após a captação dos vídeos e organização da informação na forma de matrizes numéricas, é realizada a etapa de rotulação. A rotulação é um processo realizado manualmente por um especialista, por meio da análise da sequência de quadros dos vídeos. Durante a análise, o especialista atribui (rotula) a cada quadro de vídeo, a fase do gesto que está ocorrendo naquele “momento” de gesticulação. A rotulação pode ser realizada de duas formas: observando o comportamento das duas mãos do gesticulador, que é a maneira tradicional; ou analisando o comportamento das mãos individualmente e, neste caso, gerando dois rótulos por quadro.

Na figura 12 são ilustradas três situações de rotulação: (a) o comportamento das duas mãos definem uma fase de *stroke*; (b) e (c) o comportamento de cada uma das mãos define uma fase diferente do gesto, sendo que no caso (b) a mão esquerda do gesticulador define a fase de *stroke* e a mão direita está em posição de descanso, e no caso (c) ocorre o contrário². Para os casos (b) e (c), considerando a rotulação tradicional, o especialista precisa adotar uma linha de rotulação, por exemplo, assumindo a fase mais expressiva (*stroke* é uma fase mais expressiva do que o período de descanso e as fases de preparação ou retração) ou assumindo o comportamento da mão dominante do gesticulador. Ainda, se ambas as mãos expressam o comportamento de uma mesma fase do gesto, então os dois rótulos são iguais. Um exemplo ilustrativo da organização da rotulação dos quadros de vídeos é apresentada na tabela 4.

Figura 12 – Exemplos para análise de rotulação considerando ambas as mãos e gerando um rótulo, ou considerando cada uma das mãos, gerando dois rótulos

Fonte: Jallysson Miranda Rocha, 2018.

Neste trabalho, a rotulação observando o comportamento das duas mãos foi escolhida para uso, e dois especialistas trabalharam nesse processo produzindo duas rotulações para

² Alternativamente, um especialista pode determinar que a mão direita, neste caso, está em fase de *hold*. Essa interpretação não foi adotada na rotulação realizada neste trabalho.

Tabela 4 – Ilustração da organização das rotulações junto aos quadros de vídeo. Essa ilustração segue o exemplo de quadros de vídeo apresentado na figura 12

x1	y1		Selo de tempo	Rótulo duas mãos	Rótulo mão direita	Rótulo mão esquerda
		descriptores				
	...		...456	<i>Stroke</i>	<i>Stroke</i>	<i>Stroke</i>
	...		...457	<i>Stroke</i>	<i>Stroke</i>	<i>Stroke</i>
	...		...458	<i>Stroke</i>	<i>Stroke</i>	<i>Stroke</i>
	...		...459	<i>Stroke</i>	<i>Stroke</i>	<i>Stroke</i>
	...		...460	<i>Stroke</i>	<i>Stroke</i>	<i>Stroke</i>
	...		...764	<i>Stroke</i>	<i>Descanso</i>	<i>Stroke</i>
	...		...765	<i>Stroke</i>	<i>Descanso</i>	<i>Stroke</i>
	...		...766	<i>Stroke</i>	<i>Descanso</i>	<i>Stroke</i>
	...		...767	<i>Stroke</i>	<i>Descanso</i>	<i>Stroke</i>
	...		...768	<i>Stroke</i>	<i>Descanso</i>	<i>Stroke</i>
	...		...777	<i>Stroke</i>	<i>Stroke</i>	<i>Descanso</i>
	...		...778	<i>Stroke</i>	<i>Stroke</i>	<i>Descanso</i>
	...		...779	<i>Stroke</i>	<i>Stroke</i>	<i>Descanso</i>
	...		...780	<i>Stroke</i>	<i>Stroke</i>	<i>Descanso</i>
	...		...781	<i>Stroke</i>	<i>Stroke</i>	<i>Descanso</i>

Fonte: Jallysson Miranda Rocha, 2018.

um mesmo vídeo em alguns casos. Embora cinco vídeos tenham sido rotulados usando a observação do comportamento das mãos em separado, realizar esse trabalho é bastante custoso e foi dado prioridade para a rotulação tradicional. Uma rotulação de um vídeo com cerca de 4.000 quadro pode tomar de três a quatro horas de trabalho do especialista. Na tabela 5³ estão listados os vídeos do *corpus* que foram rotulados e por quais especialistas.

Tabela 5 – Rotulações (Rot) e Coeficientes de Concordância *Krippendorff's Alpha* (CA) para oito vídeos. Legenda: G - gesticulador; E1 e E2 – especialistas; Rot1 - rotulação observando a mão direita; Rot2 – rotulação observando a mão esquerda; Rot3 - rotulação observando as duas mãos

G	Contexto	Sessão	Rot1	CA	Rot2	CA	Rot3	CA
G1	C1.1	1	E1	–	E1	–	E1	–
G1	C2.1	1	–	–	–	–	E1 e E2	0,80
G1	C3.2	1	E1 e E2	0,80	E1 e E2	0,76	E1 e E2	0,75
G1	C3.3	3	E1 e E2	0,71	E1 e E2	0,80	E1 e E2	0,79
G1	C5.1	1	E2	–	E2	–	E1 e E2	0,75
G1	C5.2	4	E2	–	E2	–	E1 e E2	0,73
G4	C1.1	1	–	–	–	–	E1 e E2	0,93
G4	C2.1	1	–	–	–	–	E2	–

Fonte: Jallysson Miranda Rocha, 2018.

³ Há contextos nos quais o objeto motivador varia, por exemplo para a contação de histórias, duas tirinhas diferentes foram usadas. Nesses casos, a codificação inclui um segundo número indicando a escolha do objeto motivador.

As rotulações foram feitas por dois especialistas de forma a aumentar a confiabilidade do processo de rotulação. Para os vídeos rotulados pelos dois especialistas, foi calculado o coeficiente de concordância *Krippendorff's Alpha*, conforme listado na tabela 5. Os valores dos coeficientes indicam que a concordância nas rotulações pode ser qualificada como de qualidade “substancial” (valor de coeficiente de concordância no intervalo $[0, 6; 0, 8)$) ou “perfeita” (valor de coeficiente de concordância no intervalo $[0, 8; 1, 0]$) segundo escala apresentada em Artstein e Poesio (2008). Ainda segundo os mesmos autores, apenas se a concordância alcança o intervalo de $[0, 8; 1, 0]$ seria possível derivar conclusões assertivas; e valores de concordância no intervalo $[0, 6; 0, 8)$ permitiriam “tentativas de conclusões”, ou levantamento de hipóteses.

Tabela 6 – Quantidade de quadros por fases do gesto segundo rotulação realizada observando as duas mãos, por cada especialista, para oito vídeos. Legenda: G - gesticulador; C - contexto; Se - sessão; E1 e E2 – especialistas

G	C	Se	Descanso		Preparação		<i>Hold</i>		<i>Stroke</i>		Retração	
			E1	E2	E1	E2	E1	E2	E1	E2	E1	E2
G1	C1.1	1	5333	-	500	-	392	-	1429	-	337	-
G1	C2.1	1	1246	1298	525	817	765	437	1572	1552	361	365
G1	C3.2	1	1104	1090	256	444	222	308	987	691	214	250
G1	C3.3	3	894	878	111	209	0	73	427	209	127	190
G1	C5.1	1	635	573	560	1125	632	469	1991	1480	300	471
G1	C5.2	4	666	866	411	725	555	200	1608	1256	235	427
G4	C1.1	1	4136	4067	137	400	97	134	1299	960	177	281
G4	C2.1	1	-	227	-	646	-	132	-	545	-	95

Fonte: Jallysson Miranda Rocha, 2018.

No entanto, altos valores de concordância de rotulação são obtidos principalmente para vídeos de contextos que geram grandes períodos de descanso, pois nesses casos, o trabalho de rotulação é simplificado. O problema da discordância na rotulação das fases do gesto se dá principalmente nos quadros de transição entre fases, principalmente quando as fases de preparação e retração estão envolvidas nesta transição. Desta forma, a confiabilidade das rotulações pode ser mais bem aferida se houver uma redução do número de quadros referentes a períodos de descanso.

A figura 13 ilustra o problema em discussão. Na figura, cada círculo na linha “acordo” se refere a um quadro de vídeo, tomado em intervalos de 50 quadros, no qual os especialistas concordaram em suas rotulações. A cor verde no preenchimento do círculo indica que aquele quadro se refere a um quadro de período de descanso e a cor cinza indica que o quadro se refere a uma das outras fases do gesto. A ausência de um círculo na linha

indica que houve uma discordância nas rotulações, e tal discordância é indicada nas duas linhas abaixo (E1 e E2). Na figura 13(a), a primeira discordância ocorreu em um quadro rotulado como *stroke* pelo E1 e como *preparação* pelo E2.

Para a construção da figura 13(b), foram suprimidos quadros referentes a períodos de descanso para os quais houve concordância na rotulação dos dois especialistas, e que envolviam uma mão em movimento e a outra em posição de descanso. A supressão desses quadros evidencia a discordância nas rotulações dos especialistas para os trechos de vídeos mais complexos, e mostra que há uma alta subjetividade na tomada de decisão no processo de rotulação. O problema da subjetividade de rotulação é explorado no contexto dos classificadores SVM e os efeitos do problema são retomados no capítulo 5.1.

A figura 14 dá destaque a um trecho menor do vídeo analisado na figura 13. Nela é possível verificar com mais clareza algumas situações nas quais ocorre a discordância de rotulação.

Após a supressão dos quadros de vídeos considerados mais simples para rotulação, os coeficientes de concordância foram recalculados. Na tabela 7 esses novos coeficientes são apresentados para que se tenha uma ideia mais precisa sobre a complexidade inerente à subjetividade de rotulação das fases do gesto. A qualidade da concordância de rotulação é menor, caindo um ou dois níveis segundo análise realizada conforme orientado em Artstein e Poesio (2008).

Tabela 7 – Coeficientes de Concordância *Krippendorff's Alpha* (CA) para vídeos rotulados por dois especialistas após supressão de quadros de vídeo de fácil rotulação. Legenda: G - gesticulador; E1 e E2 – especialistas; Se - sessão; Rot1 - rotulação observando a mão direita; Rot2 – rotulação observando a mão esquerda; Rot3 - rotulação observando as duas mãos; # – quantidade de quadros de vídeos suprimidos

G	Contexto	Se	Rot1	#	CA	Rot2	#	CA	Rot3	#	CA
G1	C2.1	1	–	–	–	–	–	–	E1 e E2	1189	0,8
G1	C3.2	1	E1 e E2	955	0,47	E1 e E2	1141	0,43	E1 e E2	1013	0,52
G1	C3.3	3	E1 e E2	815	0,46	E1 e E2	904	0,33	E1 e E2	795	0,28
G1	C5.1	1	–	–	–	–	–	–	E1 e E2	527	0,54
G1	C5.2	4	–	–	–	–	–	–	E1 e E2	545	0,56
G4	C1.1	1	–	–	–	–	–	–	E1 e E2	4045	0,49

Fonte: Jallysson Miranda Rocha, 2018.

Figura 13 – Amostragem de quadros rotulados para a análise discordância para um vídeo: (a) com todos os quadros; (b) para um vídeo com redução de períodos de descansos

(a) *sem redução*

(b) *com redução*

Fonte: Jallysson Miranda Rocha, 2018.

4.4 Pre-processamento e extração de características

O conteúdo de interesse no *corpus* diz respeito à gesticulação realizada durante a conversação natural ou durante o discurso. Assim, o que de fato é usado para análise das fases dos gestos são as coordenadas espaciais x , y , z dos pontos das articulações do corpo do gesticulador captados pelo sensor.

Figura 14 – Destaque para as discordâncias de rotulação

Fonte: Jallysson Miranda Rocha, 2018.

A representação dos gestos presentes nos vídeos é, portanto, uma matriz numérica na qual cada linha corresponde ao um quadro do vídeo e um identificador numérico e as colunas correspondem às coordenadas espaciais de cada ponto de interesse rastreado no corpo do gesticulador. Assim, a matriz é composta de N linhas, em que N é número de quadros no vídeo. A tabela 8 mostra um esquema que ilustra resumidamente como um vídeo está representado – importante notar que não há qualquer forma de identificação da identidade do gesticulador. Uma informação relacionada a um selo de tempo⁴ faz a correspondência da linha da matriz com um arquivo de imagem (o quadro do vídeo) e suporta o trabalho de rotulação. A matriz é complementada com as rotulações durante a fase de rotulação. Esse é o formato que deve usado em eventuais publicações do *corpus*.

Tabela 8 – Matriz numérica que representa um vídeo

Quadro	X_mão direita	Y_mão direita	Z_mão direita	...	Selo de tempo	Rot1	Rot2	Rot3
1	4,5	4,0	1,1	..	201710081610	<i>Descanso</i>	<i>Descanso</i>	<i>Descanso</i>
2	5,5	4,5	1,2	..	201710081620	<i>Stroke</i>	<i>Stroke</i>	<i>Stroke</i>
..	6,5	6,0	1,3	..	..	..	..	..
..	8,0	7,0	1,4	..	..	..	..	..
N	9,5	8,0	1,5	..	201710082690	<i>Hold</i>	<i>Rest</i>	<i>Hold</i>

Fonte: Jallysson Miranda Rocha, 2018.

As coordenadas espaciais representam a trajetória que os pontos de interesse captados pelo sensor percorrem no espaço 3D. No entanto, esta trajetória não necessariamente

⁴ “Devido a uma limitação do dispositivo, não é possível gravar simultaneamente o dado e a imagem, portanto, existe uma discrepância de milésimos ou até centésimos de segundo entre a imagem e o dado correspondente” (MADEO, 2013).

representa o gesto manual executado, pois nela está também embutido o deslocamento corporal do gesticulador. Para que o gesto seja representado corretamente, é necessário anular esse deslocamento por meio de uma operação de normalização usando os pontos do peito e da cabeça como referência, por meio da seguinte operação aplicada às coordenadas espaciais de todos os pontos de interesse, quadro a quadro do vídeo:

$$p'_{i,j} = \frac{p_{i,j} - p_{i,t}}{\|p_{i,c} - p_{i,t}\|},$$

em que $p_{i,j} \in \{x_{i,j}, y_{i,j}, z_{i,j}\}$ são as coordenadas tridimensionais de cada um dos j , com $j = \{1, \dots, 8\}$, pontos de articulação captados para o quadro i ; $p_{i,t}$ e $p_{i,c}$, ambos $\in \{x_{i,j}, y_{i,j}, z_{i,j}\}$, são as coordenadas espaciais dos pontos do tronco e da cabeça, respectivamente; $p'_{i,j}$ são as coordenadas tridimensionais de cada um dos j pontos de articulação captados para o quadro i normalizadas; e $\|\cdot\|$ é a distância euclidiana. Na operação de normalização proposta, as coordenadas do tronco atuam como uma referência para anular deslocamentos do corpo nos eixos x , y e z , e a distância entre tronco e cabeça atuam como uma referência para anular efeitos de aproximação e distanciamento do gesticulador em relação ao sensor.

Os efeitos da operação de normalização podem ser observados na figura 15. No quadro (a) estão plotadas as coordenadas x e y originais referentes a um trecho de vídeo. O quadro (b) mostra a mesma informação porém após a realização da operação de normalização. A anulação dos deslocamentos indesejados pode ser observada na plotagem dos pontos referentes ao tronco. No quadro (a) é perceptível que há uma movimentação do tronco e portanto uma movimentação do corpo do gesticulador. No quadro (b) essa movimentação é anulada.

O uso das coordenadas x , y e z (coordenadas espaciais) para caracterização dos gestos geram uma representação que pode ser interpretada como um sinal (MADEO, 2013). Esse sinal é construído pelo sequenciamento das coordenadas para um ponto de articulação (representando uma sequência de quadros). Com o intuito de minimizar a presença de ruído neste sinal, foi aplicado o filtro de suavização *wavelet*. Na figura 16, o primeiro quadro mostra o sinal original da coordenada x da mão direita; o segundo quadro mostra o mesmo o sinal com a aplicação do filtro. Pequenas suavizações podem ser percebidas no sinal filtrado.

Figura 15 – Efeitos da normalização aplicada às coordenadas espaciais x e y em um trecho de vídeo

(a) *sem normalização*

(b) *normalizado*

Fonte: Jallysson Miranda Rocha, 2018.

A partir dos dados normalizados, considerando opcionalmente a aplicação do filtro *wavelet*, a extração de características temporais foram realizadas⁵. Essas características dizem respeito à velocidade vetorial e escalar e aceleração vetorial e escalar, do movimento realizado por cada ponto de articulação do gesticulador durante a execução do gesto.

⁵ As características temporais representam aspectos dinâmicos das fases do gestos, escondendo padrões de comportamento dos gesticuladores associados a posicionamento das mãos. Por exemplo, há gesticuladores que executam a sua posição de descanso junto aos quadris, enquanto outros executam a posição de descanso junto à cintura ou ao tronco. Para a criação de modelos classificadores universais, tal aspecto do gesto não deveria ser considerado.

Figura 16 – Filtro *wavelet* aplicado na coordenada tridimensional x da mão direita.

Fonte: Jallysson Miranda Rocha, 2018.

A velocidade vetorial é dada por:

$$v_{l,j}^{vet} = \frac{p_{i+d,j} - p_{i,j}}{d},$$

em que $p_{i,j} \in \{x_{i,j}, y_{i,j}, z_{i,j}\}$ são as coordenadas tridimensionais de cada um dos J , com $J = \{1, \dots, 6\}$, pontos de articulação captados para o quadro i ; d é um deslocamento nos quadros do vídeo; e $v_{l,j}^{vet}$ é a velocidade vetorial em cada coordenada tridimensional de cada um dos J , com $J = \{1, \dots, 6\}$, pontos de articulação captados para o quadro i , sendo que $l = \{i + d, \dots, N\}$. A velocidade vetorial é, portanto, calculada em função do deslocamento de um ponto de articulação em um eixo do espaço $3D$ sobre o tempo dado em quantidade de quadros (d). Nos trabalhos de Ramakrishnan (2011) e Madeo (2013) são utilizados diferentes tamanhos de d para observar o efeito de d no ruído gerado no cálculo da velocidade, considerando $d > 1$. Neste trabalho d foi usado igual a 3, considerando um dos melhores valores testado no trabalho de Madeo (2013).

Além da velocidade vetorial calculada para uma coordenada, também foi calculada a velocidade vetorial média, que transforma os três valores das coordenadas (x, y e z) de um ponto de articulação em um único valor, substituindo $p_{n,i} - p_{n,i-d}$ por $\sqrt{(x_i - x_{i-d})^2 + (y_i - y_{i-d})^2 + (z_i - z_{i-d})^2}$, que é o mesmo que calcular o deslocamento (hipotenusa) entre o quadro de interesse e o quadro anterior igual ao tamanho de d .

A partir da velocidade vetorial, a aceleração vetorial é calculada como:

$$a_{k,j}^{vet} = \frac{v_{l+t,j}^{vet} - v_{l,j}^{vet}}{t},$$

em que $a_{k,j}^{vet}$ é a aceleração vetorial em cada coordenada tridimensional de cada um dos J , com $J = \{1, \dots, 6\}$, pontos de articulação captados para o quadro i , sendo que $k = \{l+t, \dots, N\}$. Como em Ramakrishnan (2011), a aceleração é calculada a partir de uma unidade de tempo $t = 1$, ou seja, considerando um deslocamento de um quadro de vídeo. É calculada também a aceleração vetorial média dos pontos de articulação, calculada a partir da diferença entre a velocidade vetorial média nos quadros de vídeo, com $t = 1$.

As características referentes à velocidade e aceleração escalares são obtidas a partir das vetoriais correspondentes, da seguinte forma:

$$m_{i,J}^{esc} = \sum_{j \in \{x_{i,J}, y_{i,J}, z_{i,J}\}} (m_{i,j}^{vet})^2,$$

em que $m_{i,j}^{vet}$ é a medida vetorial adequada, ou seja, velocidade ou aceleração nos eixos x , y e z do ponto de articulação considerado no quadro i ; J é o ponto de articulação, com $J = \{1, \dots, 6\}$; e $m_{i,J}^{esc}$ é a medida escalar calculada. As medidas escalares foram também calculadas desconsiderando o valor da coordenada z .

A figura 17 mostra um sinal (representação gestual) construído a partir da medida de velocidade escalar (quarto superior) e aceleração escalar (quarto inferior) para o ponto de articulação “mão direita”, considerando um trecho de um dos vídeos captados.

Por fim, considerando o dado gestual como um sinal variante no tempo, como uma série temporal, e com o objetivo de embutir mais informação sobre a dinâmica do gesto em sua representação vetorial, neste trabalho, assim como em Madeo (2013), foi escolhido usar uma técnica de reconstrução do espaço de fase do sinal, usando a informação do próprio sinal junto a um atraso de tempo. Esse método permite a geração de um espaço de características alternativo ao original, criando uma representação espacial do tempo no qual aspectos temporais possam ser explorados mesmo a partir da aplicação de técnicas de análise que não incorporam raciocínio temporal (MADEO, 2013), como é o caso da SVM tradicional usada neste trabalho.

Considerando uma reconstrução do espaço de fases com atraso de 1 quadro obtém-se o que também se conhece por janelas retangulares. A figura 18 mostra uma janela retangular na qual um quadro de interesse é representado por uma sequência de descrições

Figura 17 – Sinal correspondente à velocidade (quadro superior) e aceleração (quadro inferior) escalares da mão direita

Fonte: Jallysson Miranda Rocha, 2018.

de três quadros anteriores a ele e três quadros posteriores a ele (o quadro de interesse está no meio da janela). O quadro de interesse é aquele que está sob análise, portanto a rotação da fase associada a ele é a rotulação associada ao dado janelado. Para representar o próximo quadro de interesse, é preciso deslizar a janela em 1 quadro à frente (o que é equivalente a reconstruir o espaço de fases com atraso de 1). Nesse caso, disse que a janela contém informação sobre o passado e o futuro da informação contida no quadro de interesse. No caso desse exemplo, considerando que um quadro é representado pela velocidade escalar da mão direita (v_{md}) e velocidade escalar da mão esquerda (v_{me}), a representação vetorial para o quadro de interesse, com janela de tamanho sete, tem dimensão 14.

Figura 18 – Uma janela retangular para construir uma representação vetorial do tempo em relação ao quadro de interesse

Fonte: Jallysson Miranda Rocha, 2018.

4.5 *Corpus versus conjunto de dados*

O *corpus* de vídeos produzido neste trabalho conta com 89 vídeos (cf. descrito na tabela 3), referentes a captação de gesticulação de sete usuários e considerando cinco contextos: entrevista (20 vídeos), improvisação (22 vídeos), contação de história (22 vídeos), descrição de cena (21 vídeos) e aula expositiva (quatro vídeos). Esse *corpus* totaliza duas horas e 48 minutos de gravação de vídeo. O áudio dos vídeos está em posse do grupo de pesquisa porém não foi utilizado neste trabalho.

O esforço de rotulação executado resultou em oito vídeos rotulados, por dois especialistas. De forma mais específica, tem-se:

- sete vídeos rotulados pelo especialista E1;
- sete vídeos rotulados pelo especialista E2;
- seis vídeos rotulados pelos dois especialistas.

A rotulação considerando cada uma das mãos separadamente foi realizada em dois vídeos, pelos dois especialistas e três vídeos pelo especialista E1 e quatro para o especialista E2. Os contextos presentes no conjunto de vídeos rotulados abrangem a entrevista, a improvisação, a contação de histórias e a aula expositiva. Apenas para os contextos de contação de histórias e aula expositiva, a rotulação foi realizada considerando duas sessões de um mesmo gesticulador.

Uma ideia mais geral das características do *corpus* produzido, considerando os vídeos rotulados, pode ser obtida da análise dos gráficos da figura 19. Note que a quantidade

absoluta de quadros (figura a esquerda) rotulados como descanso e *stroke* são bem maiores que a quantidade absoluta de quadros rotulados para outras fases, principalmente na visão do especialista E1. Trata-se portanto de um contexto de classificação com classes desbalanceadas. Também é notável a subjetividade de rotulação expressa para as fases do gesto de “preparação” e “*stroke*”, o que torna evidente que o estudo sobre as rotulações precisa ser aprofundado. O gráfico mostrado na figura 19 fornece uma ideia da proporção (figura a direita) de cada fase dentro do *corpus* segundo a rotulação dos dois especialistas.

Figura 19 – Quantidade (figura a esquerda) e porcentagem (figura a direita) de quadros por fase do gesto no *corpus*, considerando os seis vídeos rotulados pelos especialistas E1 e E2

Fonte: Jallysson Miranda Rocha, 2018.

Os gráficos apresentados na figura 20 mostram como o padrão de gesticulação de um gesticulador (G1) muda em diferentes sessões de captação. Nos gráficos, as barras em azul dizem respeito à proporção de quadros rotulados em cada fase do gesto em dois vídeos diferentes, captados em sessões diferentes porém considerando um mesmo contexto de gesticulação – contação de história. As barras em amarelo representam o mesmo tipo de informação, porém para o contexto de aulas expositivas. O gráfico da esquerda se refere à rotulação do especialista E1 e o gráfico da direita à rotulação do especialistas E2.

Por fim, a variação no padrão de gesticulação de um mesmo gesticulador considerando mudança de contexto é ilustrado nos gráficos da figura 21. Nesses gráficos percebe-se que ainda com a subjetividade de rotulação inerente ao problema, o período de descanso e as fases de preparação e retração foram percebidas pelos dois especialistas de forma similar. Principalmente nestas três fases, a observação da mudança de padrão de gesticulação entre os contextos fica evidente. Para as fases de *hold* e *stroke*, houve uma interpretação diferente por parte dos especialistas.

Figura 20 – Padrão de gesticulação do gesticulador G1 em sessões diferentes para os contextos: contação de histórias (azul), aulas expositivas (amarelo)

Fonte: Jallysson Miranda Rocha, 2018.

Figura 21 – Padrão de gesticulação do gesticulador G1 em contextos diferentes: improvisação (azul), contação de histórias (laranja), aula expositiva (cinza)

Fonte: Jallysson Miranda Rocha, 2018.

O conjunto de dados gerado a partir do *corpus* de vídeo é composto por matrizes numéricas referentes aos 89 vídeos, considerando os atributos descritivos espaciais (coordenadas x, y, z) dos pontos de articulação captados pelo sensor, e considerando atributos descritivos temporais derivados da relação das coordenadas espaciais com a taxa de captação dos quadros de vídeos. Os atributos temporais são: velocidades e acelerações vetoriais por coordenadas, velocidades e acelerações vetoriais médias e velocidades e aceleração, e velocidades e acelerações escalares.

Além dos atributos descritivos, matrizes de dados janeladas foram construídas considerando janelas de tamanhos 10, 20, 30, 40, 50, 60, 70, 80, 90 e 100 com quadros de interesse no início, meio e fim das janelas.

Enfim, todas as versões dos conjuntos de dados que foram criados a partir de vídeos que estão rotulados possuem os rótulos associados, podendo ser usadas em tarefas de classificação. Os demais conjuntos de dados, referentes aos vídeos não rotulados, estão disponíveis para uso em tarefas de agrupamento.

Em um total de 89 vídeos (oito rotulados e 81 não rotulados) foram geradas 2.400 matrizes de dados para os vídeos rotulados considerando todas as representações e variações de tamanhos de janelas e também 81 matrizes de dados dos vídeos não rotulados, sendo disponibilizadas para estudos de análise de segmentação de gestos.

5 Experimentos, resultados e discussão

Neste capítulo são descritos os procedimentos de experimentação usados neste trabalho para estudo da relação entre o desempenho de classificadores, usados para segmentação das fases de gestos, e os diferentes contextos de gesticulação natural presentes no *corpus* rotulado (descrito no capítulo 4.4). Os resultados obtidos bem como análises decorrentes destes resultados são apresentados.

Este capítulo segue organizado da seguinte forma: na seção 5.1, é apresentada a organização dos experimentos, incluindo as estratégias de treino e teste adotadas para construção e avaliação dos modelos classificadores, a parametrização usada no pré-processamento dos conjuntos de dados e nos modelos de SVM; na seção 5.2 são apresentados os resultados obtidos com os experimentos realizados neste trabalho, acompanhados de discussões delineadas sobre tais resultados. O capítulo é concluído com as considerações finais que resumem as principais constatações observadas nos experimentos.

5.1 Organização dos experimentos

A fim de alcançar os objetivos de análise pretendidos para este trabalho, três tipos de experimentos foram executados:

- #1: : construção de modelos classificadores para resolução da tarefa de segmentação das fases do gesto usando o conjunto de dados *Gesture Phase Segmentation*¹, construído por Madeo (2013). Esse experimento foi executado em paralelo à construção do novo *corpus*. Embora o conjunto de dados *Gesture Phase Segmentation* não apresente dados adequados e suficientes para suportar a análise objetivada neste trabalho, ele provê um conjunto de dados que possui características similares ao novo conjunto de dados.
- #2: execução de um estudo sobre a subjetividade de rotulação e construção de classificadores SVM para segmentação de um mesmo vídeo porém considerando variações de rotulação. As variações foram criadas a partir das rotulações dos dois especialistas e de combinações destas rotulações. Esse experimento foi executado sobre o vídeo referente ao gesticulador G1, no contexto de discurso de contação de histórias. Este

¹ Disponível em: *UCI Repository* (LICHMAN, 2013) - (<https://archive.ics.uci.edu/ml/datasets/Gesture+Phase+Segmentation>).

experimento teve o objetivo de ilustrar a influência da subjetividade de rotulação sobre os resultados de um classificador e também indicar a rotulação que deveria ser usada na execução do experimento do tipo #3.

#3: construção de modelos classificadores para resolução da tarefa de segmentação das fases do gesto usando o conjunto de dados elaborado neste trabalho (cf. capítulo 4.4). Esse experimento foi executado com o intuito de analisar a hipótese delineada neste trabalho, referente à relação do desempenho dos classificadores e o padrão de gesticulação apresentado em diferentes contextos de discurso.

5.1.1 Estratégia de treino e teste de classificadores

Neste trabalho, a segmentação da fases do gesto é realizada sobre sequências de quadros de vídeo. Cada vídeo existente nos conjuntos de dados usados é considerado uma instância de treino/teste dos modelos classificadores construídos para resolução da tarefa. Doravante neste texto, cada vídeo será considerado como um conjunto de dados sobre o qual um modelo pode ser treinado e testado.

Para construir e avaliar os modelos classificadores usando SVM foi aplicada a estratégia *Holdout*, na qual divide-se os conjuntos de dados em duas partes: a primeira, composta de 70% ou 60% dos quadros de um vídeo, usada para treinamento dos modelos; a segunda, composta dos quadros de vídeos restantes (30% ou 40% do vídeo), usada para teste. A primeira combinação (70/30) foi usada no experimento do tipo #1, seguindo o que já havia sido aplicado em trabalhos anteriores. A segunda combinação (60/40) foi usada nos experimentos do tipo #2 e #3.

Os vídeos são compostos por sequências de quadros que possuem uma dependência temporal entre si. Por isso, a divisão dos conjuntos de dados em parte para treino e parte para teste é realizada de forma sequencial, ou seja, não há escolha aleatória de quadros no procedimento de divisão das partes. Embora não seja mandatório, neste trabalho, conjuntos de treinamento provenientes desta estratégia de divisão são compostos pela parte inicial dos vídeos, e os conjuntos de testes são compostos pela parte final dos vídeos.

Na estratégia descrita, um único vídeo é usado no treinamento e teste de modelos. Há ainda outras duas formas de apresentar os conjuntos de dados para o treinamento e para o teste:

- com dois vídeos, de forma que um, em sua totalidade, seja usado como conjunto de dados de treinamento e o outro, também em sua totalidade seja usado como conjunto de dados de teste;
- com três ou mais vídeos, sendo que, pelo menos, dois sejam usados para treinamento (os dois vídeos são concatenados em sequência para formar o conjunto de treinamento) e um seja usado para teste dos modelos. A figura 22 ilustra a maneira como vários vídeos são organizados para compor um ambiente de construção de modelos neste trabalho.

Figura 22 – Composição de conjuntos de treinamento usando n vídeos: $n - 1$ vídeos para treinamento e um vídeo para teste. Cada vídeo de m quadros é usado em sua totalidade

Fonte: Jallysson Miranda Rocha, 2018.

5.1.2 Parametrização

Nos experimentos do tipo #1, o conjunto de parâmetros que produziram os melhores resultados no trabalho de Madeo (2013) foi utilizado. Assim, as seguintes escolhas foram adotadas junto a uma representação para rótulos binária e estratégia de composição de resultados um contra todos:

- representação do gesto usando velocidade e aceleração escalares derivadas das coordenadas espaciais (x , y e z) de quatro pontos de articulação do gesticulador (mãos e pulsos direito e esquerdo);
- janelamento de tamanho 69 com quadro de interesse no meio da janela;
- SVMs com função *kernel* de base radial, com abertura em 16 e margem em 100.

Os experimentos do tipo #2 e #3 foram executados a partir de um estudo de parametrização realizado sobre o novo conjunto de dados. Esse estudo foi dividido em três passos:

1. construção de modelos SVM sobre o vídeo rotulado pelo especialista E1 no contexto de contação de histórias (segunda história do gesticulador G1), considerando variações referentes às características descritivas do gesto (coordenadas espaciais, velocidade e aceleração vetorial e escalar), aos tamanhos de janela e aos parâmetros da função *kernel* de base radial usada na SVM;
2. análise do resultado dos modelos gerados, por meio da medida de avaliação *F-score* gerada conforme descrito no capítulo 2.1 (usando matrizes de confusão multiclasse construídas com o resultado dos modelos da estratégia um contra todos);
3. construção de novos modelo SVM, considerando os parâmetros dos melhores modelos de acordo com o resultado dos passos 1 e 2, para todos os vídeos rotulados pelo especialista E1 disponíveis com todas as fases do gesto e descanso, com diferentes estratégias de composição de conjuntos de treinamento e teste.

No primeiro passo, a exploração dos parâmetros para a construção do modelos com SVM foi realizada por meio de uma busca em grade considerando os espaços de busca descritos nas tabelas 9 e 10. A exploração foi realizada na análise de apenas um vídeo e utilizando apenas uma rotulação (**R1**) realizada pelo especialista mais experiente (E1).

Tabela 9 – Espaço de busca relacionado à representação do gesto

Pontos de articulação	mãos, pulsos e cotovelos direitos e esquerdos
Filtro (suavizar do sinal)	<i>Wavelet</i> com função <i>db2</i> e nível variando de um a dez
Características	coordenadas x,y,z , velocidade e aceleração escalar e vetorial
Número de quadros na janela	10, 20, 30, 40, 50, 60, 70, 80, 90 e 100
Posição de q_{int} na janela	início, meio e fim

Fonte: Jallysson Miranda Rocha, 2017.

Tabela 10 – Espaço de busca referente aos parâmetros da SVM

Função <i>kernel</i>	Função de base radial
Parâmetro do <i>kernel</i>	2^{-5} a 2^5 , variando o expoente em um passo de tamanho 1
Margem	1, 100, 500, 1000 e 1 a 500, variando passos de tamanho 50

Fonte: Jallysson Miranda Rocha, 2017.

Modelos classificadores foram gerados no passo 1 a partir da combinação dos valores que cada um dos parâmetros pode assumir. Então, modelos que obtiveram valores de $F\text{-score} \geq 0,6$ para pelo menos duas das fases do gesto foram selecionados. Os parâmetros mais frequentes para esse subconjunto de modelos foram considerados os melhores parâmetros para o problema de segmentação de gestos em estudo, e estão listados na tabela 11. Destes, dois casos foram considerados os mais consistentes para uso no passo 3 (cf. discutido mais à frente neste capítulo), e estes estão destacados em verde na tabela. Nos casos em que não houve valores destacados como mais frequentes, a escolha não foi realizada.

Tabela 11 – Melhores parâmetros para a tarefa de segmentação das fase do gesto em estudo neste trabalho

CD	PA	TJ	QI	PK	M
esp	md, md-pd	40	meio	0,031 e 0,062	?
vel vet	?	20	meio	8,16	100, 500, 1000
vel esc	md-pd, md-pd-cd	40, 50	meio	8,16	500, 1000
ace vet	md, md-pd-me-pe	10,40	início e fim	8,16	500, 1000
ace esc	?	?	?	?	100, 500, 1000

CD - característica descritiva do gesto; PA: ponto de articulação; TJ - tamanho da janela; QI - quadro de interesse; PK - parâmetro da função *kernel*; M - margem; esp - coordenadas espaciais; vel - velocidade; ace - aceleração; vet - vetorial; esc - escalar; md - mão direita; me - mão esquerda; pd - pulso direito; cd - cotovelo direito; ? - não determinado.

Fonte: Jallysson Miranda Rocha, 2018.

Os pontos de articulação que mais aparecem nos melhores modelos são mão e pulso direito. Considerando que o gesticulador do vídeo utilizado é destro, essa constatação confirma a importância da mão dominante no padrão de gesticulação e indica que a observação apenas da mão dominante pode ser suficiente para segmentação das fases do gesto. Outra constatação frequente é o uso do quadro de interesse no meio da janela, indicando que informação anterior e posterior ao quadro é mais discriminante nesse problema do que informação apenas do passado ou apenas do futuro.

Especificamente, para o experimento do tipo #2, a fim de propiciar o estudo do impacto da subjetividade de rotulação nos resultados de classificação, foram criadas três novas rotulações para um dos vídeos do gesticulador G1 no contexto de contação de histórias, gerando variações de conjunto de dados a serem submetidos a análise de modelos SVM. As rotulações originais são aqui identificadas como **R1** e **R2**. A primeira é a rotulação original do especialista E1 e a segunda é a rotulação original do especialistas E2. A três novas rotulações (**R3**, **R4** e **R5**) tiveram o objetivo de reduzir a discordância

entre os especialistas e, conseqüentemente, reduzir o nível de subjetividade inerente à segmentação das fases do gesto. As novas rotulações são decorrentes de decisões tomadas em conjunto pelos dois especialistas, considerando: (a) a discordância na rotulação de quadros referentes a períodos de descanso; (b) a discordância na rotulação de quadros referentes à fase *stroke*.

No caso (a), a subjetividade gera discordâncias de rotulação que resultam na inserção de fases do gesto dentro dos períodos de descanso. Para minimizar esse problema, dois padrões de correção foram seguidos, em comum acordo com os especialistas a partir da análise do vídeo:

R3 há uma mudança na posição de descanso erroneamente interpretada como um gesto.

Nesse caso é realizada a substituição dos rótulos de fases do gesto por rótulos de período de descanso.

R5 o gesticulador inicia um gesto, mas recua como se desistisse de realizá-lo. Nesse caso, os respectivos quadros são rotulados como fases de preparação e retração.

No caso (b), ocorrem situações semelhantes às anteriores, contudo dentro da fase *stroke*. Frequentemente, as discordâncias ocorrem porque o gesticulador transita de um gesto significativo para outro, possibilitando duas interpretações – um único (e longo) *stroke*, dois *strokes* separados por uma fase de preparação (ou retração) ou dois *strokes* separados com fases de preparação e retração. Para minimizar esse problema, três padrões de correção foram seguidos, em acordo com os especialistas a partir da análise do vídeo:

R3 é realizada a substituição dos rótulos das fases preparação e retração por rótulos da fase *stroke*.

R4 acata-se a fase de preparação como a fase de transição entre os dois *strokes* e o rótulo preparação é atribuído aos quadros considerados como transição entre os *strokes*.

R5 acata-se as fases de retração e preparação como fases de transição entre os dois *strokes* e os rótulos referentes a essas fases são adequadamente atribuídos aos quadros considerados como transição entre os *strokes*.

5.2 Resultados e discussões

Nesta seção estão organizados os resultados obtidos com os três tipos de experimentos executados neste trabalho. Os resultados seguem acompanhados de discussões que representam a busca pelo cumprimento dos objetivos deste trabalho.

5.2.1 Resultados e análises: experimento do tipo #1

O conjunto *Gesture phase segmentation* é composto de sete vídeos obtidos a partir de contações de histórias. Para formar o conjunto de dados, três gesticuladores (*A*, *B* e *C*) contam de uma a três histórias (1, 2 e 3). Os dados originalmente obtidos pelo sensor *kinect* dizem respeito às coordenadas espaciais dos pontos de articulação: mãos, pulsos, tronco e cabeça do gesticulador. Para aproximar o ambiente de teste deste experimento aos objetivos deste trabalho, o conjunto de dados *Gesture phase segmentation* foi usado considerando:

- três usuários diferentes contando a mesma história (experimento #1.1);
- o mesmo usuário contando duas histórias diferentes (experimento #1.2);
- o mesmo usuário contando três histórias diferentes (experimento #1.3).

Em termos de variação do contexto do discurso, os experimentos #1.2 e #1.3 constituem os ambientes de testes mais adequados aos objetivos deste trabalho.

Experimento #1.1

O estudo realizado com esse experimento diz respeito ao treinamento e teste de um modelo classificador para cada um dos gesticuladores contando a mesma história. Os modelos são compostos por cinco classificadores binários (um para cada fase do gesto) compostos usando a estratégia um contra todos.

Assim foram gerados três grupos de cinco classificadores binários, um grupo para analisar cada um dos vídeos disponíveis no conjunto de dados. A estratégia de treino e teste bem como a variação dos parâmetros para os modelos seguiram o exposto na seção 5.1.

O melhor resultado foi obtido a partir da análise do vídeo correspondente à história 1 contada pelo gesticulador A. Esse resultado é apresentado na tabela 12, em termos da medida de avaliação de *F-score*, obtidas a partir da matriz de confusão (tabela 13)².

Tabela 12 – Melhores resultados entre os classificadores gerados no experimento #1.1.
Melhor resultado obtido para o gesticulador A na história 1

Contexto - história 1					
Gesticulador	Descanso	Preparação	Hold	Stroke	Retração
A	0,746	0,431	0,0	0,628	0,454
B	0,533	0,452	0,351	0,068	0,0
C	0,344	0,404	0,588	0,0	0,139

Fonte: Jallysson Miranda Rocha, 2018.

Tabela 13 – Matriz de confusão para o melhor resultado obtido no experimento #1.1

		Classificado como					Total
		D	P	H	S	R	
Classe real	D	182	7	0	10	27	226
	P	5	22	0	7	28	62
	H	0	0	0	0	9	9
	S	68	11	0	147	77	303
	R	7	0	0	1	62	70
	Total	262	40	0	165	203	

Fonte: Jallysson Miranda Rocha, 2018.

Os melhores resultados de classificação foram obtidos no período de **descanso** e na fase de **stroke**, *F-score* de 0,746 e 0,628. No que diz respeito às fases, esse resultado corrobora com o que já foi obtido em trabalhos anteriores realizados sobre esse conjunto de dados. Considerando três gesticuladores diferentes (A, B e C) no treinamento e teste do modelo classificador na história 1, os resultados foram diferentes nas fases do gesto e período de descanso. Embora, estivessem realizando uma atividade semelhante os três gesticuladores tiverem padrões de gesticulação diferentes para o período de descanso e as fases de *hold*, *stroke* e *retração*.

² Desde que a estratégia utilizada pode gerar uma resposta na qual um determinado quadro do vídeo é classificado em mais de uma classe (quando mais de um classificador no grupo responde com classificação na classe positiva), a magnitude da saída fornecida pelo SVM foi usada como base para tomar a decisão final de classificação.

Experimento #1.2

O segundo experimento consiste em criar o modelo classificador usando como conjunto de treinamento um vídeo completo, referente à contação de uma história pelo gesticulador *A*, e então usar um segundo vídeo referente à contação de outra história, pelo mesmo gesticulador *A*. A partir desse experimento é possível ensaiar algumas análises referentes à variação de contexto, neste caso, referente a duas variáveis: o conteúdo da história e o momento de captação do vídeo. Esse experimento foi realizado a partir de vídeos do gesticulador *A* porque no experimento #1.1, o melhor resultado de segmentação foi obtida para esse gesticulador. Há uma suposição que as fases são mais bem definidas na gesticulação desse gesticulador. Há também a motivação referente ao fato de haver três vídeos captados com esse usuário, o que possibilitou variar os vídeos usados como treinamento e como teste.

O melhor resultado de segmentação (tabela 14 e 15) foi obtido com o treinamento e criação do modelo usando a história 1 e com o teste usando a história 2. Esse resultado corrobora com outros testes, de diferentes naturezas, realizados pelo grupo de pesquisa. Em trabalhos prévios já foi constada uma facilidade maior na análise da história 1, seguida pela história 2 e então pela história 3, quando se dá a implementação da segmentação das fases do gesto via modelos classificadores. Muito embora não tenha ainda sido possível, nem mesmo nestes experimentos, compreender o motivo para esse fato.

Tabela 14 – Melhor resultado entre os modelos classificadores gerados no experimento #2 para o gesticulador *A* com treinamento na história 1 e teste na história 2

Classe	Precisão	Revocação	<i>F-score</i>
Descanso	71,629	89,107	0,794
Preparação	56,911	35,714	0,439
<i>Hold</i>	100,00	4,762	0,091
<i>Stroke</i>	72,191	59,629	0,653
Retração	24,286	53,125	0,333

Fonte: Jallysson Miranda Rocha, 2018.

Os resultados em termos de *F-score* considerando esse experimento e o experimento #1.1 são semelhantes. Então, diante desse fato, ensaia-se uma inferência de que a mudança de história, porém com manutenção de gesticulador e contexto de discurso, não insere mudanças significativas no padrão de gesticulação.

Tabela 15 – Matriz de confusão para o melhor resultado obtido no experimento #2

		Classificado como					Total
		<i>D</i>	<i>P</i>	<i>H</i>	<i>S</i>	<i>R</i>	
Classe real	<i>D</i>	409	10	0	28	12	459
	<i>P</i>	45	70	0	48	33	196
	<i>H</i>	32	0	2	2	6	42
	<i>S</i>	76	43	0	257	55	431
	<i>R</i>	9	0	0	21	34	64
Total		571	123	2	356	140	

Fonte: Jallysson Miranda Rocha, 2018.

Explorando um pouco mais os modelos classificadores obtidos nesse experimento percebe-se que ao considerar o treino com a história 1 e um teste com a história 3 há uma queda no desempenho (Tabela 16). Interessante notar que, neste caso, há uma variação maior no ambiente de teste, pois a história 3 foi capturada em uma sessão diferente daquela na qual a história 1 foi capturada (o que não ocorre com a história 1 e a história 2, que foram capturadas na mesma sessão). A história 3, na realidade, foi capturada em um dia diferente da história 1 e 2.

Tabela 16 – Resultado do modelo treinado com a história 1 e testado com a história 3

Classe	Precisão	Revocação	<i>F-score</i>
Descanso	53,666	90,894	0,675
Preparação	52,381	23,656	0,326
<i>Hold</i>	0,000	0,000	0,000
<i>Stroke</i>	68,952	31,963	0,437
Retração	29,863	56,186	0,390

Fonte: Jallysson Miranda Rocha, 2018.

Experimento #3

No experimento #1.3 é feita uma combinação de vídeos com contação de histórias realizadas em sessões diferentes (história 2 e história 3) para realizar o treinamento do modelo, e realiza o teste usando a história 1. O resultado do melhor classificador obtido é mostrado na tabela 17 e a matriz de confusão na tabela 18.

Neste caso houve uma melhora no desempenho do classificador nas fases mais fáceis de segmentar quando comparado aos outros dois experimentos. Contudo há uma degradação para as classes mais difíceis. Não foi possível tirar conclusões a partir da observação desse comportamento devido à simplicidade do experimento, e devido aos

Tabela 17 – Melhor resultado entre os modelos classificadores gerados no experimento #1.3 para o treinamento na história 2 e 3 e teste na história 1

Classe	Precisão	Revocação	<i>F-score</i>
Descanso	89,720	65,866	0,760
Preparação	41,085	44,167	0,426
<i>Hold</i>	0,000	0,000	0,000
<i>Stroke</i>	72,245	93,322	0,814
Retração	28,125	16,770	0,210

Fonte: Jallysson Miranda Rocha, 2018.

Tabela 18 – Matriz de confusão para o melhor resultado obtido no experimento #1.3

		Classificado como					Total
		<i>D</i>	<i>P</i>	<i>H</i>	<i>S</i>	<i>R</i>	
Classe real	<i>D</i>	384	33	47	69	50	583
	<i>P</i>	24	53	3	32	8	120
	<i>H</i>	12	0	0	15	0	27
	<i>S</i>	4	23	0	531	11	569
	<i>R</i>	4	20	22	88	27	161
	Total	428	129	72	735	96	

Fonte: Jallysson Miranda Rocha, 2018.

resultados antagônicos (melhoras por um lado e pioras por outro). O esperado era uma melhora para todas as fases ou a degradação do classificador devido à variação de padrão que pressupõe-se haver para os vídeos das histórias obtidos em sessões diferentes, porém a hipótese não se mostrou verdadeira.

5.2.2 Resultados e análises: experimento do tipo #2

Inicialmente, modelos de SVM explorando todo o espaço de busca de parâmetros foram executados usando a rotulação **R1**. Então, um subespaço de parâmetros, no qual os modelos alcançaram os melhores resultados de segmentação das fases do gesto foi estabelecido (tabela 11). Dentro desse espaço de busca reduzido foi possível realizar um estudo sobre o impacto de diferentes rotulações sobre o desempenho de classificação. Novos classificadores foram gerados, usando o mesmo conjunto de treinamento e teste, porém considerando as rotulações **R2**, **R3**, **R4** e **R5**.

Em geral, nessa nova etapa de construção de classificadores, os melhores resultados foram obtidos sob a representação espacial dos gestos (coordenadas x , y , z), sinal suavizado

e janela de 40 quadros. Os desempenhos de classificação são listados na tabela 19, em termos dos melhores e piores *F-scores* obtidos.

Tabela 19 – Desempenho dos classificadores para cada fase do gesto, em termos de *F-score* máximo e mínimo.

Característica	Descanso		Rotulação		<i>Hold</i>		<i>Stroke</i>		Retração	
	max	min	max	min	max	min	max	min	max	min
R1	0,79	0,77	0,17	0,15	0,00	0,00	0,75	0,71	0,74	0,70
R2	0,78	0,71	0,43	0,26	0,38	0,24	0,60	0,54	0,49	0,27
R3	0,78	0,76	0,17	0,14	0,00	0,00	0,78	0,71	0,73	0,52
R4	0,79	0,76	0,25	0,19	0,03	0,00	0,69	0,60	0,69	0,53
R5	0,79	0,77	0,22	0,17	0,03	0,00	0,70	0,61	0,51	0,38

Fonte: Jallysson Miranda Rocha, 2018.

Tais resultados mostram que, em geral, não houve uma sensibilidade alta às adequações de rotulação realizadas. Os resultados obtidos com a rotulação **R1** são semelhantes àqueles obtidos com as rotulações **R3**, **R4** e **R5**. Porém, há algumas variações no resultado que nos permitem levantar algumas hipóteses:

- As rotulações **R1** e **R3** tendem a simplificar o problema, aproximando o problema de segmentação das fases do gesto ao problema de encontrar uma unidade gestual. O desbalanceamento de classes no conjunto de treinamento é acentuado.
- As rotulações **R4** e **R5** admitem maior variabilidade de fases em uma unidade gestual, e contribuem para o aumento de quadros de preparação e retração. Sob essa rotulação os classificadores aprendem um pouco mais sobre a fase de preparação, porém há indícios que isso ocorre em detrimento do aprendizado para *stroke* e retração. No entanto, principalmente sob a rotulação **R5**, a perda de desempenho para retração é acentuada e não parece refletir em aprendizado para preparação.
- Finalmente, ao observar os resultados sob a rotulação **R2**, percebe-se que o desempenho do classificador é baixo para todas as fases da unidade gestual, contudo, há uma distribuição mais uniforme do aprendizado entre as classes.

A análise dos erros introduzidos na seção 3.2.1 (EF, ET, EI, EXC, INS e SUB) foi realizada para os oito melhores classificadores obtidos. Os resultados são listados na tabela 20. Os erros EF, ET e EI são calculados em termos de porcentagem de quadros classificados erroneamente no vídeo completo. Os erros INS, EXC e SUB são calculados em termos de porcentagem de quadros envolvidos nesses erros.

Tabela 20 – Análise qualitativa dos melhores modelos obtidos sob cada rotulação

Rotulação	EF	ET	EI	EXC	INS	SUB
R1	35, 15	1, 55	33, 61	4, 46	0, 00	9, 38
R2	42, 53	11, 48	31, 06	1, 46	0, 00	10, 84
R3	34, 88	3, 19	31, 69	5, 56	0, 00	8, 47
R4	39, 34	4, 55	34, 79	5, 56	0, 00	7, 65
R5	39, 34	4, 74	34, 61	5, 56	0, 00	7, 65

Fonte: Jallysson Miranda Rocha, 2018.

Dado que os dados são desbalanceados (há uma quantidade maior de quadros de descanso e de *stroke* do que de outras fases), os resultados da tabela 20 devem ser considerados apenas para fins de comparação do impacto das diferentes rotulações.

Em termos de erros que ocorrem nas transições entre fases, há um destaque negativo para o caso da rotulação **R2**. Embora o erro de transição seja aceitável e possa ser desconsiderado na análise do classificador, a porcentagem alta de erros desse tipo indica que o classificador não conseguiu capturar o padrão de rotulação do especialista, sugerindo que se trata de um especialista com mais subjetividade na tomada de decisão. No classificador criado com a rotulação **R1**, verifica-se que o padrão de rotulação foi mais facilmente aprendido. Há portanto um indicativo que o classificador tem, de fato, alta sensibilidade à rotulação nos quadros de transição. Ainda, levanta-se a hipótese de que o classificador pode atuar também como um meio de analisar a qualidade do trabalho do especialista.

Os erros EXC, INS e SUB se mantiveram parecidos para todas as rotulações, o que não permitiu levantar hipóteses sobre o impacto da subjetividade das rotulações. Contudo, nota-se que erros de substituição nas fases de preparação e retração diminuíram para o caso da rotulação **R5** (Tabela 21). Essa rotulação pode favorecer o aprendizado dessas fases por torná-la mais representativa no conjunto de treinamento, mas também pode estar contribuindo para uma melhor definição do padrão quando impõe que essas fases sejam representadas entre mudanças de posições de descanso e entre *strokes* consecutivos.

5.2.3 Resultados e análises: experimento do tipo #3

Este experimento foi executado com o intuito de explorar o novo *corpus*/conjunto de dados, criado neste trabalho e apresentado no capítulo 4. A intenção em explorar

Tabela 21 – Detalhamento dos erros de substituição por fase do gesto

Rotulação	Descanso	Preparação	<i>Hold</i>	<i>Stroke</i>	Retração
R1	0,00	36,89	63,11	0,00	0,00
R2	0,00	22,69	47,90	5,88	23,53
R3	0,00	30,11	69,89	0,00	0,00
R4	0,00	22,62	77,38	0,00	0,00
R5	0,00	10,71	77,38	0,00	11,90

Fonte: Jallysson Miranda Rocha, 2018.

esse novo conjunto de dados é produzir e avaliar modelos classificadores para análise de influência do contexto de discurso nos resultados de segmentação das fases do gesto. Estratégias de treino e teste e parametrização discutidas na seção 5.1 foram usadas nesse experimento. Todas as análises realizadas neste experimento são provenientes de modelos gerados a partir dos vídeos rotulados pelo especialista E1 (rotulação **R1**). Essa rotulação foi escolhida por ter sido a que permitiu análises mais consistências no experimento do tipo #2.

A quantidade de possibilidades de combinação de gesticuladores, sessões e contextos de discurso proporcionada pelo novo conjunto de dados foi limitada pela quantidade de vídeos rotulados que foi possível produzir durante a execução do trabalho. Porém, mesmo com uma quantidade de vídeos limitada, foi necessário estabelecer uma estratégia de experimentação que permitiu comparar os resultados dos modelos classificadores nos diferentes contextos de discurso. Um esquema gráfico referente às combinações exploradas é apresentado no apêndice D.

A fim de organizar a apresentação dos resultados do experimento do tipo #3, o restante dessa seção segue dividido da seguinte forma:

- Análise #3.1: apresentação de resultados obtidos com os primeiros modelos de SVM, que exploraram o espaço de parâmetros completo e geraram a escolha do subespaço de parâmetros usados para construção dos demais modelos. Esses resultados foram obtidos com o uso do vídeo obtido com o gesticulador G1, no contexto contação de histórias³.
- Análise #3.2: apresentação de resultados obtidos em modelos que foram treinados e testados em um mesmo vídeo, considerando três novos vídeos, referentes aos contextos

³ Esse foi o vídeo que também foi explorado no experimento do tipo #2.

improvisação e entrevista. Vídeos de dois gesticuladores (G1 e G4) são usados nessa análise.

- Análise #3.3: apresentação de resultados obtidos em modelos que foram treinados em um vídeo e testados em outro vídeo. 3 combinações diferentes com quatro vídeos foram usadas nesta análise. Essas combinações envolveram os vídeos de entrevista (primeiro bloco), improvisação (primeiro tema) e contação de histórias (segunda história) do gesticulador G1 e o vídeo de entrevista (primeiro bloco) do gesticulador G4.
- Análise #3.4: apresentação de resultados obtidos em modelos que foram treinados em dois vídeos diferentes e testados em um terceiro diferente. Os contextos envolvidos nas combinações foram primeiro bloco da entrevista (gesticulador G1 e G4), primeiro tema de improvisação (gesticulador G1) e contação de histórias (segunda história para o gesticulador G1).
- Análise #3.5: apresentação de resultados obtidos em modelos que foram treinados em três vídeos diferentes e testados em um quarto diferente. Os contextos envolvidos nas combinações foram primeiro bloco de entrevista (G1 e G4), primeiro tema de improvisação (G1) e contação de histórias (segunda história de G1).

Análise #3.1

Inicialmente, como resultado da combinação dos demais parâmetros descritos na seção 5.1 no passo 1. Os resultados de *F-score* dos modelos criados foram analisados por meio de um procedimento que selecionou os modelos que alcançaram $F\text{-score} \geq 0,6$ em pelo menos duas fases do gesto (cf. tabela 22), considerando-os como os mais consistentes para suportar a escolha do subespaço de parâmetros a ser usado na construção dos demais modelos necessários para análise pretendida nesse trabalho.

O resumo apresentado na tabela 22 mostra que nenhum dos modelos gerados foi capaz de alcançar o *F-score* mínimo para mais do que três fases do gesto. Além disso, o período de descanso e a fase de *stroke* são frequentes nesse processo, indicando que a segmentação de fases pouco representadas como *hold*, e de transição (preparação e retração) configuram uma grande dificuldade para a análise automática das fases do gesto. Esses resultados mostram também que a extração da velocidade escalar a partir de um sinal

Tabela 22 – Resultados gerais do processo de filtragem dos melhores modelos classificadores. Legenda: Esp - espacial; Vet vet - velocidade vetorial; Vel esc - velocidade escalar; Ace vet - aceleração vetorial; Ace esc - aceleração escalar; D, P, H, S, R – descanso, preparação, *hold*, *stroke* e retração

Característica	<i>F-score</i> mínimo	Fases	Modelos (quantidade)
Esp sem filtro	0,7	D, S, R	8
Esp com filtro	0,7	D, S, R	8
Vel vet sem filtro	0,67	D, R	13
Vel vet com filtro	0,67	D, R	13
Vel esc sem filtro	0,6	D, P	6
Vel esc com filtro	0,73	D, P, S	15
Ace vet sem filtro	0,6	D, S	3
Ace vet com filtro	0,62	D, S	14

Fonte: Jallysson Miranda Rocha, 2018.

que sofre a ação de suavização do filtro *wavelet* gerou a representação mais discriminante neste experimento.

Diante desses resultados, o uso o não do sinal suavizado com aplicação do filtro *wavelet* não pode ser determinada. Para definir a continuidade do uso do filtro *wavelet* nos próximos experimentos, foram analisados os valores de *F-score* em cada uma das fases do gesto e nos períodos de descanso, obtidos nos melhores modelos para cada característica selecionada no procedimento de seleção dos melhores modelos. Na tabela 23 são apresentados os valores máximos e mínimos de *F-score* para cada fase do gesto e para o período de descanso obtido nos modelos selecionados para cada característica.

Os valores em destaque (cor verde) na tabela 23 indicam os melhores *F-scores* obtidos para cada fase quando houve diferença de valores para os resultados com e sem filtro, considerando o *F-score* mínimo estabelecido e cada uma das características em análise. Esses destaques mostram que há uma melhora no desempenho dos classificadores quando o sinal é suavizado com a aplicação do filtro. Pesa em favor do uso do filtro, os resultados obtidos nos melhores modelos construídos usando a características de velocidade escalar. Nesse conjunto de modelos, inclusive os valores mínimos de *F-scores* obtidos estão acima de 0,7. Outro bom resultado foi também obtido nos modelos que foram construídos sobre a representação com coordenadas espaciais. E neste último caso, o uso de filtro não levou à melhorias de desempenho de classificação. No entanto, pesa em favor da escolha do uso de velocidade escalar com filtro a quantidade maior de bons modelos obtidos (15 modelos cf. tabela 22).

Tabela 23 – Valores máximos e mínimos de F -score obtidos nos melhores modelos de cada característica em cada fase do gesto e no período de descanso. Legenda: Esp - espacial; Vet vet - velocidade vetorial; Vel esc - velocidade escalar; Ace vet - aceleração vetorial; Ace esc - aceleração escalar; max e min – máximo e mínimo

Característica	Descanso		Preparação		Hold		Stroke		Retração	
	max	min	max	min	max	min	max	min	max	min
Esp sem filtro	0,79	0,77	0,17	0,15	0,00	0,00	0,75	0,71	0,75	0,71
Esp com filtro	0,79	0,77	0,17	0,15	0,00	0,00	0,75	0,71	0,74	0,70
Vel vet sem filtro	0,72	0,67	0,22	0,16	0,02	0,00	0,63	0,51	0,69	0,67
Vel vet com filtro	0,71	0,68	0,39	0,23	0,05	0,00	0,58	0,52	0,70	0,67
Vel esc sem filtro	0,71	0,70	0,64	0,60	0,01	0,00	0,46	0,38	0,04	0,00
Vel esc com filtro	0,75	0,73	0,79	0,73	0,00	0,00	0,84	0,75	0,17	0,00
Ace vet sem filtro	0,67	0,65	0,19	0,15	0,10	0,00	0,62	0,61	0,07	0,04
Ace vet com filtro	0,78	0,72	0,35	0,16	0,06	0,00	0,65	0,62	0,31	0,15
Ace esc sem filtro	0,65	0,64	0,03	0,00	0,06	0,00	0,55	0,11	0,00	0,00
Ace esc com filtro	0,76	0,72	0,00	0,00	0,00	0,00	0,51	0,18	0,04	0,00

Fonte: Jallysson Miranda Rocha, 2018.

Enfim, diante das análises delineadas, foi possível reduzir o espaço de parâmetros a serem usados nos demais experimentos. Esses dados estão resumidos na tabela 11, mas seguem detalhados nas tabelas 24 e 25 em uma organização que permite analisar em detalhes as combinações de parâmetros de maior sucesso. O conjunto de parâmetros selecionado são aqueles usados nos melhores modelos para: coordenadas espaciais sem uso de filtro e velocidade escalar sem uso de filtro.

Tabela 24 – Parâmetros usados nos melhores modelos de SVM obtidos com representação espacial sem uso de filtro. Legenda: C - margem; TJ - tamanho de janela; PR - posição do rótulo da janela

C	Parâmetro de <i>kernel</i>	Ponto articulação	TJ	PR
1	0,0625	mão direita	40	meio
100	0,0625	mão direita	40	meio
500	0,0625	mão direita	40	meio
1000	0,0625	mão direita	40	meio
1	0,03125	mão direita e pulso direito	40	meio
100	0,03125	mão direita e pulso direito	40	meio
500	0,03125	mão direita e pulso direito	40	meio
1000	0,03125	mão direita e pulso direito	40	meio

Fonte: Jallysson Miranda Rocha, 2018.

Os modelos dos quais os parâmetros das tabelas 24 e 25 foram extraídos, mostram que determinados valores para alguns parâmetros são frequentes ou unânimes. A posição do quadro de interesse no meio da janela, e o uso de dados provenientes da mão dominante

Tabela 25 – Parâmetros usados nos melhores modelos de SVM obtidos com representação por velocidade escalar sem uso de filtro. Legenda: C - margem; TJ - tamanho de janela; PR - posição do rótulo da janela

C	Parâmetro de <i>kernel</i>	Ponto articulação	TJ	PR
500	16	mão direita e pulso direito	40	meio
1000	16	mão direita e pulso direito	40	meio
500	16	mão direita e pulso direito	50	meio
1000	8	mão direita e pulso direito	50	meio
1000	16	mão direita e pulso direito	50	meio
500	16	mão direita, pulso direito e cotovelo direito	40	meio

Fonte: Jallysson Miranda Rocha, 2018.

do gesticulador (no caso do vídeo usado para esse experimento), são escolhas unânimes. O tamanho da janela em 40 quadros aparece para os dois tipos de representação, embora para representação por velocidade, o tamanho 50 também seja adequado. Em relação ao parâmetro da função *kernel*, cada um dos tipos de representação exige uma faixa específica. E no caso da margem, valores altos foram os melhores para a representação por velocidade.

Análise #3.2

Essa análise proporcionou verificar o desempenho de modelos classificadores treinados com os parâmetros selecionados na análise #3.1 em dois contextos considerando o mesmo gesticulador, e para outro gesticulador considerando um contexto comum com aqueles analisados do primeiro gesticulador. Desta forma, três vídeos estão envolvidos na análise apresenta nesta seção⁴. Esta análise, portanto, permite a comparação do desempenho do classificador mediante o seguinte ambiente de teste:

- V1.11.1 - (**G1_C1.1_S1**) - G1: primeiro gesticulador; C1: contexto entrevista (primeiro bloco); S1: primeira sessão;
- V1.21.1 - (**G1_C2.1_S1**) - G1: primeiro gesticulador; C2: contexto improvisação (primeira improvisação); S1: primeira sessão;
- V4.11.1 - (**G4_C1.1_S1**) - G4: quarto gesticulador; C1: contexto entrevista (primeiro bloco); S1: primeira sessão;

Os resultados em termos de *F-score* máximo e mínimo obtidos com os modelos treinados e testados no mesmo vídeo (considerando os três vídeos) esta mostrado na

⁴ A rotulação usada para esse experimento é a do especialista E1.

tabela 26. A tabela é separada em duas partes: na primeira parte estão os resultados obtidos usando a representação espacial; na segunda parte estão os resultados obtidos usando a representação por velocidade escalar.

Tabela 26 – Valores máximos e mínimos de F -score de cada vídeo entre os melhores modelos nos dados espaciais e velocidade escalar em cada fase do gesto e período de descanso. Legenda: Esp - espacial; Vel esc - velocidade escalar; max - máximo valor; min - mínimo valor

Vídeo - Esp	Descanso		Preparação		Hold		Stroke		Retração	
	max	min	max	min	max	min	max	min	max	min
V1.11.1	0,83	0,79	0,39	0,3	0,07	0,0	0,61	0,49	0,53	0,17
V1.21.1	0,47	0,45	0,11	0,0	0,27	0,12	0,62	0,58	0,11	0,03
V4.11.1	0,63	0,57	0,24	0,09	0,0	0,0	0,37	0,05	0,29	0,16

Vídeo - Vel esc	Descanso		Preparação		Hold		Stroke		Retração	
	max	min	max	min	max	min	max	min	max	min
V1.11.1	0,82	0,82	0,25	0,19	0,0	0,0	0,42	0,39	0,18	0,1
V1.21.1	0,52	0,48	0,22	0,14	0,36	0,21	0,73	0,66	0,16	0,0
V4.11.1	0,58	0,57	0,17	0,07	0,0	0,0	0,03	0,03	0,13	0,02

Fonte: Jallysson Miranda Rocha, 2018.

Nesse experimento já é possível fazer algumas análises referentes à relação do desempenho dos classificadores aplicados a vídeos de gesticulação realizada em contextos diferentes. Os vídeos V1.11.1 e V4.11.1 são provenientes da gesticulação realizada durante uma entrevista realizada com dois usuários diferentes. Nestes vídeos, os períodos de descanso são mais bem reconhecidos pelos classificadores, principalmente considerando a representação espacial. No entanto, não há bom desempenho para as demais fases.

A mudança do contexto, representada pelo vídeo V1.21.1, para um contexto no qual há mais protagonismo do gesticulador, levou a um desempenho melhor para o reconhecimento da fase expressiva do gesto - o *stroke*. Esse destaque ocorre para os dois tipos de representação.

Em geral, o desempenho dos modelos não pode ser considerado adequado, e assim, a discussão aqui delineada precisa ser tomada como asserções iniciais que devem ser investigadas de forma mais aprofundada tanto em relação à quantidade de vídeos explorados quanto em relação ao treinamento dos modelos.

Análise #3.3

Outro tipo de análise possível é realizada mediante o treinamento do modelo em um vídeo e teste no outro vídeo. A capacidade de generalização dos modelos é colocada em teste. Contudo, no caso deste experimento, há uma vantagem para o aprendizado uma vez que um vídeo completo, e portanto uma quantidade maior de quadros de vídeo, é usado no treinamento. Os vídeos utilizados nesse experimento foram: V1.11.1, V1.21.1, V4.11.1 e o vídeo usado na análise #3.1 - V1.32.1 - gesticulador G1, contexto de contação de história (segunda história) e S1: primeira sessão. Os resultados obtidos para a combinação de treinamento no vídeo V1.11.1 e teste nos demais são apresentados na tabela 27.

A principal constatação observada é que as fases de preparação, *hold* e retração não são generalizadas pelos modelos. Na análise #3.2 já havia um baixo rendimento para essas fases, e nesta análise o pequeno rendimento é perdido. O aprendizado se torna mais consistente para os períodos de descanso e para a fase de *stroke*. No caso do período de descanso, o desempenho é melhor no vídeo de mesmo contexto. Isso se deve principalmente ao fato desses resultados terem sido obtidos com características espaciais – nas entrevistas os gesticuladores mantiveram posições de descanso muito similares. Já para o *stroke*, prevaleceu a intensidade de movimento, que aumenta nos contextos de contação de histórias e improvisação.

Tabela 27 – Valores máximos e mínimos de F -score do treinamento dos modelos em V1.11.1 e teste nos outros vídeos nos dados espaciais em cada fase do gesto e período de descanso. Legenda: Esp - espacial; max - máximo valor; min - mínimo valor

Vídeo - Esp	Descanso		Preparação		Hold		Stroke		Retração	
	max	min	max	min	max	min	max	min	max	min
V1.21.1	0,0	0,0	0,0	0,0	0,0	0,0	0,52	0,52	0,0	0,0
V1.32.1	0,0	0,0	0,0	0,0	0,0	0,0	0,53	0,53	0,0	0,0
V4.11.1	0,64	0,12	0,24	0,13	0,0	0,0	0,33	0,23	0,26	0,13

Fonte: Jallysson Miranda Rocha, 2018.

O padrão apresentado na tabela 27 se repete quando o vídeo V4.11.1 é usado para treinamento dos modelos, e inverte quando os vídeos V1.11.1 ou V1.21.1 são usados para treinamento. Neste último caso, as fases com F -score em 0,0 aparecem para os vídeos do contexto de entrevista. Em todas as combinações, o desempenho dos classificadores para a fase de *stroke* é semelhante.

O uso da representação em velocidade escalar, neste experimentos, produz resultados diferentes. Com essa representação, o período de descanso é mais bem representado e o melhor desempenho de reconhecimento se dá com o teste no vídeo de mesmo contexto que o vídeo de treino. Já a fase de *stroke* é mais bem reconhecida nos testes com os vídeos de mesmo usuário, não sendo o contexto um fator positivo para o desempenho do classificador. Esses resultados podem ser observados na tabela 28

Tabela 28 – Valores máximos e mínimos de F -score do treinamento dos modelos em V1.11.1 e teste nos outros vídeos nos dados representados por velocidade escalar em cada fase do gesto e período de descanso. Legenda: Vel esc - velocidade escalar; max - máximo valor; min - mínimo valor

Vídeo - Vel esc	Descanso		Preparação		<i>Hold</i>		<i>Stroke</i>		Retração	
	max	min	max	min	max	min	max	min	max	min
V1.21.1	0,49	0,48	0,22	0,16	0,01	0,0	0,52	0,41	0,11	0,06
V1.32.1	0,72	0,68	0,25	0,2	0,08	0,0	0,61	0,56	0,15	0,08
V4.11.1	0,83	0,82	0,0	0,0	0,09	0,0	0,25	0,18	0,08	0,06

Fonte: Jallysson Miranda Rocha, 2018.

Análise #3.4

Neste experimento, dois vídeos são usados no treinamento dos modelos, e o teste é realizado em um terceiro. O treinamento dos modelos, neste caso, recebe uma quantidade maior de quadros e um conjunto de padrões mais diversificado.

As tabelas 29 e 30 apresentam os melhores resultados de segmentação das fases do gesto para os modelos executados sobre vídeos representados via características descritivas espaciais e via velocidade escara, respectivamente. Estas tabelas são divididas em quatro partes sendo que cada uma delas diz respeito ao teste em um vídeo em especial. Em cada uma das linhas é apresentado o resultado obtido mediante modelos treinados com a combinação de dois vídeos, conforme informado na primeira coluna das tabelas. Os melhores resultados estão destacados na cor verde.

De forma geral, a quantidade de vídeos e combinação de contextos possível dentro dessa quantidade não permitiu análises variadas e aprofundadas, mas algumas observações podem ser feitas.

Em geral, não foram percebidas melhoras nos resultados de segmentação em relação às análises realizadas com treino e teste no mesmo vídeo, considerando as coordenadas

Tabela 29 – Valores máximos e mínimos de F -score do treinamento dos modelos em coordenadas espacial e teste nos vídeos com a linha em destaque de cinza para cada parte da tabela. Legenda: Esp - espacial; max - máximo valor; min - mínimo valor

Vídeo de teste: V4.11.1										
Treino - Esp	Descanso		Preparação		Hold		Stroke		Retração	
	max	min	max	min	max	min	max	min	max	min
V1.11.1 e V1.21.1	0,64	0,11	0,24	0,13	0,0	0,0	0,33	0,23	0,26	0,13
V1.11.1 e V1.32.1	0,64	0,12	0,24	0,12	0,0	0,0	0,33	0,23	0,26	0,13
V1.21.1 e V1.32.1	0,0	0,0	0,0	0,0	0,0	0,0	0,37	0,37	0,0	0,0
Vídeo de teste: V1.32.1										
Treino - Esp	Descanso		Preparação		Hold		Stroke		Retração	
	max	min	max	min	max	min	max	min	max	min
V1.11.1 e V1.21.1	0,72	0,56	0,2	0,12	0,34	0,3	0,62	0,48	0,23	0,15
V1.11.1 e V4.11.1	0,0	0,0	0,0	0,0	0,0	0,0	0,53	0,53	0,0	0,0
V1.21.1 e V4.11.1	0,81	0,69	0,18	0,11	0,38	0,31	0,62	0,58	0,21	0,13
Vídeo de teste: V1.21.1										
Treino - Esp	Descanso		Preparação		Hold		Stroke		Retração	
	max	min	max	min	max	min	max	min	max	min
V1.11.1 e V1.32.1	0,67	0,25	0,18	0,14	0,4	0,36	0,59	0,54	0,2	0,18
V1.11.1 e V4.11.1	0,0	0,0	0,0	0,0	0,0	0,0	0,52	0,52	0,0	0,0
V1.32.1 e V4.11.1	0,69	0,45	0,17	0,13	0,4	0,34	0,6	0,56	0,2	0,09
Vídeo de teste: V1.11.1										
Treino - Esp	Descanso		Preparação		Hold		Stroke		Retração	
	max	min	max	min	max	min	max	min	max	min
V1.21.1 e V1.32.1	0,0	0,0	0,0	0,0	0,0	0,0	0,3	0,29	0,0	0,0
V1.21.1 e V4.11.1	0,86	0,47	0,36	0,07	0,02	0,0	0,46	0,28	0,21	0,03
V1.32.1 e V4.11.1	0,69	0,45	0,17	0,13	0,4	0,34	0,6	0,56	0,2	0,09

Fonte: Jallysson Miranda Rocha, 2018.

espaciais. Para os resultados obtidos com a representação espacial foi observado um padrão referente à incapacidade de reconhecimento de períodos de descanso, e nas fases de preparação, *hold* e retração, em alguns casos. Essa incapacidade chama a atenção principalmente porque ocorre em um padrão que é o de mais fácil reconhecimento no problema tratado nesta dissertação - o descanso. O destaque a este comportamento se deve ao fato dele ocorrer quando os dois vídeos usados no treinamento são referentes a contextos com um padrão de gesticulação diferente do vídeo no qual os modelos são testados. De forma mais específica, esse comportamento ocorre quando:

- treino com os vídeos de contação de histórias e improvisação - teste no vídeo de entrevista (primeira e última partes da tabela 29);
- treino com os vídeos de entrevista - teste no vídeo de entrevista (segunda e terceira partes da tabela 29);

Nesses mesmos modelos, há alguma capacidade de reconhecimento para a fase de *stroke*, o que indica que a fase expressiva do gesto, considerando coordenadas espaciais, manifesta um padrão que é menos sensível à mudança de contexto.

Além disso, quando vídeos de contextos distintos são usados no treinamento dos modelos, o desempenho no teste aumenta, chegando próximo dos resultados obtidos quando treino e teste são realizados no mesmo vídeo. Em especial, nos períodos de descanso, o poder de generalização aumenta.

Para os casos de teste com vídeos do contexto de contação de histórias e de improvisação, houve uma pequena melhora na capacidade de reconhecimento de *holds*.

Tabela 30 – Valores máximos e mínimos de *F-score* do treinamento dos modelos em velocidade escalar e teste nos vídeos com a linha em destaque de cinza para cada parte da tabela. Legenda: Vel esc - velocidade escalar; max - máximo valor; min - mínimo valor

Vídeo de teste: V4.11.1										
Treino - Vel esc	Descanso		Preparação		Hold		Stroke		Retração	
	max	min	max	min	max	min	max	min	max	min
V1.11.1 e V1.21.1	0,83	0,76	0,03	0,01	0,03	0,0	0,46	0,27	0,05	0,02
V1.11.1 e V1.32.1	0,84	0,82	0,04	0,0	0,1	0,0	0,38	0,25	0,09	0,01
V1.21.11 e V1.32.1	0,78	0,71	0,1	0,08	0,05	0,0	0,48	0,42	0,04	0,01
Vídeo de teste: V1.32.1										
Treino - Vel esc	Descanso		Preparação		Hold		Stroke		Retração	
	max	min	max	min	max	min	max	min	max	min
V1.11.1 e V1.21.1	0,76	0,71	0,27	0,23	0,09	0,04	0,7	0,66	0,06	0,04
V1.11.1 e V4.11.1	0,68	0,66	0,16	0,1	0,01	0,0	0,59	0,55	0,1	0,01
V1.21.11 e V4.11.1	0,71	0,7	0,16	0,07	0,12	0,02	0,67	0,66	0,03	0,0
Vídeo de teste: V1.21.1										
Treino - Vel esc	Descanso		Preparação		Hold		Stroke		Retração	
	max	min	max	min	max	min	max	min	max	min
V1.11.1 e V1.32.1	0,5	0,49	0,22	0,18	0,0	0,0	0,58	0,5	0,12	0,06
V1.11.1 e V4	0,48	0,47	0,15	0,11	0,0	0,0	0,46	0,4	0,06	0,02
V1.32.1 e V4	0,49	0,49	0,06	0,04	0,01	0,01	0,56	0,54	0,05	0,0
Vídeo de teste: V1.11.1										
Treino - Vel esc	Descanso		Preparação		Hold		Stroke		Retração	
	max	min	max	min	max	min	max	min	max	min
V1.21.1 e V1.32.1	0,86	0,83	0,28	0,21	0,1	0,05	0,45	0,42	0,12	0,02
V1.21.1 e V4.11.1	0,87	0,86	0,22	0,1	0,06	0,0	0,49	0,47	0,14	0,01
V1.32.1 e V4.11.1	0,86	0,86	0,11	0,03	0,0	0,0	0,49	0,46	0,08	0,01

Fonte: Jallysson Miranda Rocha, 2018.

Para as característica de velocidade escalar, o agrupamento dos vídeos no treinamento gerou algumas melhorias no desempenho dos classificadores no período de descanso

e na fase do gesto de *stroke*, quando comparados aos desempenhos dos modelos treinados e testados no mesmo vídeo.

No período de **descanso** os testes nos vídeos *V1.11.1* e *V4.11.1* tiveram aumentos de até 0,05 e 0,26 no seu *F-score* respectivamente. Para a fase de *stroke* os modelos aumentaram seu desempenho em até 0,07 no vídeo *V1.11.1* e 0,45 em *V4.11.1*.

O aumento no vídeo *V4.11.1* foi bastante superficial em relação ao aumento no desempenho do vídeo *V1.11.1*, uma das razões levantadas é dos vídeos serem de gesticuladores diferentes. Os vídeos de contação de histórias e de improvisação foram obtidos com o mesmo gesticulador que o vídeo *V1.11.1*, assim, aspectos do padrão de gesticulação do gesticulador pode estar contribuindo com a melhoria. Esse fato indica que há uma sensibilidade dos classificador em relação à generalizar para gesticuladores diferentes.

Analizando os testes agora em outros vídeos de contextos bem parecidos, que é improvisação e induz ao gesticulador contar uma história no vídeo *V1.21.1* e o contexto de contação de história em si no vídeo *V1.32.1*, o período de descanso não teve nenhum aumento no desempenho para o vídeo *V1.21.1* em termo de valores máximos, apenas um aumento bem superficial nos valores mínimos e ao mesmo tempo perdeu desempenho também na fase de *stroke*. Já o vídeo *V1.32.1* possuiu aumento de até 0,09 no período de descanso e aumento de até 0,09 na fase de *stroke*.

Um outro ponto a considerar é a queda do desempenho dos modelos classificadores na fase de *hold* para o vídeo *V1.21.1*. Como a fase de *hold* é muito semelhante ao período de descanso em termos de velocidade da mão, quando as mãos estão praticamente paradas, ficou evidenciado que os modelos não foram capazes de distinguir os *holds* e, fazer o treinamento com vídeos de contextos diferentes pode, muito provavelmente, ter levado à execução de *holds* em um contexto com padrão semelhante ao período de descanso de outro contexto.

Também em comparação aos modelos treinados e testados no mesmo vídeo, o único caso que levou a uma leve perda na capacidade do reconhecimento da fase de *stroke* foi usar vídeos de entrevista para treino e vídeo de improvisação no teste. Entretanto, trata-se de uma perda pequena (0,02), a qual deve ser investigada a fim de verificar a sua significância estatística.

Análise #3.5

A análise realizada nesse experimento também faz uso de resultados obtidos com modelos treinados em vários vídeos (três nesse caso), e testados em um quarto vídeo. Os resultados são mostrados nas tabelas 31 e 32 respectivamente para representação usando coordenadas espaciais e velocidade escalar. Nessas tabelas, o destaque em verde segue para os melhores resultados obtidos.

Tabela 31 – Valores máximos e mínimos de *F-score* do treinamento dos modelos em coordenadas espaciais e teste nos vídeos com a linha em destaque de cinza para cada parte da tabela. Legenda: Esp - espacial; max - máximo valor; min - mínimo valor

Vídeo de teste: V4.11.1										
Treino - Esp	Descanso		Preparação		Hold		Stroke		Retração	
	max	min	max	min	max	min	max	min	max	min
Outros vídeos	0,64	0,12	0,24	0,13	0,0	0,0	0,33	0,24	0,26	0,12
Vídeo de teste: V1.32.1										
Treino - Esp	Descanso		Preparação		Hold		Stroke		Retração	
	max	min	max	min	max	min	max	min	max	min
Outros vídeos	0,72	0,56	0,17	0,11	0,37	0,3	0,63	0,51	0,23	0,11
Vídeo de teste: V1.21.1										
Treino - Esp	Descanso		Preparação		Hold		Stroke		Retração	
	max	min	max	min	max	min	max	min	max	min
Outros vídeos	0,67	0,11	0,17	0,13	0,4	0,36	0,59	0,54	0,19	0,12
Vídeo de teste: V1.11.1										
Treino - Esp	Descanso		Preparação		Hold		Stroke		Retração	
	max	min	max	min	max	min	max	min	max	min
Outros vídeos	0,86	0,44	0,35	0,08	0,02	0,0	0,46	0,27	0,23	0,03

Fonte: Jallysson Miranda Rocha, 2018.

O comportamento observado é semelhante àquele obtido na análise #3.4: resultados melhores do que o obtido no treino e teste em um único vídeo são obtidos para o período de descanso tanto no uso de representação por coordenada espacial quanto no uso de representação por velocidade escalar; alguma melhora é observada no reconhecimento da fase de *hold* para os vídeos de contação de histórias e de improvisação, na representação com coordenadas espaciais; e resultados melhores na fase de *stroke* são observados na representação por velocidade para os vídeos de entrevista.

Sobre as fases de preparação e retração, foi possível perceber que a representação com velocidade não é adequada. Nesses experimentos com treinamento envolvendo vários

Tabela 32 – Valores máximos e mínimos de F -score do treinamento dos modelos em velocidade escalar e teste nos vídeos com a linha em destaque de cinza para cada parte da tabela. Legenda: Vel esc - velocidade escalar; max - máximo valor; min - mínimo valor

Vídeo de teste: V4.11.1										
Treino - Vel esc	Descanso		Preparação		Hold		Stroke		Retração	
	max	min	max	min	max	min	max	min	max	min
Outros vídeos	0,83	0,76	0,04	0,0	0,05	0,0	0,47	0,29	0,03	0,01
Vídeo de teste: V1.32.1										
Treino - Vel esc	Descanso		Preparação		Hold		Stroke		Retração	
	max	min	max	min	max	min	max	min	max	min
Outros vídeos	0,7	0,68	0,15	0,11	0,05	0,0	0,64	0,62	0,05	0,0
Vídeo de teste: V1.21.1										
Treino - Vel esc	Descanso		Preparação		Hold		Stroke		Retração	
	max	min	max	min	max	min	max	min	max	min
Outros vídeos	0,48	0,48	0,16	0,09	0,0	0,0	0,51	0,49	0,09	0,05
Vídeo de teste: V1.11.1										
Treino - Vel esc	Descanso		Preparação		Hold		Stroke		Retração	
	max	min	max	min	max	min	max	min	max	min
Outros vídeos	0,87	0,87	0,2	0,11	0,06	0,0	0,5	0,47	0,12	0,02

Fonte: Jallysson Miranda Rocha, 2018.

vídeos, houve uma queda de desempenho. Isso também mostra que há alguma dependência de gesticulador na identificação do padrão dessas fases.

5.3 Considerações finais

As análises delineadas neste capítulo tomaram como base um extenso trabalho de calibração de parâmetros referentes à representação dos gestos e também referentes à máquina de vetores suporte descritas no experimento do tipo #1.

Então, assumindo um espaço menor de geração de classificadores, rotulações diferentes para um mesmo contexto de gesticulação natural foram avaliadas sob aspectos quantitativos e qualitativos no experimento do tipo #2.

Finalmente, experimentos envolvendo diferentes contextos e estratégias de treino e teste de modelos classificadores foram realizados e analisados no experimentos do tipo #3.

Um resumo das principais constatações realizadas nesses experimentos e análises é apresentado na conclusão deste trabalho.

6 Conclusões

Neste trabalho de mestrado, a área de segmentação automática das fases do gesto na gesticulação natural foi explorada em termos da aplicação de aprendizado de máquina, mais especificamente, máquinas de vetores suporte, com o intuito de realizar um estudo sobre a interferência do contexto de gesticulação nos padrões de caracterização das fases e no desempenho dos modelos de automação da segmentação. Para isso, a tarefa de segmentação das fases foi modelada como um problema de classificação no qual cada quadro de um vídeo foi o objeto de análise e quatro fases de gestos e período de descanso compuseram as classes a serem preditas. A designação de uma sequência de quadros de vídeo para uma das classes configura-se como a identificação de uma fase do gesto.

O objetivo delineado neste estudo foi fundamentado no fato que diferentes pesquisadores consideram diferentes formas de perceber as fases do gesto e, além disso, aplicar a definição das fases em uma instância real do discurso é uma tarefa altamente subjetiva e, como defendido na literatura (RAMAKRISHNAN, 2011; RAMAKRISHNAN; NEFF, 2013), dependente do contexto de produção do discurso (ou da gesticulação que acompanha o discurso).

A busca pelo objetivo principal do trabalho dependeu, primeiramente, do alcance de dois objetivos específicos. O primeiro, o desenvolvimento de um ambiente de experimentação (um *corpus* de vídeos) especialmente projetado para análise de diferentes contextos de discurso foi descrito no capítulo 4 deste texto. O *corpus* produzido conta 89 vídeos, obtidos a partir da colaboração de sete gesticuladores, e conta com a representação de cinco contextos de discurso diferentes. Desses vídeos, 8 foram rotulados por dois especialistas e, considerando todas as representações vetoriais usadas e janelamentos, um total de 2.400 conjuntos de dados foi gerado (sendo que cada conjunto de dados diz respeito a um vídeo). Os conjuntos de dados são de domínio público e podem ser obtidos por meio do grupo de pesquisa ao qual este trabalho está atrelado.

O segundo objetivo específico, a definição de um protocolo de avaliação para os modelos classificadores no contexto de segmentação das fases do gesto, ou seja, seguindo a visão de um especialista da área, foi alcançado por meio da proposição do uso combinado de medidas de: avaliação de classificadores provenientes da área de aprendizado de máquina (apresentado no capítulo 2) , aplicação de coeficientes de concordância para determinação

da rotulação de fases do gesto mais adequada para o *corpus* de vídeo e avaliação de segmentação das fases a partir de erros internos, de transição (ou de borda), e de inserção, exclusão e substituição de fases (apresentados no capítulo 3). Complementa a definição desse protocolo, a sugestão de formas para combinação de vídeos de diferentes contextos dentro de uma estratégia de treinamento e teste de um modelo classificador (usada nos experimentos do tipo #3 apresentados no capítulo 5).

Finalmente, o objetivo principal do trabalho foi alcançado a partir da realização dos experimentos apresentados no capítulo 5. Esses experimentos envolveram: a exploração de um espaço de busca de parametrizações, tanto do SVM quanto da representação dos gestos; o estudo sistemático das rotulações geradas para os vídeos; a exploração dos modelos de classificação junto aos vídeos obtidos a partir de diferentes contextos de discurso.

6.1 Contribuições

Os resultados e análises produzidos nos experimentos do tipo #2 confirmaram que o contexto da gesticulação natural representa um desafio para o reconhecimento automático de padrões. Os classificadores mostraram desempenhos diferenciados sob alguns aspectos quando consideradas as diferentes rotulações. O impacto da subjetividade de rotulação é evidente nas transições entre as fases, que é, segundo os especialistas, o ponto de maior dificuldade de rotulação manual. A partir dos resultados obtidos levantou-se a hipótese de que o classificador possa ser usado como um parâmetro de qualidade de rotulação. Para além da questão da subjetividade, as diferentes rotulações permitiram também fundamentar a indicação de uso das fases de preparação e retração na demarcação de mudanças de posição de descanso e de *strokes* consecutivos, o que não é necessariamente considerado na prática de rotulação das fases do gesto. Os resultados obtidos neste experimento geraram a publicação do artigo¹:

- Rocha, J. M.; Feitosa, R. A.; Lima, C. A. M.; PERES, S. M.. Segmentação das fases do gesto com máquinas de vetores suporte: o impacto da subjetividade e de diferentes rotulações. In: Encontro Nacional de Inteligência Artificial e Computacional (ENIAC), Uberlândia. 2017, p. 1-12.

¹ Além deste artigo, o autor do presente trabalho apresentou sua proposta de dissertação no Workshop de Teses e Dissertações de Sistemas de Informação do Simpósio Brasileiro de Sistemas de Informação, em 2016; e devido à sua participação num grupo de pesquisa, também foi co-autor de um artigo correlato ao tema do presente trabalho, publicado em um evento internacional em 2018.

Os resultados e análises produzidos nos experimentos do tipo #3 ilustraram o comportamento dos modelos classificadores para diversas situações referentes ao uso de contextos de discursos diferentes nas estratégias de treino e teste. A partir do que foi observado, foi confirmada a dificuldade de segmentação das fases de preparação, retração e *hold*, já discutida em trabalhos anteriores. Diante do desempenho obtidos pelos modelos, conclusões podem ser delineadas para a segmentação do período de descanso e para a fase de *stroke*. Segundo as análises, é possível dizer que o padrão de gesticulação manifestado nos períodos de descanso sofrem interferência do contexto de discurso; já a caracterização da fase de *stroke* está mais relacionada ao gesticulador do que ao contexto de discurso. Desta forma, a hipótese estabelecida para este trabalho foi confirmada para o período de descanso e refutada para a fase de *stroke*.

Até onde foi possível verificar na literatura da área, este é o primeiro estudo que propôs uma análise sistemática e automatizada do problema da subjetividade de análise nas fases do gesto. Trata-se ainda de um estudo preliminar que precisa ser aprofundado e realizado em diferentes contextos de gesticulação natural. Contudo, o estudo já permitiu levantar algumas hipóteses.

6.2 Trabalhos futuros

O estudo realizado possui algumas limitações que precisam ser consideradas na continuidade desta pesquisa. A primeira limitação encontrada está no conjunto limitado de vídeos rotulados disponíveis para a realização de um estudo aprofundado no tema. O tempo necessário para realização das rotulações foi um impeditivo para que mais vídeos fossem preparados para serem submetidos à tarefas de classificação. A segunda limitação diz respeito ao desempenho alcançado pelos modelos classificadores. De forma geral, os desempenhos de segmentação podem ser melhorados a partir da aplicação de outros algoritmos de classificação e de outras estratégias de segmentação. Além disso, os desempenhos baixos em termos de *F-score* enfraquecem as conclusões obtidas, impedem que análises mais detalhadas possam ser realizadas. Como exemplo, a falta da análise referente à qualidade de segmentação das fases do gesto sob o ponto de vista do especialistas (erros internos, de transição, de inserção, exclusão e substituição) nos resultados do experimento do tipo #3 é devido ao baixo desempenho dos modelos obtidos.

Diante das limitações levantadas para esse trabalho, as seguintes ações são sugeridas para continuidade desta pesquisa:

- Uso de aprendizado supervisionado e de transferência de aprendizado para superar a falta das rotulações: há um trabalho contemporâneo ao presente estudo, executado no âmbito do mesmo grupo de pesquisa, que explorou a aplicação de Mapas Auto-Organizáveis (rede neural artificial de aprendizado não supervisionado) no problema de análise das fases do gesto. Neste trabalho foi observado que o aprendizado não-supervisionado pode gerar modelos de agrupamentos que podem ser usados na segmentação das fases do gesto. Porém, o nível de sucesso de fazer a segmentação via aprendizado supervisionado é ainda um problema em aberto. Alternativamente, técnicas de transferência de aprendizado são comumente aplicados em situações nas quais há a presença de dados rotulados e dados não rotulados, sendo portanto, uma possibilidade a ser estudada.
- Exploração de técnicas de aprendizado supervisionado a exemplo do que foi realizado no presente trabalho, porém considerando modelos hierárquicos, *ensembles*, misturas de especialistas ou aprendizado profundo - alternativas que potencializam as capacidades de classificadores e podem gerar resultados de segmentação melhores para execução de uma análise de interferência de contexto mais robusta. Além disso, a aplicação de testes estatísticos para verificação da significância do desempenho dos classificadores se faz também necessária para fortalecer as conclusões obtidas.

Para além das limitações do presente trabalho, porém fazendo uso de um recurso aqui obtido e considerado fora do escopo deste estudo, sugere-se o uso dos áudios associados aos vídeos para ajudar no processo de rotulação. Especialistas da área defendem que a rotulação é mais precisa se for feita com o máximo de informação possível. Até o presente momento, as pesquisas realizadas pelo grupo de pesquisa correlato a este estudo consideraram apenas as imagens dos quadros de vídeos para fazer a rotulação. Uma justificativa para isso é que a informação passada aos modelos classificadores seriam restritas a características extraídas dos quadros de vídeo. No entanto, um dos problemas associados à segmentação das fases do gesto via algoritmos de aprendizado supervisionado é justamente a subjetividade de rotulação. A minimização da subjetividade, ainda que com o uso de informação extra, pode ajudar na melhoria dos desempenhos dos classificadores e na melhoria do estudo da relação destes desempenhos com os contextos de discurso.

Referências²

- ALLWOOD, J. et al. The mumin coding scheme for the annotation of feedback, turn management and sequencing phenomena. Language Resources and Evaluation, Springer Netherlands, v. 41, n. 3, p. 273–287, 2007. Citado na página 12.
- ALVARES, A. R.; ROMAN, N. T. Agreecalc: Uma ferramenta para análise da concordância entre múltiplos anotadores. In: Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology. [S.l.: s.n.], 2013. p. 1–10. Citado na página 18.
- ARTSTEIN, R.; POESIO, M. Inter-coder agreement for computational linguistics. Computational Linguistics, MIT Press, v. 34, n. 4, p. 555–596, 2008. Citado 4 vezes nas páginas 18, 33, 43 e 44.
- BHAME, V.; SREEMATHY, R.; DHUMAL, H. Vision based hand gesture recognition using eccentric approach for human computer interaction. In: IEEE. Advances in Computing, Communications and Informatics (ICACCI, 2014 International Conference on. [S.l.], 2014. p. 949–953. Citado na página 34.
- BRUGMAN, H.; RUSSEL, A. Annotating multimedia/multi-modal resources with elan. In: Proceedings of the 4th International Conference on Language Resources and Evaluation. [S.l.: s.n.], 2004. p. 2065–2068. Citado na página 12.
- DARDAS, N. H.; SILVA, J. M.; SADDIK, A. E. Target-shooting exergame with a hand gesture control. Multimedia tools and applications, Springer, v. 70, n. 3, p. 2211–2233, 2014. Citado na página 34.
- DENG, L. Y. et al. Real-time hand gesture recognition by shape context based matching and cost matrix. JNW, v. 6, n. 5, p. 697–704, 2011. Citado na página 34.
- FAWCETT, T. An introduction to roc analysis. Pattern recognition letters, Elsevier, v. 27, n. 8, p. 861–874, 2006. Citado na página 21.
- GARNHAM, A. Psycholinguistics: Central Topics. [S.l.]: Routledge, 1994. Citado na página 27.
- GIBBON, D. Gesture theory is linguistics: On modelling multimodality as prosody. In: PACLIC. [S.l.]: Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation, 2009. p. 9–18. Citado na página 29.
- GIBBON, D. et al. A computational model of arm gestures in conversation. In: In Proceedings of Eurospeech. [S.l.: s.n.], 2003. p. 813–816. Citado na página 12.
- HAYKIN, S. Multilayer perceptrons. Neural networks: a comprehensive foundation, Prentice Hall New Jersey, v. 2, p. 156–255, 1999. Citado na página 25.
- HUSSAIN, S. M. A.; RASHID, A. H.-u. User independent hand gesture recognition by accelerated dtw. In: IEEE. Informatics, Electronics & Vision (ICIEV), 2012 International Conference on. [S.l.], 2012. p. 1033–1037. Citado na página 34.

² De acordo com a Associação Brasileira de Normas Técnicas. NBR 6023.

JACOB, M. G.; WACHS, J. P. Context-based hand gesture recognition for the operating room. Pattern Recognition Letters, Elsevier, v. 36, p. 196–203, 2014. Citado na página 34.

KENDON, A. Gesticulation and speech: Two aspects of the process of utterance. The relationship of verbal and nonverbal communication, v. 25, p. 207–227, 1980. Citado 3 vezes nas páginas 12, 27 e 28.

KENDON, A. Gesture: Visible action as utterance. [S.l.]: Cambridge University Press, 2005. Citado na página 28.

KETTEBEKOV, S. Exploiting prosodic structuring of coverbal gesticulation. In: . [S.l.]: Proceedings of the 6th international conference on Multimodal interfaces, 2004. p. 105–112. Citado na página 28.

KETTEBEKOV, S. et al. Prosody based co-analysis of deictic gestures and speech in weather narration broadcast. In: . [S.l.]: Workshop on Multimodal Resources and Multimodal System Evaluation, 2002. p. 57–62. Citado na página 28.

KETTEBEKOV, S.; YEASIN, M.; SHARMA, R. Prosody based audiovisual coanalysis for coverbal gesture recognition. IEEE transactions on multimedia, IEEE, v. 7, n. 2, p. 234–242, 2005. Citado na página 28.

KIPP, M. Multimedia annotation, querying, and analysis in anvil. In: _____. Multimedia Information Extraction. [S.l.]: John Wiley e Sons, Inc., 2012. cap. 19, p. 351–367. Citado na página 12.

KITA, S.; GIJN, I. V.; HULST, H. Van der. Movement phases in signs and co-speech gestures, and their transcription by human coders. In: International Gesture Workshop. [S.l.]: I. Wachsmuth and M. Frohlich, eds., Gesture and Sign Language in Human-Computer Interaction, Lecture Notes in Computer Science,, 1998. p. 23–35. Citado 2 vezes nas páginas 13 e 28.

KITCHENHAM, B. Procedures for performing systematic reviews. Keele, UK, Keele University, v. 33, n. 2004, p. 1–26, 2004. Citado na página 17.

KRIPPENDORFF, K. Computing krippendorff’s alpha reliability. Departmental papers (ASC), 2007. Citado 2 vezes nas páginas 32 e 33.

KUROKI, K. et al. A remote conversation support system for deaf-mute persons based on bimanual gestures recognition using finger-worn devices. In: Pervasive Computing and Communication Workshops (PerCom Workshops), 2015 IEEE International Conference on. [S.l.]: IEEE, 2015. p. 574–578. Citado na página 34.

KWOK, J. T.-Y. Support vector mixture for classification and regression problems. In: IEEE (Ed.). Pattern Recognition. [S.l.]: Proceeding 13th International Conference on Pattern Recognition - ICPR, 1998. p. 255–258. Citado na página 23.

LIANG, Y.; MINGZHI, D.; YING, D. A high-performance training-free approach for hand gesture recognition with accelerometer [j]. Multimedia tools and applications, v. 1, n. 72, p. 843–864, 2013. Citado na página 34.

LICHMAN, M. UCI Machine Learning Repository. 2013. Disponível em: <<http://archive.ics.uci.edu/ml>>. Citado na página 56.

LIU, J.; KAVAKLI, M. A survey of speech-hand gesture recognition for the development of multimodal interfaces in computer games. In: Multimedia and Expo (ICME). [S.l.]: IEEE International Conference on, 2010. p. 1564–1569. Citado 2 vezes nas páginas 12 e 27.

LORENA, A. C.; CARVALHO, A. C. de. Uma introdução às support vector machines. Revista de Informática Teórica e Aplicada, v. 14, n. 2, p. 43–67, 2007. Citado 2 vezes nas páginas 23 e 24.

LU, Z. et al. A hand gesture recognition framework and wearable gesture-based interaction prototype for mobile devices. IEEE transactions on human-machine systems, IEEE, v. 44, n. 2, p. 293–299, 2014. Citado na página 34.

MADEO, R. Máquinas de Vetores Suporte e a Análise de Gestos: incorporando aspectos temporais. Dissertação (Mestrado) — Universidade de São Paulo, 2013. Citado 9 vezes nas páginas 30, 33, 46, 47, 49, 50, 56, 58 e 97.

MADEO, R. C. B.; LIMA, C. A. M.; PERES, S. M. Studies in automated hand gesture analysis: an overview of functional types and gesture phases. Language Resources and Evaluation, v. 51, n. 2, p. 547–579, 2017. Citado na página 27.

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. International Journal of Computer Science and Information Technology, v. 5, n. 4, p. 1–20, 2013. Citado na página 27.

MARASOVIĆ, T.; PAPIĆ, V. User-dependent gesture recognition on android handheld devices. In: IEEE. Software, Telecommunications and Computer Networks (SoftCOM), 2014 22nd International Conference on. [S.l.], 2014. p. 398–402. Citado na página 34.

MARICCHIOLO, F.; GNISCI, A.; BONAIUTO, M. Coding hand gestures: A reliable taxonomy and a multi-media support. In: _____. Cognitive Behavioural Systems. [S.l.]: Springer Berlin Heidelberg, 2012. p. 405–416. Citado na página 12.

MARTELL, C. H. An extensible, kinematically-based gesture annotation scheme. In: _____. Advances in Natural Multimodal Dialogue Systems. [S.l.]: Springer Netherlands, 2005. cap. Form, p. 79–95. Citado na página 12.

MATSUBARA, T.; MORIMOTO, J. Bilinear modeling of emg signals to extract user-independent features for multiuser myoelectric interface. IEEE Transactions on Biomedical Engineering, IEEE, v. 60, n. 8, p. 2205–2213, 2013. Citado na página 34.

MCCLEARY, L.; VIOTTI, E. Sign-gesture symbiosis in brazilian sign language narrative. Meaning, form, and body, Fey Parrill, Vera Tobin and Mark Turner eds, p. 181–201, 2009. Citado na página 26.

MCNEILL, D. Hand and mind: What gestures reveal about thought. [S.l.]: University of Chicago press, 1992. Citado 2 vezes nas páginas 12 e 27.

MCNEILL, D. Gesture and thought. [S.l.]: University of Chicago Press, 2008. Citado na página 29.

MEYER, D.; WIEN, F. T. Support vector machines. R News, v. 1, n. 3, p. 23–26, 2001. Citado na página 24.

MICHALSKI, R. S. A theory and methodology of inductive learning. In: _____. Machine Learning: An Artificial Intelligence Approach. [S.l.]: Springer Berlin Heidelberg, 1983. p. 83–134. Citado na página 20.

MITCHELL, T. M. Machine learning. Burr Ridge, McGraw Hill, v. 45, p. 37, 1997. Citado na página 20.

MITRA, S.; ACHARYA, T. Gesture recognition: A survey. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, v. 37, n. 3, p. 311–324, 2007. Citado na página 27.

MOLINA, J. et al. Real-time user independent hand gesture recognition from time-of-flight camera video using static and dynamic models. Machine vision and applications, Springer, v. 24, n. 1, p. 187–204, 2013. Citado na página 34.

MONI, M.; ALI, A. S. Hmm based hand gesture recognition: A review on techniques and approaches. In: Computer Science and Information Technology. [S.l.]: 2nd IEEE International Conference on Computer Science and Information Technology, 2009. p. 433–437. Citado na página 27.

OKADA, S. et al. Context-based conversational hand gesture classification in narrative interaction. In: ACM. Proceedings of the 15th ACM on International conference on multimodal interaction. [S.l.], 2013. p. 303–310. Citado na página 34.

PANTIC, M. et al. Human computing and machine understanding of human behavior: a survey. In: Artificial Intelligence for Human Computing. [S.l.]: Springer, 2007. p. 47–71. Citado na página 26.

PAULSON, B.; CUMMINGS, D.; HAMMOND, T. Object interaction detection using hand posture cues in an office setting. International journal of human-computer studies, Elsevier, v. 69, n. 1-2, p. 19–29, 2011. Citado na página 34.

PICKERING, C. A. The search for a safer driver interface: a review of gesture recognition human machine interface. Computing Control Engineering Journal, v. 16, n. 1, p. 34–40, 2005. Citado 2 vezes nas páginas 12 e 27.

POWERS, D. M. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. Journal of Machine Learning Technologies, Bioinfo Publications, v. 2, n. 1, p. 37–63, 2011. Citado na página 21.

QUEK, F. et al. Multimodal human discourse: gesture and speech. ACM Transactions on Computer-Human Interaction (TOCHI), v. 9, n. 3, p. 171–193, 2002. Citado na página 27.

RAMAKRISHNAN, A. S. Segmentation of hand gestures using motion capture data. Dissertação (Mestrado) — University of California, Davis, 2011. Citado 5 vezes nas páginas 30, 33, 49, 50 e 82.

RAMAKRISHNAN, A. S.; NEFF, M. Segmentation of hand gestures using motion capture data. In: Proceedings of the 2013 international conference on Autonomous agents and multi-agent systems. [S.l.: s.n.], 2013. p. 1249–1250. Citado 3 vezes nas páginas 30, 33 e 82.

- REFICE, M. et al. Automatic classification of gestures: a context-dependent approach. In: Computer Science and Information Systems (FedCSIS), 2011 Federated Conference on. [S.l.]: IEEE, 2011. p. 743–750. Citado na página 34.
- SCHÖLKOPF, B.; SMOLA, A. J. Learning with kernels: support vector machines, regularization, optimization, and beyond. [S.l.]: MIT press, 2002. Citado na página 23.
- SEMMLOW, J. Biosignal and biomedical image processing: Matlab-based applications. DOI, v. 10, 2004. Citado na página 97.
- SHEN, J. et al. Gpu-based realtime hand gesture interaction and rendering for volume datasets using leap motion. In: IEEE. Cyberworlds (CW), 2014 International Conference on. [S.l.], 2014. p. 85–92. Citado na página 34.
- SOWA, T. The recognition and comprehension of hand gestures-a review and research agenda. In: Modeling Communication with Robots and Virtual Humans. [S.l.]: Springer, 2008. p. 38–56. Citado na página 27.
- SUYKENS, J. A. K. et al. Least squares support vector machines. World Scientific, 2002. Citado na página 23.
- VAPNIK, V. Statistical Learning Theory. 1998. Citado 2 vezes nas páginas 23 e 24.
- VAPNIK, V.; LERNER, A. Pattern recognition using generalized portrait method. Automation and remote control, v. 24, p. 774–780, 1963. Citado na página 23.
- WILSON, A. D.; BOBICK, A. F.; CASSELL, J. Recovering the temporal structure of natural gesture. In: Automatic Face and Gesture Recognition. [S.l.]: Proceedings of the Second International Conference on, 1996. p. 66–71. Citado na página 28.
- ZHANG, X. et al. A framework for hand gesture recognition based on accelerometer and emg sensors. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, IEEE, v. 41, n. 6, p. 1064–1076, 2011. Citado na página 34.
- ZHU, C.; SHENG, W. Wearable sensor-based hand gesture and daily activity recognition for robot-assisted living. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, IEEE, v. 41, n. 3, p. 569–573, 2011. Citado na página 34.

Apêndice A – Termo de consentimento livre e esclarecido

TERMO DE CONSENTIMENTO LIVRE E ESCLARECIDO

Você está sendo convidado (a) para participar da pesquisa intitulada **Construção de corpus - vídeos com gesticulação natural** sob a responsabilidade dos pesquisadores Jallysson Miranda Rocha e Sarajane Marques Peres.

Nesta pesquisa nós estamos buscando coletar **conjunto de vídeos** referentes a um ou mais indivíduos realizando gesticulação natural em diferentes momentos, documenta-los e converte-los em formato numérico para disponibilização pública. O objetivo da coleta dos vídeos é permitir a análise de gestos de uma mesma pessoa em contextos diferentes.

O Termo de Consentimento Livre e Esclarecido será obtido pelo pesquisador Jallysson Miranda Rocha, que iniciará a coleta dos vídeos após aprovação no comitê de ética, tendo como local de coleta o laboratório de Inteligência Artificial do grupo de pesquisa GrIA na Escola de Artes, Ciências e Humanidades da Universidade de São Paulo (EACH-USP). A coleta poderá ocorrer nos períodos da manhã, tarde ou noite a depender da disponibilidade do participante (gesticulador), sendo necessário que esses participem de pelo menos duas sessões de captação de vídeo por contexto, e que elas ocorram em períodos diferentes do dia (manhã, tarde ou noite), no mesmo dia ou em dias diferentes.

Na sua participação você deverá contar uma história baseada em uma “tirinha” que lhe será mostrada, ou improvisar uma história sobre um tema dentre algumas opções que lhe serão fornecidas, ou descrever uma cena dentre algumas opções que lhe serão fornecidas, ou participar de uma entrevista sobre seu cotidiano na EH ou participar de um diálogo sobre um tema dentre algumas opções que lhe serão fornecidas, ou proferir uma aula – sendo de sua escolha dois ou mais contextos de participação. Sua fala ocorrerá em frente a um ou dois aparelhos *Kinect* da *Microsoft*, e você poderá permanecer em pé ou sentar-se, como preferir. Sua fala e sua gesticulação serão gravadas em um vídeo, por meio do qual serão identificadas as posições de suas mãos, pulsos, tronco e cabeça, que por sua vez serão gravadas na forma numérica. Também

será gravado o áudio de sua fala. Esse estudo tem interesse em criar um conjunto de dados (planilha numérica) para um ambiente de experimentação de estudos de análise de gestos naturais. Apenas os responsáveis por este projeto terão acesso às suas imagens e ao som de sua fala (gravados em vídeo). Alguns quadros dos vídeos podem ser usados para documentação e publicação, e neste caso, a face do participante será descaracterizada. Em nenhum momento você será identificado. Os resultados da pesquisa serão publicados e ainda assim a sua identidade será preservada.

Você não terá nenhum gasto e ganho financeiro por participar na pesquisa.

Os riscos mais prováveis consistem em **estresse emocional e fadiga, devido à necessidade de permanecer falando e gesticulando por um período de 1 a 5 minutos**. Os benefícios deste projeto **não envolvem diretamente os participantes, porém é um estudo que visa melhorar a qualidade da análise dos gestos em contexto de gesticulação natural, contribuindo com áreas de pesquisa como a Linguística e a Interação Humano-Computador**.

Você é livre para deixar de participar da pesquisa a qualquer momento sem nenhum prejuízo ou coação.

Uma via original deste Termo de Consentimento Livre e Esclarecido ficará com você.

Qualquer dúvida a respeito da pesquisa, você poderá entrar em contato com: **Jallysson Miranda Rocha, telefone: (11) 948149135; Sarajane Marques Peres, telefone: (11)98711-0994 ou (11)30918897; Rua Arlindo Bettio, 1000 - CEP: 03828 – 000, Vila Guaraciaba, São Paulo-SP**. Poderá também entrar em contato com o Comitê de Ética na Pesquisa com Seres-Humanos – Escola de Artes, Ciências e Humanidades da Universidade de São Paulo: **Rua Arlindo Bettio, 1000 - CEP: 03828-000, Vila Guaracicaba, São Paulo-SP**.

São Paulo, 07 de Fevereiro de 2016.

Assinatura dos pesquisadores

Eu aceito participar do projeto citado acima, voluntariamente, após ter sido devidamente esclarecido.

Participante da pesquisa

Apêndice B – Roteiro de atividades dos participantes

Primeira ação	Gestculador - 1		Gestculador - 2		Gestculador - 3	
	Sessão	Repertir captura	Sessão	Repertir captura	Sessão	Repertir captura
Apresentar o objetivo da captura de dados	1		1		1	
Pegar assinaturas nos documentos	1		1		1	
Segunda ação	Gestculador - 1		Gestculador - 2		Gestculador - 3	
	Sessão	Repertir captura	Sessão	Repertir captura	Sessão	Repertir captura
Posicionar o gesticulador	1..6		1..8		1..6	
Explicar que eventualmente o sensor falha mas ele não precisa se preocupar com isso	1		1		1	
Terceira ação	Gestculador - 1		Gestculador - 2		Gestculador - 3	
	Sessão	Repertir captura	Sessão	Repertir captura	Sessão	Repertir captura
Fazer a entrevista com bloco 1 e 2	1	Sessão - 3	1	Sessão - 2	1	Sessão - 1
Fazer a entrevista com bloco 2 e 3	2	Sessão - 4	1	Sessão - 2	2	Sessão - 2
Quarta ação	Gestculador - 1		Gestculador - 2		Gestculador - 3	
	Sessão	Repertir captura	Sessão	Repertir captura	Sessão	Repertir captura
Tema 1 - Brincadeira de criança	1	Sessão - 1	1	Sessão - 5	1	Sessão - 3
Tema 2 - Situação inusitada em uma viagem	2	Sessão - 2	2	Sessão - 6	1	Sessão - 4
Tema 3 - Cena de um filme	3	Sessão - 3	3	Sessão - 7	2	Sessão - 3
Tema 4 - Situação inusitada no ambiente de trabalho	4	Sessão - 4	4	Sessão - 8	2	Sessão - 4
Quinta ação	Gestculador - 1		Gestculador - 2		Gestculador - 3	
	Sessão	Repertir captura	Sessão	Repertir captura	Sessão	Repertir captura
Tirinha 1 - Turna da Monica (corda)	1	Sessão - 2	1	Sessão - 1	1	Sessão - 4
Tirinha 2 - Os Simpsons (jardinagem)	1	Sessão - 3	2	Sessão - 2	2	Sessão - 5
Tirinha 3 - Homem Aranha (. e o lixo)	2	Sessão - 3	3	Sessão - 3	3	Sessão - 6
Sexta ação	Gestculador - 1		Gestculador - 2		Gestculador - 3	
	Sessão	Repertir captura	Sessão	Repertir captura	Sessão	Repertir captura
Cena 1 - filme (Namoro ou Liberdade)	1	Sessão - 4	1	Sessão - 2	1	Sessão - 1
Cena 2 - Jogo (Super Mario World)	2	Sessão - 5	1	Sessão - 3	2	Sessão - 2
Cena 3 - Quadro (O Estádio Olímpico)	3	Sessão - 6	2	Sessão - 3	3	Sessão - 3
Sétima ação	Gestculador - 1		Gestculador - 2		Gestculador - 3	
	Sessão	Repertir captura	Sessão	Repertir captura	Sessão	Repertir captura
Conversar um pouco com o gesticulador sobre assuntos aleatórios	1..6		1..8		1..6	

Apêndice C – Roteiro de perguntas do contexto de entrevista

Construção de corpus - vídeos com gesticulação natural

Pesquisador responsável: Jallysson Miranda Rocha

Orientadora participante: Dr. (a). Sarajane Marques Peres

Campus/Curso: Escola de Artes, Ciências e Humanidades / Programa de pós-graduação em Sistemas de Informação

Local da coleta de dados: Escola de Artes, Ciências e Humanidades da Universidade de São Paulo

Entrevista com os participantes

Bloco 1 - perguntas:

- 1 - Você é aluno (professor) da graduação ou da pós-graduação?
- 2 - Qual curso você faz (trabalha)?
- 3 - Qual a data de seu ingresso na EACH?
- 4 - Como foi o início de seu curso (trabalho) aqui na EACH? A escola e seu curso corresponderam às suas expectativas?
- 5 - O que você mais gosta no curso que você escolheu cursar (trabalhar)? Nos fale mais sobre isso.

Bloco 2 - perguntas:

- 1 - Precizou mudar de endereço para estudar (trabalhar) aqui?
- 2 - Já conhecia outros alunos (professores) que estudam ou estudavam (trabalham ou trabalhavam) aqui antes de você vir para cá?
- 3 - Está participando de alguma atividade acadêmica extra-curricular (extra) no momento? Se sim, qual ou quais são? Se não, se motiva a participar de alguma? Qual?
- 4 - Já participou de alguma outra entrevista aqui na faculdade? Se sim, conte-nos como foi.
- 5 - Tem interesse em cursar outro curso após concluir o atual (trabalhar em alguma área diferente da atual)? Nos fale um pouco sobre suas motivações para isso.

Bloco 3 - perguntas:

- 1 - Você tem alguma sugestão para incrementar/melhorar o nosso campus? Qual?
- 2 - Você costuma fazer refeições aqui no campus? Se sim, onde? Na cantina ou no restaurante universitário?
- 3 - Como você costuma vir para a EACH? De transporte público ou particular? Enfrenta dificuldades no trânsito?
- 4 - Onde você costuma estudar aqui na EACH e como você se organiza nesse ambiente de estudo (como você se organiza em sua sala de trabalho)?

Apêndice D – Esquema gráfico de combinações de vídeos

Apêndice E – Filtro *Wavelet*

É importante minimizar os ruídos nos dados, obtidos durante a captura. Uma das técnicas que pode ser utilizada é a transformada *Wavelet*. Essa transformada descreve um sinal no qual a frequência se altera no decorrer do tempo. A aplicação do filtro pode ser realizada em sinais contínuos ou discretos. Para sinais contínuos, específico desse projeto, a transformada *Wavelet* pode ser obtida por:

$$W(a, b) = \int_{-\infty}^{\infty} s(t) \frac{1}{\sqrt{|a|}} \psi * \left(\frac{t - b}{a} \right) dt,$$

em que $*$ indica a operação de conjugação complexa, b representa uma função sobre $s(t)$ e a varia a escala de tempo de uma função de sondagem ψ , uma função é somada ou multiplicada visando modificar o sinal $s(t)$ (MADEO, 2013). O termo $\frac{1}{\sqrt{|a|}}$ é um fator de normalização que garante que a energia seja a mesma para todos os valores de a e b . Os coeficientes de *Wavelet* $W(a, b)$ descrevem a correlação entre o sinal original e o sinal com várias translações e escalas, ou seja, a similaridade entre o sinal e o mesmo sinal em determinada combinação de escala a e oposição b (SEMMLOW, 2004).

O filtro gera dois coeficientes de aproximação (cA) e de detalhe respectivamente (cD).