Análise das fases do gesto: a influência do contexto na execução das fases e na indução de modelos de aprendizado de máquina supervisionado

Resumo

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 ajeitar 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 melhorara 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 descanso, preparação, *stroke*, *hold* e retração. Nesse contexto 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 foi aplicado 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 pela pesquisa aqui proposta, a qual objetiva estudar a variabilidade do padrão inerente à cada uma das fases do gesto quando a manifestação delas se dá a partir de um mesmo indivíduo, porém em diferentes contextos de produção do discurso.

Abstract

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 mes-

sage 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 Psycholinguistics, 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: rest, preparation, stroke, hold and retraction. In this context, 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 research proposed herein, whose purpose is to study the variability of the inherent pattern of each gesture phase when its execution is made by the same person, but in different contexts.

1. Introdução

Recentemente, a área de pesquisa em análise de gestos tem recebido uma forte atenção de diferentes setores da sociedade. Muitas pesquisas na área de análise de gestos, manuais ou não manuais[1] tem focado no desenvolvimento de novos métodos para interação humano computador, suportando a criação de aplicações das mais diversas naturezas. Jogos (Liu e 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 como interessados em análise de gestos. Algumas linhas de pesquisa da Linguística e da Psicolinguística, que se preocupam em interpretar gestos, fazem uso de ferramentas computacionais (Gibbon et al, 2003; Martell, 2002; Maricchiolo et al, 2012; Allwood et al, 2004; Kipp,

[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.

2012; Brugman e Russel, 2004) para suportar o seu trabalho de análise de discurso baseado em gestos, ou baseado em gestos coocorrente com a fala.

Dentre as diferentes possibilidades de estudo dos gestos está aquela que presume que os gestos são compostos por unidades gestuais e estas, por sua vez, compostas por fases gestuais (Kendon, 1980). Diferentes pesquisadores consideram diferentes formas de perceber essas fases do gesto e, além disso, aplicar a definição das fases em uma instância real do discurso (segmentar as fases do gesto) é uma tarefa altamente subjetiva e, como defendido na literatura (Ramakrishnan, 2011), dependente do contexto de produção do discurso, o que constitui a principal motivação para o projeto aqui proposto: **compreender se de fato diferentes contextos influenciam na análise automática de gestos manuais e como se dá essa influência, se ela existir**.

Neste projeto, a segmentação automática das fases do gesto é tratada como uma tarefa de aprendizado de máquina supervisionado, considerando aspectos temporais inerentes ao padrão a ser aprendido (como em Madeo (2013)[2], onde a técnica Máquinas de Vetores Suporte foi aplicada), a subjetividade de rotulação do conjunto de dados a ser usado na indução do modelo (como em Wagner et al (2014), onde a técnica Multilayer Perceptron foi aplicada) e a variabilidade do contexto onde os dados foram obtidos – variável onde está a contribuição pretendida para o presente projeto.

Com os estudos realizados até o momento, pelo grupo de pesquisa correlato a esta proposta, já foi possível constatar que o aprendizado indutivo tem potencial para realizar a tarefa de segmentação das fases do gesto com uma eficiência promissora, conforme mostrado na Tabela 1.

Tabela 1: Alguns resultados obtidos para segmentação das fases do gestos usando SVM (Madeo, 2013) e MLP (Wagner et al, 2014)[3].

Referência	Problema resolvido	Vetor de Características	Fase do gesto	F-score
(Madeo, 2013)	Segmentação de unidades gestuais (dependente do contexto e do usuário)	Velocidade e aceleração das mãos e dos pulsos	Unidade gestual	0,87
	Segmentação das fases do gesto (dependente do contexto e do usuário)		Descanso	0,91
			Preparação	0,75
			Stroke	0,79
Wagner et al. (2014)	Segmentação de unidades gestuais (dependente do contexto e do usuário)	Velocidade e aceleração das mãos e dos pulsos; coordenada x,y,z das mãos e pulsos.	Unidade gestual	0,92

[2]Dissertação de mestrado realizada com apoio da FAPESP (Processo 2011/04608-8).

[3]O entendimento pleno dos números apresentados nessa tabela depende da menção a muitas variáveis e restrições referentes aos experimentos e análises realizados nos trabalhos citados.

Porém, os testes se concentraram em um único domínio de discurso (contação de histórias[4]) e faz-se necessário analisar como se dará a eficiência da abordagem em domínios de discurso onde os gestos possuem uma natureza diferente (gesticulação durante a conversação natural ou gesticulação de professor durante o exercício de ministrar uma aula, por exemplo); e como deve ser tratada a subjetividade da tomada de decisão sobre a análise do desempenho do classificador, uma vez que codificadores humanos também inserem subjetividade no processo.

A fim de melhor esclarecer as motivações para o presente projeto e o planejamento de execução do mesmo, o restante deste documento é organizado da seguinte forma: os objetivos do projeto bem como a hipótese a ser estudada são apresentados na Seção 2; a fundamentação teórica correlata ao domínio de aplicação e ao domínio técnico do projeto, motivando a realização do projeto e fornecendo base para entendimento de como o problema será resolvido, está resumida na Seção 3; o método que se pretende implementar para alcançar os objetivos, incluindo a forma de análise dos resultados, cronograma de execução e resultados esperados, é apresentada na Seção 4; e na sequência as referências bibliográficas são listadas.

2. Objetivos e Hipótese

O objetivo deste projeto é estudar 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; esse estudo está baseado 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, alguns objetivos específicos são:

- desenvolver o ambiente de experimentação (um conjunto de dados ou uma série de conjuntos de dados) onde 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;

[4]Gesture Phases Database –conjunto de dados criado pelo grupo de pesquisa correlato a este projeto, e disponibilizado no UCI Repository (Lichman, 2013) - https://archive.ics.uci.edu/ml/datasets/Gesture+Phase+Segmentation

- definir um protocolo de avaliação do desempenho quantitativo e qualitativo 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.

A hipótese a ser verificada no projeto de pesquisa aqui proposto é:

> "O padrão de gesticulação natural, embora analisado em relação a um mesmo executor, é dependente do contexto onde 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 tem 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 do classificadores."

3. Fundamentação teórica

Esta seção segue organizada em quatro partes. Na primeira é apresentado um resumo dos conceitos fundamentais da área de análise de gestos, relacionados principalmente a problemas correlatos à área de Linguística e Psicolinguística. A segunda parte trata, resumidamente, da área de Aprendizado de Máquina, com atenção especial à técnica de Máquinas de Vetores Suporte (técnica escolhida para construção dos classificadores). Finalmente, na última parte, é fornecida uma visão geral da relação estabelecida entre Aprendizado de Máquina e o problema de segmentação das fases do gesto, definindo como o problema de segmentação é modelado para ser resolvido por um classificador e como o classificador deve ser avaliado.

a. Área de Aplicação: análise 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: a Linguística e a Psicolinguística. Nessas áreas aplica-se uma maneira singular de análise que é principalmente pautada na a análise de padrões espaço-temporais gerados pela realização de "movimentos" e "gestos"[5] e pela inserção e análise da significação destas variáveis no contexto comportamental (McCleary e 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 2010, foram publicadas algumas revisões de literatura em análise de gestos. Juntas, essas revisões cobrem um conjunto de estudos 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 and 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, Peres e Wagner (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.

[5]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.

O contexto do presente projeto se localiza na 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 et al. 1996) ou considerando seu conteúdo semântico (Kettebekov et al. (2002), Kettebekov(2004) e Kettebekov et al.(2005)). Alguns estudos, como os apresentados por Kendon (1980) e por Kita et al. (1998) são baseados no estudo da divisão dos gestos em fases. Veja uma taxonomia referente a elementos componentes de um gesto em McNeill (1992).

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 et al. (1998), é também sugerido que as fases *pré-stroke hold*,

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 et al. (1998) quanto Kendon (2005) consideram o conceito de frase gestual. Contudo, desde que não existe um consenso sobre este conceito[6], ele não é usualmente aplicado nas tarefas de segmentação de gestos.

b. Aprendizado de máquina e Máquinas de Vetores Suporte

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, onde 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-ia como chave de melhoria para métodos computacionais capazes de adquirir conhecimento. Formalmente, o AM indutivo assume um subconjunto **X,** de um universo de discurso χ ,formado por *m* veto-

[6]Kita et al. (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.

res de dados *n*-dimensionais $\boldsymbol{x}_i$(i = 1, ..., *m*), associados a um conjunto **C** de *k*rótulos c_j(j = 1 .. *k*) de acordo com uma função *F* desconhecida. Considerando *m* como uma amostra representativa do universo de discurso χonde o conjunto **X** ocorre, é possível realizar uma busca em um espaço de funções hipóteses **H** e encontrar uma função ***F'*** tal que $\boldsymbol{F'(X) \cong F(X)}$ e$\boldsymbol{F'(\chi) \cong F(\chi)}$.

Máquina de Vetores Suporte (SVM) 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 e 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), Tin-Yau(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 sobreajuste 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 VC (Lorena e Carvalho, 2007).

SVM é 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).

Considerando um conjunto de treinamento com **N** exemplos, definidos por $\{\boldsymbol{x}_i, y_i\}_{i=1}^{n}$onde $\boldsymbol{x}_i$ 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 $h(\boldsymbol{x}_i) = \boldsymbol{w}^T \varphi(\boldsymbol{x}_i) + b$, onde $\boldsymbol{w}$ é o conjunto ótimo de pesos, *b* é o bias ótimo, e φé o mapeamento não linear aplicado nas entradas. SVM otimiza o hiperplano maximizando as distâncias entre este hiperplano e os dados mais próximos ($\boldsymbol{x}_i$), que corresponde a minimizar $\boldsymbol{w}$usando, por exemplo[7]:

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

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

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

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

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

Aplicando o método Lagrangeano na equação (1), obtém-se

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

sujeito a

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

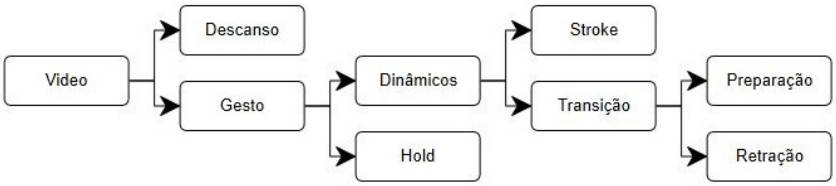
$$C \geq \alpha_i \geq 0,\quad i = 1, 2, \dots, N\ ,$$

onde $\boldsymbol{\alpha}$ são os multiplicadores de Lagrange. Enfim, resolvendo o problema da equação (2) é possível resolver o problema na equação (1), desde que $\boldsymbol{w}$ possa ser definido em termos de $\boldsymbol{\alpha}$ (Haykin, 1999). Na equação (2), uma função kernel$K(\boldsymbol{x_i}, \boldsymbol{x_j})$ pode ser usada para representar o produto interno $\langle\varphi(\boldsymbol{x_i})\cdot\varphi(\boldsymbol{x_j})\rangle$, executando um mapeamento não linear implícito.

c. Máquinas de Vetores Suporte e Segmentação Automática de Fases do Gesto

Em trabalhos anteriores, desenvolvidos no âmbito do grupo de pesquisa no qual essa proposta está inserida, o problema de segmentação automática das fases do gesto foi abordado como um problema dividido em problemas menores, modelados como tarefas de classificação e resolvidos usando SVM considerando aspectos temporais. O objeto de análise do classificador é um vídeo, representado por uma sequência de frames. O problema de segmentação de fases de gesto, então, consiste em receber a representação de um frame, ou de um conjunto de frames (quando uma abordagem baseada em representação por janelas deslizantes é aplicada) da sequência como entrada e classificá-lo como pertencente a uma das classes: posição de descanso, preparação, *stroke*, *hold* e retração. Contudo, para abordar o problema considerando uma complexidade gradual, o mesmo foi divido em subtarefas, como ilustrado pela Figura 1. Um gesto é inicialmente analisado por um classificador binário que rotula os frames do vídeo do gesto como "posição de descanso" ou "gesto". Posteriormente, outro classificador binário analisa o "gesto", rotulando seus frames como "frames dinâmicos" ou "*hold*", e assim por diante.

Figura1: Estratégia para classificação de fases de gestos.



Fonte: Madeo (2013)

A informação sobre o vídeo é captada via o dispositivo Xbox Kinect™, e a informação sobre o contexto dos frames é representada por meio de características referentes à velocidade, aceleração e coordenadas espaciais, organizadas em vetores que representam janelas do vetor de. O uso das janelas insere uma característica temporal na representação do gesto.

Alternativamente, classificadores binários construídos para analisar cada um dos sub-problemas (fases) podem ser construídos considerando a abordagem *oneXone* ou *oneXall* (Galar et al., 2011). Essas estratégias podem servir como uma alternativa à abordagem hierárquica minimizando o problema de propagação de erros de classificação de níveis mais altos para níveis mais baixos na hierarquia – como observado nos trabalhos anteriores (Madeo, 2013) desenvolvidos no contexto histórico em que se encaixa o presente projeto.

d. 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

$$Acurácia = 1 - \xi_{generalização}$$

$$\xi_{generalização}(f(\varkappa), D) = \sum_{<x,y> \in \varkappa} D(x,y) \times \wp(f(x), y)$$

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 e $\wp$ é uma função de perda binária

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

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 são derivadas a medida *F-score* a as curvas ROC (*Receiver Operating Characteristics*).

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 se seus frames 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 frame a frame e trata-se de uma abordagem baseada em aprendizado supervisionado, medidas para análise de desempenho de classificadores deveriam considerar que: (a) erros de classificação cometidos na transição entre fases devem ser analisados de forma diferenciada dos erros cometidos dentro de uma fase – gerando medidas de acurácia e matrizes de confusão ponderadas; (b) a comparação da resultado do classificador à segmentação realizada por um especialista deve ser feita de forma a considerar que a precisão da segmentação humana, assim como a sua consistência, está sujeita a imprecisões nas regiões de transições entre fases. Ramakrishna (2011) e Madeo (2013) adotaram como medida adicional de avaliação dos classificadores o número de seguimentos classificados corretamente, considerando que porcentagens de frames (consecutivos e não consecutivos) corretamente classificados dentro de uma fase. Ainda, Chen et al. (2004) analisar erros de segmentação em relação a erros de inserção e deleção, que no caso deste projeto significa encontrar uma fase onde ela não existe ou não encontrar uma fase onde ela existe.

4. Método

Este projeto se constitui como uma pesquisa experimental, que envolverá 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 serão realizados estudos exploratórios que visam a obtenção de informações sobre pesquisas correlatas ao tema deste projeto, as quais sustentarão a decisão sobre a realização (ou atualização) de levantamento refinado de bibliografia - potencialmente por meio de

revisão sistemática da literatura (Kitchenham, 2004). Esse levantamento deve ser referente a representação de dados gestuais, estratégias de análise de fases de gestos e avaliação de resultados de análise de gestos que não estejam ainda contemplados pelo arcabouço bibliográfico já constituído no grupo de pesquisa correlato a este projeto a saber: Madeo, Lima e Peres (2012), Madeo, Peres e Lima (2012), Madeo, Peres e Lima (submetido para revisão)[8].

A definição de um protocolo de coleta de dados ser faz 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. Essa definição deverá ser feita com apoio de especialistas em Linguística do Departamento de Letras da Faculdade de Filosofia, Letras e Ciências Humanas (FFLCH) da USP, já colaboradores do grupo de pesquisa correlato (Prof. Dr. Leland McCleary, Profa. Dra. Evani Viotti e Prof. Dr. Felipe Venâncio Barbosa). Esse protocolo deverá estabelecer condições para obtenção de dados, tais como: a restrição do espaço (3D) de movimentação do usuário; o contexto de comunicação; a afinidade do usuário com o contexto de comunicação; o período de captação de dados; o tempo de captação de dados.

A coleta de dados será feita com o suporte do sensor Kinect™ 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. O grupo de pesquisa dispõe de dois sensores Kinect, de forma que é possível obter dados a partir do sensoriamento de duas perspectivas espaciais diferentes.

Após a coleta de dados, esses serão organizados como um *corpus* e documentados em termos de meta-dados, de informações 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 deverão ser rotulados por pelo menos dois rotuladores, e análises de concordância (Artstein e Poesio, 2008; Alvares e Roman, 2013) serão realizadas sobre essa rotulação.

[8]Madeo, R. C. B, Peres, S. M e Lima, C. A. M. Studies in Automated Hand Gesture Analysis: An overview on functional types and gesture phases. Submetido para o periódico Language Resources and Evaluation .Atualmente está na segunda revisão.

A partir da existência do corpus rotulado, os dados deverão ser pré-processados e representados em termos de uma representação vetorial, de forma adequada para apresentação à técnica SVM. Um conjunto de experimentos com diferentes combinações de representação de dados, uso de janelas deslizantes e projeto de SVM (com variação de kernel e de parâmetros de kernel) será realizado para produção de classificadores.

Na sequência, os resultados dos classificadores serão avaliados, seguindo um protocolo de avaliação de desempenho quantitativa e qualitativo a ser definido como parte das contribuições esperadas deste projeto. Este protocolo fará uso das medidas de avaliação citadas na *Seção 3.d.* Os resultados dos classificadores mais bem avaliados sob tal protocolo serão analisados de forma a buscar a comprovação ou refutação da hipótese deste projeto.

5. Cronograma

Os cronogramas ilustrados abaixo trazem a distribuição das tarefas referentes a este projeto considerando um desenvolvimento em dois anos.

Ano 1												
Atividade	mês 1	mês 2	mês 3	mês 4	mês 5	mês 6	mês 7	mês 8	mês 9	mês 10	mês 11	mês 12
Estudos exploratórios												
Refinamento / atualização de levantamentos bibliográficos												
Definição de protocolo de coleta de dados												
Coleta de dados												
Organização e documento de dados												
Rotulação do corpus												
Implementações necessárias para experimentação												

Ano 2												
Atividade	mês 1	mês 2	mês 3	mês 4	mês 5	mês 6	mês 7	mês 8	mês 9	mês 10	mês 11	mês 12
Estudos exploratórios												
Refinamento / atualização de levantamentos bibliográficos												
Organização e documento de dados												
Rotulação do corpus												
Implementações necessárias para experimentação												
Execução dos experimentos												
Definição de protocolo de avaliação de qualidade de classificadores												
Execução da avaliação dos classificadores												
Análise dos dados produzidos pelos classificadores para comprovação ou refutação da hipótese												

6. Resultados esperados

A meta deste projeto é apresentar e validar uma metodologia de análise automática de gesticulação natural que permita avaliar a influência do contexto de comunicação sobre os padrões dos gestos produzidos pelo comunicador. A apresentação desta metodologia poderá suportar estudos na área de Linguística e Psicolinguísticas, que se baseiam em fases do gesto para elucidar questões inerentes ao sistema linguístico, como um sistema multimodal onde gestos representam um elemento provido de significado, de informação e de capacidade de comunicação.

A metodologia de análise automática de gestos estará pautada em técnicas de aprendizado de máquina supervisionado e requererá um estudo sobre diferentes formas de representação de dados e modelagem de classificadores. Tais realizações devem contribuir para a área de análise automática de gestos em geral. Uma vez que representações e estratégias de análise de gestos propostas para um fim específicos podem transitar entre diferentes domínios de aplicação, os progressos deste projeto se configuram como uma contribuição para a área de reconhecimento de padrões, mineração (análise) de dados e visual computacional.

Em termos de construção de artefatos, esse projeto deverá resultar em um corpus de dados gestuais documentado, dois protocolos (uma para coleta de dados e um para avaliação de classificadores) e no aprimoramento de ferramentas de captura de dados gestuais por meio do sensor Kinect e de segmentação automática das fases do gesto.

Em termos de publicações científicas, é esperada a inserção dos resultados dessa pesquisa em veículos de disseminação científica nas áreas de: reconhecimento de padrões, linguística computacional, aprendizado de máquina, visão computacional e sistemas de informação. Eventualmente, divulgação científica nas áreas de Linguística e Psicolinguística poderão também ser produzidas.

Referências Bibliográficas

Allwood J, Cerrato L, Dybkjaer L, Jokinen K, Navarretta C, Paggio P (2004) The mumin multimodal coding scheme. Tech. rep., University of Copenhagen, URL http://www.cst.dk/mumin/

Alvares, A. R., Roman, N. T. (2013)AgreeCalc: Uma Ferramenta para Análise da Concordância entre MúltiplosAnotadores. In: Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology, p. 1-10.

Artstein, R., and Poesio, M. (2008). Inter-coder agreement for computational linguistics. Computational Linguistics, vol. 34, n. 4, p. 555–596

Brugman H, RusselA (2004) Annotating multimedia/multi-modal resources with ELAN. In: Proceedings of the 4th International Conference on Language Resources and Evaluation, p. 2065-2068

Chen L, Liu Y, Harper MP, Shriberg E (2004) Multimodal model integration for sentence unit detection. In: Proceedings of the 6th international conference on Multimodal interfaces, ACM Press, New York, New York, USA, pp. 121-128.

Fawcett, T. (2006) An introduction to ROC analysis. Pattern Recognition Letters, vol. 27, pp. 861-874.

Galar, M., Fernández, A., Barrenechea, E., Herrera, F.. (2011) An overview of ensemble methods for binary classifiers in multi-class problems: experimental study on one-vs-one and on-vs-all schemes. Pattern Recognition, 44, p. 1761-1776.

Garnham, A. (1994). Psycholinguistics: Central Topics. In Routledge.

Gibbon D, Gut U, Hell B, Looks K, Thies A, Trippel T (2003) A computational model of arm gestures in conversation. In: Proceedings of Eurospeech, p. 813-816.

Haykin, S. (1999). Neural Networks: A Comprehensive Foundation. In Prentice Hall, Upper Saddle River, NJ, 2nd edition.

Kendon, A. (1980). Gesticulation and speech: Two aspects of the process of utterance. In M.R. Key, ed., The Relationship of verbal and nonverbal communication, p. 207-227. Mounton Publishers.

Kendon, A. (2005). Gesture: Visible Action as Utterance. In Cambridge University Press.

Kettebekov, S. (2004). Exploiting prosodic structuring of coverbal gesticulation. In Proceedings of the 6th International Conference on Multimodal Interfaces, p. 105-112.

Kettebekov, S., Yeasin, M., Krahnstoever, N., Sharma, R. (2002). Prosody based co-analysis of deictic gestures and speech in weather narration broadcast. In Proceedings of the Workshop on Multimodal Resources and Multimodal System Evaluation, p. 57-62.

Kettebekov, S., Yeasin, M., Sharma, R. (2005). Prosody based audiovisual coanalysis for coverbalgesture recognition. In IEEE Transactions on Multimedia, vol. 7, n. 2, p. 234-242.

Kipp M (2012) Multimedia Annotation, Querying, and Analysis in Anvil, JohnWiley & Sons, Inc., p. 351-367

Kita, S., van Gijn, I., van der Hulst, H. (1998). Movement phases in signs and co-speech gestures, and their transcription by human coders. In I. Wachsmuth and M. Frohlich, eds., Gesture and Sign Language in Human-Computer Interaction, Lecture Notes in Computer Science, vol. 1371, p. 23-35. Springer Berlin/Heidelberg.

Kitchenham, B. (2004). Procedures for performing systematic reviews. Technical Report TR/SE-0401, Keele UK Keele University.

Lichman, M. (2013) UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Liu, J.,Kavakli, M. (2010). A survey of speech-hand gesture recognition for the development of multimodal interfaces in computer games. In Proceedings of the IEEE International Conference on Multimedia and Expo, p. 1564-1569.

Lorena, A. C., Carvalho, A. C. P. L. F. (2007). Uma introdução às Support Vector Machines. In Revista de Informática Teórica e Aplicada, v. 14, n. 2.

Madeo C. B. R., Lima A. M. C., Peres, S. M. (2012). A Review on Temporal Reasoning using Support Vector Machines.In Proceedings of 19th International Symposium on Temporal Representation and Reasoning, p. 114-121, Leicester, UK.

Madeo C. B. R., Peres, S. M., Lima A. M. C. (2012). Overview on Support Vector Machines applied to Temporal Modeling. In Anais do IX Encontro Nacional de Inteligência Artificial, Curitiba, Brasil.

Madeo C. B. R. (2013) Máquinas de Vetores Suporte e a Análise de Gestos: incorporando aspectos temporais. Dissertação de mestrado – Universidade de São Paulo.

Madeo, R. C. B., Peres, S. M., Wagner, P. K. (2013) A review of temporal aspects of hand gestures analysis applied to discourse analysis and natural conversation. International Journal of Computer Science and Information Technology, v. 3. P 1-20.

Maricchiolo F, Gnisci A, Bonaiuto M (2012) Coding hand gestures: a reliable taxonomy and a multi-media support. In: Cognitive Behavioural Systems, Springer, p. 405-416

Martell C (2002) Form: An extensible, kinematically-based gesture annotation scheme. In: Proceedings of the 2002 International Conference on Language Resources and Evaluation, Istanbul, Turkey

McCleary, L. E.,Viotti, E. C. (2009). Sign-gesture symbiosis in Brazilian Sign Language narrative. In Fey Parrill, Vera Tobin and Mark Turner eds., Meaning, form, and body, p. 181-201, Stanford, CA. CSLI Publications.

McNeill, D. (1992). Hand and mind: What the hands reveal about thought. In University of Chicago Press.

Michalski, R. S.. (1983) A Theory and Methodology of Inductive Learning. Artificial Intelligence, p. 111-161.

Mitchell, T. M (1997).Machine Learning. In McGraw-Hill.

Mitra, S., Acharya, T. (2007). Gesture recognition: A survey. In IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, vol. 37, n. 3, p. 311-324.

Moni, M., Ali, A. B. M. S. (2009). HMM based hand gesture recognition: A review on techniques and approaches. In 2nd IEEE International Conference on Computer Science and Information Technology, p. 433-437.

Pantic, M., Pentland, A., Nijholt, A., Huang, T. (2007). Human computing and machine understanding of human behavior: a survey. In Lecture Notes in Artificial Intelligence, vol. 44512007, p. 47-71, Berlim. Springer-Verlag.

Pickering, C. (2005). The search for a safer driver interface: a review of gesture recognition human machine interface. In Computing Control Engineering Journal, vol. 16, n. 1, p. 34-40.

Powers, D. M. W. (2011) Evaluation: From Precision, Recall and F-measure to ROC, Informedness, Markedness& correlation. Journal of Machine Learning Technologies, v. 2, n. 1, pp 37-63.

Quek, F., McNeill, D., Bryll, R., Duncan, S., Ma, X., Kirbas, C., McCullough, K., Ansari, R. (2002). Multimodal human discourse: gesture and speech. In ACM Transactions on Computer-Human Interaction, vol. 9, n. 3, p. 171-193.

Ramakrishnan A. S. (2011) Segmentation of hand gestures using motion capture data. Master's thesis, University of California.

Schölkopf, B.,Smola, A. J. (2002). Learning with kernels: Support Vector Machines, Regularization, Optimization and Beyond. In MIT Press.

Sowa, T. (2008). The recognition and comprehension of hand gestures - a review and research agenda.In I. Wachsmuth and G. Knoblich, eds., Modeling Communication with Robots and Virtual Humans, Lecture Notes in Computer Science, vol. 4930, p. 38-56. Springer Berlin / Heidelberg.

Suykens, J. A. K., Van Gestel, T., De Brabanter, J., De Moor, B., Vanthienen, J. (2002).Least Squares Support Vector Machines. In World Scientific.

Tin-Yau, K. J. (1998). Support vector mixture for classification and regression problems. In Proceeding 13th International Conference on Pattern Recognition - ICPR, p. 255-258.

Vapnik, V. N. (1998). Statistical Learning Theory. In John Willey & Sons.

Vapnik, V.N., Lerner, A. (1963).Pattern recognition using generalized portrait method. In Automation and Remote Control, v. 24.

Wagner, P. K.,Peres, S. M., Madeo, R. C. B., Lima, C. A. M.,Freitas, F. A.. (2014) Gesture Unit Segmentation Using Spatial-Temporal Information and Machine Learning. In: 27th Florida Artificial Intelligence Research Society Conference (FLAIRS), 2014, Pensacola Beach, p. 101-106.

Wilson, A., Bobick, A., Cassell, J. (1996). Recovering the temporal structure of natural gesture. In Proc. of the Second International Conference on Automatic Face and Gesture Recognition, p. 66 –71.