UNIVERSIDADE DE SÃO PAULO

Escola de Artes, Ciências e Humanidades

Relatório Técnico PPgSI-002/2012

Análise Automatizada de Gestos: uma Revisão Sistemática considerando Aspectos Temporais

Renata C. B. Madeo
Sarajane M. Peres

Outubro - 2012

Série de Relatórios Técnicos

PPgSI-EACH-USP. Rua Arlindo Béttio, 1000 - Ermelino Matarazzo - 03828-000
São Paulo, SP.
TEL: (11) 3091-8197 http://www.each.usp.br/ppgsi

Análise Automatizada de Gestos: uma Revisão Sistemática considerando Aspectos Temporais

Renata C. B. Madeo , Sarajane M. Peres

[1]Escola de Artes, Ciências e Humanidades – Universidade de São Paulo
São Paulo – SP, Brasil

{renata.si, sarajane}@usp.br

Resumo. *Recentemente, houve um aumento no número de pesquisas sobre análise automatizada de gestos. Este aumento é justificado, dentre outros motivos, pela busca de métodos mais naturais para a Interação Humano-Computador. Além disso, a análise automatizada de gestos pode ser aplicada na construção de ferramentas que auxiliem a análise de gestos feitas nas áreas de Linguística e Psicolinguística, por exemplo. Este documento apresenta uma revisão sistemática sobre análise automatizada de gestos, focando nos aspectos temporais abordados por cada trabalho. Como resultado da revisão, é proposta uma organização sistemática dos trabalhos sobre análise de gestos de acordo com o tipo de análise realizada, métodos utilizados e aplicações.*

1. Introdução

Recentemente, muitos estudos foram realizados visando desenvolver novos métodos para a análise automatizada de gestos. Grande parte desses trabalhos objetivam a criação de novos métodos de Interação Humano-Computador baseados nos gestos ou que combinem gestos a outras modalidades de interação, como fala ou direção do olhar, visando criar médotos de interação multimodal. Essa busca por novos métodos de interação é motivada pelos recentes avanços tecnológicos, que tornam a interação por métodos tradicionais, como *mouse* e teclado, inadequada para determinados objetivos, como sistemas de computação ubíqua.

Neste contexto, a análise automatizada de gestos é geralmente usada para realizar o reconhecimento de um conjunto pré-definido de gestos simples ou ainda um conjunto reduzido de gestos retirado de uma Língua de Sinais, visando possibilitar a interação.

Outro uso possível para a análise automatizada de gestos, porém, é a criação de ferramentas que auxiliem a análise de gestos feita em outras áreas, como linguística e psicolinguística. As análises realizadas nestas áreas são importantes para os pesquisadores em Sistemas de Informação pois estudam tópicos que podem suportar o desenvolvimento de novos métodos de interação, como a sincronia entre gestos e fala, por exemplo. Neste caso, as análises realizadas são focadas especificamente na "gesticulação", ou seja, nos gestos realizados junto com a fala durante uma conversa.

Ao estudar a gesticulação, é possível perceber a importância dos aspectos temporais na análise dos gestos: durante uma conversa, utilizam-se vários tipos de gestos, sem aspectos espaciais (forma) convencionalizada Wilson et al. [1996]. Porém, mesmo esses gestos apresentam uma estrutura temporal definida – as fases dos gestos Kendon [1980], Wilson et al. [1996]. O aspecto temporal é destacado também em várias definições de gestos, que definem gestos como movimentos ou sequências. Alguns exemplos incluem a definição de Corradini [2002], na qual gestos são definidos como sequências de posturas

conectadas por movimentos durante um período de tempo; e a definição de Kim et al. [2007], na qual gestos são definidos como movimentos feitos por alguma parte do corpo com o objetivo de expressar intenções. Portanto, como levantado por Wilson et al. [1996], parece razoável considerar os aspectos temporais dos gestos.

Assim, este documento apresenta uma revisão sistemática sobre análise automatizada de gestos manuais, focando nos aspectos temporais considerados em cada trabalho. Além disso, destacam-se especialmente trabalhos aplicados à análise da conversação natural e do discurso, visando aumentar a variadade de aplicações capturadas pela revisão. A análise dos resultados da revisão sistemática pretende apresentar classificações de acordo com o tipo de análise realizada, métodos utilizados e aplicações.

O documento é organizado da seguinte forma: alguns conceitos fundamentais dos estudos sobre gestos realizados por pesquisadores de Linguística são brevemente apresentados na Seção 2; a Seção 3 apresenta a revisão sistemática realizada, descrevendo o processo de condução e discutindo os resultados; e a Seção 4 apresenta algumas considerações finais.

2. Alguns conceitos dos estudos sobre gestos

Para compreender os resultados da revisão sistemática é necessário conhecer alguns conceitos dos estudos sobre gestos, realizados na área de Linguística, como informações sobre os tipos de gestos [McNeill 1992], sobre as fases dos gestos [Kendon 1980, Kita et al. 1998] e o conceito de análise perceptual [McNeill 2005].

McNeill [1992] define cinco tipos de gestos como icônicos, metafóricos, dêiticos, coesivos e *beats*. Nesta revisão, alguns trabalhos focam na classificação de alguns tipos de *beats*, que são gestos que marcam o ritmo da fala e não possuem conteúdo semântico bem definido, de forma que seu valor semiótico é indicar pontos importantes da narrativa. *Beats* têm outra particularidade: esse tipo de gesto tem apenas duas fases do movimento – dentro/fora, cima/baixo, esquerda/direita, etc.

Segundo Kita et al. [1998], a estrutura temporal do gesto considera unidades gestuais, que corresponde ao período entre duas posições de descanso das mãos. Dentro das unidades gestuais, cada gesto é composto pelas fases de preparação (opcional), fase expressiva, e retração (opcional). A fase expressiva pode ser composta por um *stroke* – movimento que expressa o conteúdo semântico do gesto – cercado por duas breves pausas opcionais chamadas de *hold* dependente; ou por um *hold* independente, que ocorre quando o conteúdo semântico do gesto é expressado por uma pausa.

McNeill [1992] relaciona as fases dos gestos a uma hierarquia fonológica, na qual o *stroke* corresponde à sílaba proeminente da sentença. Essa relação é explorada em alguns artigos desta revisão, que consideram aspectos fonológicos para melhorar o reconhecimento dos gestos [Kettebekov et al. 2002, Kettebekov 2004, Kettebekov et al. 2005].

Tais conceitos são importantes no estudo da produção, compreensão e uso dos gestos. Para realizar tais estudos, é preciso gravar e transcrever vídeos, através da transcrição do discurso e dos gestos realizados, para que só então os pesquisadores realizem algum tipo de análise qualitativa sobre os dados coletados. Em geral, a transcrição dos gestos é um processo manual que utiliza ferramentas que permitem a reprodução do vídeo *frame* a *frame* para auxiliar o processo de transcrição. Porém, mesmo com o auxílio de fer-

ramentas, o processo, também conhecido como análise perceptual, é muito trabalhoso e demorado: Quek et al. [2002] afirmam que é possível levar até 10 dias para transcrever um minuto de vídeo. Por esse motivo, alguns trabalhos focam em extrair algumas características de forma automática para facilitar a transcrição dos vídeos. Nesta revisão, esses trabalhos são classificados segundo à sua aplicação na área de Análise Psicolinguística.

3. Revisão Sistemática sobre Análise de Gestos

A revisão sistemática sobre análise de gestos foi realizada segundo o protocolo de revisão apresentado no Apêndice A. Esta seção apresenta alguns detalhes sobre o processo de condução da revisão (Seção 3.1) e a análise dos resultados obtidos (Seção 3.2).

3.1. Condução

Após planejar a revisão e construir o protocolo, seguiu-se as especificações do protocolo para conduzir a revisão. Nesta revisão, a condução consiste em: submeter as *strings* de busca às máquinas de busca selecionadas e aplicar os critérios de seleção dos estudos relevantes para a revisão.

Neste caso, as buscas retornaram 57 estudos – 30 da ACM, 19 da IEEE, 26 da CiteSeerX e 8 da SpringerLink, com sobreposição de artigos encontrados em diferentes fontes. Após a seleção, restaram 28 artigos – 17 da ACM, 11 da IEEE, 12 da CiteSeerX e 4 da SpringerLink. O processo de revisão ocorreu de 04/10/2011 a 24/11/2011.

3.2. Análise dos Resultados da Revisão

Uma análise interessante para qualquer revisão sistemática consiste em verificar a evolução da área de pesquisa abordada na revisão. Na Figura 1, podemos notar que houve uma intensificação da pesquisa na área de análise de gestos nos últimos dez anos. É possível que esse crescimento tenha acontecido devido aos avanços na tecnologia, incluindo: a evolução da capacidade de processamento dos computadores, imprescindíveis à análise de gestos quando técnicas de visão computacional são utilizadas; e o desenvolvimento de sensores, uma alternativa à visão computacional para captação de dados. Ademais, o desenvolvimento de algumas outras áreas, como a computação ubíqua, faz com que novos métodos de interação humano-computador sejam necessários, e a interação através de gestos é uma opção para esse tipo de aplicação.

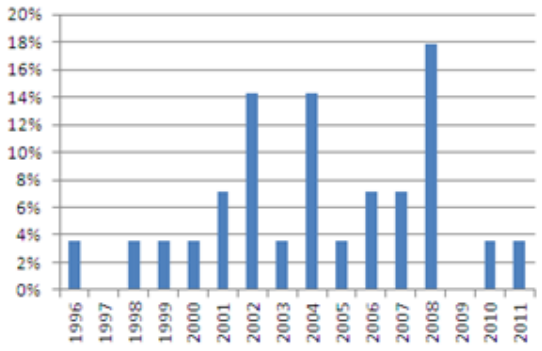

Figura 1. Distribuição dos trabalhos selecionados por ano.

Para este trabalho, a revisão foca em duas características específicas: a aplicação da análise de gestos na análise do discurso e da conversação natural; e os aspectos temporais

utilizados na análise de gestos. Entretanto, apesar do foco nessas características, também é objetivo desta revisão obter uma visão geral dos diversos tipos de aplicações e análises relacionadas.

A fim de dar maior legibilidade aos resultados produzidos nesta revisão, esta seção é organizada de forma a cobrir cada aspecto que foi analisado nos trabalhos encontrados, a saber: aplicações; métodos para aquisição de dados; métodos para extração e representação de características; métodos para análise de gestos, incluindo segmentação e reconhecimento; e métodos para análise de resultados de segmentação e reconhecimento.

3.2.1. Aplicações

Os resultados desta revisão mostram que a análise de gestos pode ser utilizada em diversas aplicações. Com o objetivo de propor uma sistematização sobre as aplicações de análise de gestos, é possível dividir os trabalhos encontrados nesta revisão em três grupos de aplicações:

Análise de um vocabulário pré-definido (AGPD):

Esses trabalhos definem um conjunto limitado de gestos para uma aplicação específica. A análise de um conjunto limitado de gestos pode ser utilizada para interação humano-computador em tarefas específicas, como o controle de uma casa inteligente [Kim et al. 2007], o reconhecimento de gestos para controle de aviação [Choi et al. 2008], ou o controle de uma apresentação de *slides* [Chen et al. 2011]. Alguns trabalhos aplicam sua análise a um conjunto de gestos simples, sem definir uma aplicação específica, visto que tais gestos poderiam ser utilizado para interação em vários contextos, como Wong e Cipolla [2006], Corradini [2002], Yuan [2008] e Li e Greenspan [2007]. Outros trabalhos visam à análise de um escopo limitado de uma língua de sinais: Lamar et al. [1999] reconhecem 34 gestos da datilologia japonesa; Werapan e Chotikakamthorn [2004] reconhecem 14 gestos da Língua de Sinais Taiwanesa; e Nguyen e Bui [2010] segmentam 23 gestos do alfabeto vietnamita.

Entretanto, como o foco desta revisão é a aplicação em problemas relacionados à análise de discurso e a análise de aspectos temporais da análise, o interesse nesses trabalhos consiste em analisar como os aspectos temporais são considerados nesse tipo de aplicação. Neste sentido, alguns trabalhos [Kim et al. 2007, Chen et al. 2011, Wong e Cipolla 2006, Li e Greenspan 2007, Werapan e Chotikakamthorn 2004] objetivam o reconhecimento de gestos contínuos, no qual é necessário realizar a tarefa de segmentação dos gestos no vídeo antes do reconhecimento, o que claramente considera aspectos temporais. Mesmo os trabalhos que realizam apenas reconhecimento consideram aspectos temporais, visto que utilizam uma definição de gestos que destaca seu aspecto temporal como uma sequência de posturas e incorporam essa definição à análise: Corradini [2002] utiliza HMM para captar a estrutura temporal dos gestos; Choi et al. [2008] usam, uma sequência de posturas para representar o gesto; e Yuan [2008] representa o gesto como uma sequência de vetores e utiliza algoritmos de aprendizado de máquina que consideram aspectos temporais para classificar os gestos. Ainda há trabalhos que consideram gestos estáticos. Os seguintes trabalhos entraram como resultado desta revisão porque, de alguma forma, também consideram aspectos temporais: Nguyen

e Bui [2010] realizam uma segmentação encontrando os gestos de transição entre os gestos estáticos; e Lamar et al. [1999] consideram uma normalização temporal para classificar 20 gestos dinâmicos e 14 gestos estáticos.

Análise de gesticulação (AG):

Trabalhos com análise de gesticulação geralmente focam em conceitos mais abstratos, como o reconhecimento de um tipo de gestos ou a análise da estrutura de um gesto. Esse tipo de análise pode ser considerada mais complexa porque, em geral, gestos que pertencem ao mesmo tipo, considerando tais conceitos mais abstratos, não compartilham semelhanças espaciais, e sim semelhanças semânticas.
Alguns trabalhos dessa categoria são os trabalhos de: Kettebekov et al. [2002] e Kettebekov et al. [2005], que reconhecem gestos com os significados "contornar"e "apontar"; Kettebekov [2004], que classifica gestos do tipo *beat* de acordo com três subtipos – *beat* simples, *beat* icônico e *beat* transicional[1]; Wilson et al. [1996], que analisam a estrutura temporal do gesto e classifica-o como gesto "bifásico"ou "trifásico"; Wilson e Bobick [2000], que utilizam a estrutura temporal do gesto para analisar gestos utilizados por maestros para conduzir orquestras; Swaminathan et al. [2008], que analisam gestos de condução de orquestra visando identificar características da música conduzida; e Eisenstein et al. [2008b], que analisam gestos co-verbais para melhorar a precisão de um sistema na análise de correferência.

Análise psicolinguística (AP):

Nesta revisão, foram encontrados alguns trabalhos que visam gerar métodos automatizados para extrair características que possam ser utilizadas na análise psicolinguística. A ideia é gerar características que possam ser relacionadas a determinadas características do discurso, de forma a automatizar algum ponto da análise.
Chen et al. [2002], Quek et al. [2002], e Quek [2004] consideram a posição da mão, a quantidade de esforço realizado pela mão em determinada janela e a presença de *holds*, extraída como descrito em Bryll et al. [2001], visando identificar relações entre tais características e reparos no discurso, estrutura do discurso e *catchment*[2], respectivamente. Já Xiong et al. [2002] visam detectar movimentos simétricos de forma automatizada, para depois relacioná-los à estrutura do discurso. Xiong et al. [2003] e Xiong e Quek [2006] apresentam trabalhos similares, porém para movimentos oscilatórios ao invés de simétricos. Outros trabalhos incluem Eisenstein [2008], que visa à segmentação do vídeo em unidades de discurso; e Chen et al. [2004], que utiliza uma abordagem multimodal para identificar os limites de uma sentença.

3.2.2. Aquisição, Extração e Representação de Dados

Nos trabalhos de análise de gestos, as tarefas de aquisição, extração e representação dos dados constituem aspectos relevantes a serem considerados.

[1] *Beat* transional é um beat que acompanha marcadores do discurso, como as conjunções "mas"e "então"; *beat* icônico é um *beat* que faz parte de um gesto icônico; e *beat* simples é o tipo mais comum, que simplesmente marca o ritmo do discurso Kettebekov [2004].

[2] *Catchment* é um conceito que declara que temas de discurso coesos são relacionados por um conjunto de imagens recorrentes e, consequentemente, a características gestuais recorrentes.

O trabalho de Mitra e Acharya [2007] divide os métodos de aquisição de dados em: baseados em sensores, que geralmente são caros e podem prejudicar a mobilidade do usuário, caso o sensor seja conectado ao computador por cabos; e baseados em visão computacional, que geralmente são menos precisos, já que podem sofrer com a oclusão de partes do corpo do usuário, entre outros problemas relacionados à tarefa de segmentação do objeto de interesse.

Durante a condução desta revisão, foram encontrados trabalhos baseados em sensores que utilizam acelerômetros [Chen et al. 2011] ou luvas com sensores [Werapan e Chotikakamthorn 2004, Nguyen e Bui 2010]. Os demais trabalhos utilizam visão computacional, sendo baseados em abordagens bidimensionais ou tridimensionais. Abordagens bidimensionais são as mais comuns e utilizam apenas uma câmera para obter as imagens, considerando apenas os eixos x e y [Xiong e Quek 2006, Kettebekov et al. 2002, Kettebekov 2004, Kettebekov et al. 2005, Corradini 2002, Wilson et al. 1996, Wilson e Bobick 2000, Choi et al. 2008, Wong e Cipolla 2006, Li e Greenspan 2007, Quek et al. 2002, Xiong et al. 2002, Quek 2004, Eisenstein et al. 2008a,b, Eisenstein 2008, Lamar et al. 1999, Yuan 2008]. Já as abordagens tridimensionais têm por objetivo obter uma posição estimada do objeto de interesse considerando os eixos x, y e z. Nesta revisão, trabalhos utilizando tais abordagens empregam: duas câmeras sincronizadas em posições diferentes para estimar a posição tridimensional [Chen et al. 2002, 2004, Bryll et al. 2001, Kim et al. 2007, Xiong et al. 2002, Quek 2004]; e marcadores para estimar a informação sobre sua profundidade [Swaminathan et al. 2008]. Note que Xiong et al. [2002] e Quek [2004] usam dois conjuntos de dados distintos em suas análises, um conjunto bidimensional e um tridimensional.

Além disso, também é interessante destacar como cada trabalho que utiliza visão computacional obtém características das imagens.

Normalmente, esse tipo de estudo utiliza um algoritmo para rastrear a posição da mão na imagem. Nesta revisão, vários trabalhos empregam um algoritmo que visa extrair um fluxo ótico do vídeo chamado *Vector Coherence Mapping* [Bryll et al. 2001, Chen et al. 2002, Xiong et al. 2002, Quek et al. 2002, Xiong et al. 2003, Quek 2004, Chen et al. 2004, Xiong e Quek 2006]. Outros trabalhos utilizam um algoritmo que integra informações sobre movimento e cor da pele em uma estrutura probabilística para determinar o caminho mais provável conectando pontos candidatos a partes do corpo ao longo do vídeo [Kettebekov et al. 2002, Kettebekov 2004, Kettebekov et al. 2005]. Outras opções incluem: algoritmos que tirem proveito do uso de luvas coloridas para auxiliar na distinção do objeto de interesse [Eisenstein et al. 2008a, Lamar et al. 1999]; algoritmos baseados na cor da pele [Corradini 2002]; *Mean Shift Algorithm*, que é um algoritmo baseado na distribuição de características visuais, como cor e textura [Choi et al. 2008]; rastreamento baseado em marcadores [Swaminathan et al. 2008].

Muitos trabalhos usam apenas a informação da posição das mãos para representar os dados [Xiong et al. 2002, 2003, Xiong e Quek 2006, Choi et al. 2008, Chen et al. 2011] ou utilizam essa informação para derivar algumas características simples, como velocidade [Werapan e Chotikakamthorn 2004, Nguyen e Bui 2010], que também pode ser associada à aceleração [Kettebekov et al. 2002, Kettebekov 2004, Kettebekov et al. 2005], ângulo entre as mãos e a cabeça [Corradini 2002], ou direção do movimento [Swaminathan et al. 2008]. A posição da mão também é utilizada por [Bryll et al. 2001]

para calcular a quantidade de esforço realizado pelas mãos com o objetivo de determinar se a mão é parte de um segmento de *hold* ou não. A presença de *hold* e a quantidade de esforço são utilizadas por Chen et al. [2002],Chen et al. [2004], Quek et al. [2002] e Quek [2004], sendo que os dois últimos também utilizam a presença de simetria extraída em [Xiong et al. 2002] como característica.

Entretanto, também é possível representar um gesto sem a necessidade de extrair a posição das mãos e da cabeça. Uma abordagem consiste em utilizar toda a imagem para representar cada *frame*. Alguns exemplos de estudos neste sentido incluem: Wilson et al. [1996], que realizam uma decomposição da imagem em autovetores, extraindo 10 coeficientes para representar cada *frame*; Eisenstein et al. [2008b], Eisenstein [2008], que extraem pontos de interesse com uma ferramenta chamada *Activity Recognition Toolbox* [Dollar et al. 2005]; Wong e Cipolla [2006], que calculam *Motion Gradient Orientation*, que é uma representação do movimento em sequências de imagens; e Yuan [2008], que usa momentos de Hu para extrair coeficientes de toda a imagem. Outra abordagem consiste em analisar a silhueta do usuário, como em: Wilson e Bobick [2000], que extraem a distribuição espacial dos pixels da silhueta do usuário através de um algoritmo de rastreamento baseado no algoritmo *Expectation-Maximization*; Kim et al. [2007], que representam o contorno da silhueta por 80 pontos e utiliza Análise de Componentes Principais para reduzir a dimensionalidade; e Li e Greenspan [2007], que representam o contorno da silhueta através de uma assinatura de borda.

3.2.3. Tipos de Análise

A partir dos resultados obtidos nesta revisão, conclui-se ser possível dividir os trabalhos de acordo com o tipo de análise realizada em três grandes grupos: reconhecimento de gestos (1), considerando gestos já segmentados; segmentação e reconhecimento de gestos (2); e outros tipos de análise (3), que, neste caso, consistem em análises psicolinguísticas. A Figura 2 apresenta a distribuição de cada tipo de análise nos artigos incluídos.

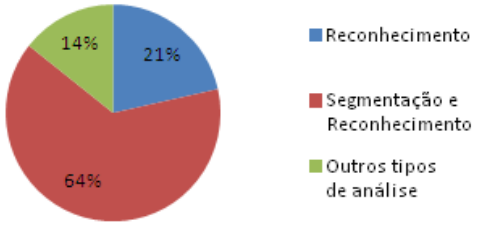

Figura 2. Artigos incluídos de acordo com o tipo de análise.

Como é possível verificar na Figura 2, 64% dos artigos se referem a estudos que realizam segmentação e reconhecimento. É possível subdividir essa categoria em três subcategorias: estudos que realizam segmentação para posterior reconhecimento (2a); estudos que utilizam os resultados do reconhecimento para segmentar o vídeo (2b); e estudos que reconhecem alguma propriedade do gesto, como simetria ou movimentos oscilatórios, e então segmentam o gesto de acordo com a ocorrência dessa propriedade. A Tabela 1 mostra a classificação para cada estudo selecionado por esta revisão.

Reconhecimento de gestos previamente segmentados

Como visto na Figura 2, 21% dos estudos encontrados nesta revisão realizam apenas a tarefa de reconhecimento, considerando que os gestos tenham sido segmentados previamente. Neste caso, gestos já segmentados são utilizados tanto para a construção dos conjuntos de treinamento quanto para os conjuntos de teste.

Esta abordagem é utilizada por Corradini [2002], Lamar et al. [1999] e Yuan [2008] para reconhecer gestos simples, que poderiam ser usados para criar novos métodos de interação humano-computador. Em Kettebekov [2004], a estratégia é utilizada para análise de gesticulação, utilizando gestos segmentados manualmente com o auxílio de uma ferramenta. Já Eisenstein et al. [2008a] segmentam manualmente as porções do vídeo que devem ser analisadas utilizando a informação proveniente da fala do usuário, visando utilizar a análise de gesticulação para ajudar na análise de correferência da fala. Similarmente, Eisenstein et al. [2008b] consideram vídeos contendo diálogos para extrair características gerais dos gestos executados no vídeo e definir se os gestos são relacionados ao assunto ou ao usuário. Neste caso, não há necessidade de segmentar os gestos, já que as características consideram o vídeo como um todo.

Métodos de aprendizado de máquina aplicados ao reconhecimento de gestos

36% dos trabalhos que contém reconhecimento de gestos utilizam *Hidden Markov Model* (HMM). Normalmente, métodos com HMM são utilizados em trabalhos que visam reconhecer conjuntos pré-definidos de gestos, como Kim et al. [2007] e Chen et al. [2011]. Entretanto, esta revisão encontrou trabalhos que utilizam HMM para reconhecer classes mais abstratas de gestos, como em Kettebekov [2004] e Kettebekov et al. [2005], e como módulo integrador em Chen et al. [2004], visando integrar modelos de reconhecimento específicos para linguagem, prosódia e gestos. Também é possível utilizar HMM para modelar a estrutura temporal dos gestos, como em Wilson e Bobick [2000], que realizam um treinamento para associar a imagem do usuário a cada estado da HMM no começo da interação com o sistema, ao invés de um treinamento *offline*. De forma similar, Wilson et al. [1996] utilizam uma máquina de estados finitos (*Finite State Machine* – FSM) para modelar a estrutura temporal dos gestos e utilizar essa informação para classificá-los em "bifásicos"ou "trifásicos". Além disso, HMM pode ser combinada a Redes Neurais Recorrentes em uma arquitetura híbrida que usa redes neurais para definir as probabilidades de transição dos estados da HMM, como em Corradini [2002].

Outros métodos de reconhecimento incluem: Programação Dinâmica Contínua (*Continuous Dynamic Programming* – CDP), em Li e Greenspan [2007]; Máquina de Vetor de Relevância (*Relevance Vector Machine* – RVM), em Wong e Cipolla [2006]; C-means associado à Maior Subsequência Comum (*Longest Common Subsequence* – LCS), em Choi et al. [2008]; árvores de decisão, em Chen et al. [2004]; e abordagens probabilísticas, principalmente análise bayesiana, em Eisenstein et al. [2008b] e Eisenstein [2008]; uma rede dinâmica bayesiana, em Swaminathan et al. [2008]; e redes neurais (*Neural Networks* – NN), como em Lamar et al. [1999] e Yuan [2008]. O último trabalho também usa: Máquinas de Vetores Suporte (*Support Vector Machines* – SVM), Alinhamento Dinâmico do Tempo (*Dynamic Time Alignment* – DTA) e Deformação Dinâmica do Tempo (*Dynamic Time Warping* – DTW). Um caso especial consiste em utilizar uma técnica chamada Fusão de Modalidade Condicional (*Conditional Modality Fusion* – CMF) para incorporar

características gestuais para melhorar a análise de correferência, reconhecendo se duas frases se referem ao mesmo objeto ou não Eisenstein et al. [2008a].

Estratégias para segmentação

Como visto na Figura 2, 64% dos trabalhos incluem segmentação. Nesse aspecto, é possível dividir os trabalhos selecionados em três formas de realizar a segmentação: baseada nos resultados do reconhecimento; através da aplicação de algum filtro ou *threshold* sobre todos os *frames*; ou através da detecção de um ponto de mudança no sinal, que representa o início ou final de um gesto.

Segmentação baseada no modelo de reconhecimento

A segmentação baseada no resultado do reconhecimento de gestos é geralmente utilizada para problemas de reconhecimento de gestos contínuos. Em geral, um modelo de reconhecimento probabilístico é utilizado e os gestos são segmentados de acordo com a sua probabilidade de existência. Exemplos de trabalho desse tipo incluem: Wong e Cipolla [2006], que geram hipóteses sobre pontos de início e fim de cada gesto, utiliza RVM para estimar a probabilidade de existência de cada classe no segmento analisado e, quando a probabilidade excede determinado valor, o gesto é considerado válido; e Li e Greenspan [2007], que utilizam CDP para detectar segmentos nos quais a ocorrência de um gesto é mais provável.

Já Eisenstein [2008] usam uma abordagem similar aplicando uma estratégia bayesiana para detecção do tópico do discurso. Nesse trabalho, um classificador bayesiano avalia a probabilidade de determinado segmento corresponder a certo tópico do discurso e, posteriormente, um conjunto de segmentos é avaliado para gerar a melhor segmentação possível de acordo com as probabilidades geradas pela estratégia. Em um problema similar, Chen et al. [2004] utilizam classificadores para detectar os limites das sentenças e segmenta o vídeo de acordo com os limites encontrados.

Para problemas relacionados à análise de gesticulação, Kettebekov et al. [2002] e Kettebekov et al. [2005] utilizam uma técnica chamada *Token Passing*, que consiste em calcular iterativamente a probabilidade de possíveis interpretações de sequências de gestos em cada segmento, enquanto Wilson et al. [1996] e Wilson e Bobick [2000] utilizam FSM e HMM, respectivamente, para processar os gestos e obter suas fases – fases de descanso, transição e *stroke* no primeiro trabalho, e fases de cima e baixo no segundo trabalho. De forma similar, Swaminathan et al. [2008] utilizam um modelo de reconhecimento para segmentar as fases de *stroke* de gestos de condução de orquestra.

Finalmente, Choi et al. [2008] realizam um agrupamento para dividir todas as posturas possíveis em grupos. Cada gesto é representado por uma sequência de posturas (descritas pelo grupo ao qual pertencem) e LCS é utilizado para classificar o gesto. Como o LCS classifica o gesto de acordo com a mais longa sequência válida de posturas, a estratégia pode ser usada para segmentação.

Segmentação baseada em filtros ou thresholds

A segunda estratégia de segmentação consiste em aplicar filtros ou *thresholds* a todos os *frames*. Tal estratégia é utilizada por Nguyen e Bui [2010] para segmentar gestos estáticos da Língua de Sinais Vietnamita através da análise da velocidade da mão: quando a velocidade da mão em um conjunto de *frames* está abaixo de determinado valor, o

segmento deve conter um gesto estático; caso contrário, tal conjunto de *frames* contém um movimento de transição.

Seguindo a mesma estratégia, alguns estudos empregam filtros ou *thresholds* visando o reconhecimento de alguma propriedade nos gestos, como simetria e movimentos oscilatórios, de forma que os segmentos são determinados de acordo com a ocorrência da propriedade de interesse. Alguns exemplos incluem: Xiong et al. [2002], que identificam segmentos simétricos quando a correlação entre os sinais das mãos direita e esquerda excede certo valor; Xiong et al. [2003], que utilizam CWT para identificar movimentos oscilatórios; Xiong e Quek [2006], que comparam os resultados da aplicação dos filtros WFT e CWT para a identificação segmentos contendo movimentos oscilatórios; e Bryll et al. [2001], que detectam *holds* através de um threshold adaptativo, utilizando a "regra da mão dominante", que declara que a quantidade de movimento necessária para definir que uma mão está em *hold* depende da quantidade de movimento realizada pela outra mão.

Segmentação baseada em detecção de um ponto de mudança

A terceira estratégia consiste em detectar um ponto de mudança em uma sequência de gestos, que representa o início e o fim de um gesto. A abordagem é utilizada por Werapan e Chotikakamthorn [2004] e Chen et al. [2011]. O primeiro define pontos de mudança quando a velocidade da mão atinge um ponto mínimo, enquanto o último se baseia na hipótese de que a aceleração muda de forma abrupta no início e no fim de um gesto.

Estratégias combinadas

Também é possível utilizar duas estratégias em conjunção. Kim et al. [2007], por exemplo, utilizam um classificador para determinar a probabilidade de existência de um gesto de cada classe em determinado *frame*. Depois, calcula um coeficiente que correspondente à diferença entre a probabilidade máxima encontrada e a probabilidade de não encontrar um gesto no *frame* e determina pontos de mudança quando tal coeficiente passa de valores positivos a negativos (ou vice-versa). Assim, tal trabalho combina a primeira e a terceira estratégia para realizar a segmentação.

Relação entre as tarefas de segmentação e reconhecimento

Alguns estudos realizam a segmentação aplicando técnicas de reconhecimento que definem para cada segmento a probabilidade de conter um gesto (2a). Esses estudos dependem de um conjunto de dados de gestos previamente segmentados para treinamento, como em Wong e Cipolla [2006], Li e Greenspan [2007] e Kettebekov [2004].

Outros realizam a segmentação e o reconhecimento de forma simultânea, geralmente aplicando um método heurístico (2b). Nesta revisão, esse caso corresponde aos estudos que realizam a segmentação de acordo com a detecção de algum tipo de propriedade, como em Xiong e Quek [2006], Xiong et al. [2003], Xiong et al. [2002] e Bryll et al. [2001]. Tais estratégias são diretamente relacionadas às estratégias de segmentação previamente descritas. Contudo, a terceira estratégia consiste em realizar a segmentação através da detecção de pontos de mudança e posteriormente realizar o reconhecimento (2c). Werapan e Chotikakamthorn [2004], por exemplo, utilizam velocidade para determinar pontos de mudança (segmentação) e então analisa o sinal utilizando Análise de Fourier e procurando picos do espectro para detectar movimentos periódicos (reconhec-

imento). De forma similar, Chen et al. [2011] e Nguyen e Bui [2010] detectam pontos de mudança para depois realizar o reconhecimento utilizando técnicas de aprendizado de máquina.

Tabela 1. Estudos com aplicação, tipo de análise, método de segmentação e método de reconhecimento.

Estudo	Aplicação	Tipo	Seg.	Rec.
Eisenstein et al. [2008a]	AG – análise de correferência.	1	–	Fusão de Modalidade Condicional
Eisenstein et al. [2008b]	AP – classifica gestos como dependentes do assunto ou da pessoa.	1	–	Modelo bayesiano
Kettebekov [2004]	AG	1	–	HMM
Corradini [2002]	AGPD – gestos simples.	1	–	HMM com NN
Lamar et al. [1999]	AG – datilologia japonesa.	1	–	T-Comb Net (NN)
Yuan [2008]	AGPD – diversos conjuntos de dados.	1	–	SVM, DTA, DTW, e NN
Nguyen e Bui [2010]	AGPD – Língua de Sinais Vietnamita.	2a	S2	Baseado na segmentação
Werapan e Chotikakamthorn [2004]	AGPD – Língua de Sinais Taiwanesa.	2a	S3	Baseado na segmentação
Chen et al. [2011]	AGPD – gestos para controlar apresentação de *slides*.	2a	S3	HMM
Chen et al. [2004]	AP – segmentação de unidade de sentença.	2b	S1	Árvores de decisão bayesianas
Eisenstein [2008]	Análise psicolinguística – segmentação de unidade de discurso.	2b	S1	Análise bayesiana
Wilson et al. [1996]	AG – classifica gestos em “bifásicos”ou “trifásicos”.	2b	S1	FSM
Wilson e Bobick [2000]	AG – gestos de maestros.	2b	S1	HMM
Choi et al. [2008]	AGPD – gestos de sinalização da aviação.	2b	S1	K-means com LCS
Wong e Cipolla [2006]	AGPD – simple gestures.	2b	S1	RVM
Kettebekov et al. [2002]	AG – gestos de *apontar* e *contornar* no contexto de narração do clima.	2b	S1	HMM
Li e Greenspan [2007]	AGPD – gestos simples.	2b	S1	CDP
Kettebekov et al. [2005]	AG – gestos de *apontar* e *contornar* no contexto de narração do clima.	2b	S1	HMM
Swaminathan et al. [2008]	AG – gestos de maestros.	2b	S1	Rede bayesiana dinâmica
Kim et al. [2007]	AGPD – gestos para controlar um ambiente de casa inteligente.	2b	S1 S3	HMM acumulativa
Xiong e Quek [2006]	AP – detecção de movimentos oscilatórios.	2c	S2	WFT e CWT
Xiong et al. [2003]	AP – detecção de movimentos oscilatórios.	2c	S2	WFT e CWT
Bryll et al. [2001]	AP – segmentação de *hold*.	2c	S2	Baseado na segmentação
Xiong et al. [2002]	AP – detecção de comportamento simétrico.	2c	S2	Baseado na segmentação
Chen et al. [2002]	AP – pesquisa sobre reparos no discurso.	3	–	–
Quek et al. [2002]	AP – pesquisa sobre estrutura do discurso.	3	–	–
Quek [2004]	AP – pesquisa sobre *catchment*.	3	–	–
Kita et al. [1998]	AP – Gramática para a segmentação das fases do gesto.	3	–	–

3.2.4. Métodos para Análise de Resultados

Grande parte dos estudos que realizam tarefas de reconhecimento utilizam taxa de acerto como métrica para analisar os resultados [Eisenstein 2008, Choi et al. 2008, Kim et al. 2007, Werapan e Chotikakamthorn 2004, Chen et al. 2011]. Entretanto, outros métodos são usados para analisar resultados de reconhecimento, como: matriz de confusão [Kettebekov 2004]; precisão e sensibilidade [Nguyen e Bui 2010, Bryll et al. 2001]; especificidade [Bryll et al. 2001]; e área sob a curva ROC (*Receiver Operating*

Characteristic) [Eisenstein et al. 2008a].

Por outro lado, estudos que realizam o reconhecimento de gestos contínuos, ou seja, que incluem a tarefa de segmentação dos gestos no vídeo, podem utilizar uma métrica que considera três tipos de erro: inserção, que ocorre quando um gesto reconhecido na verdade não existe; deleção, que ocorre quando um gesto que existe no vídeo não é reconhecido; e substituição, quando um gesto que existe no vídeo é reconhecido, porém como um gesto diferente. Tal métrica é utilizada em Li e Greenspan [2007], Kettebekov et al. [2005], Wong e Cipolla [2006], Kettebekov et al. [2002] e Chen et al. [2004], porém no último caso a métrica é adaptada: como o trabalho visa reconhecer limites de sentença, só é possível haver erros de inserção e deleção.

Outra estratégia para analisar resultados de segmentação e reconhecimento de gestos envolve a comparação de vídeos rotulados por pessoas. Essa estratégia é utilizada por Wilson et al. [1996], que usam apenas gráficos para comparar os resultados gerados pelo classificador com os vídeos rotulados manualmente. Porém, é importante notar que podem existir divergências entre rótulos criados por diferentes pessoas [Kita et al. 1998]. Assim, para avaliar um método automatizado através da comparação com os rótulos gerados por um especialista, é necessário considerar uma comparação com rótulos entre diversos especialistas, para que seja possível avaliar se a diferença entre o nível de concordância entre dois especialistas e entre os especialistas e o método automatizado é significativa.

Por fim, alguns estudos apresentam os resultados por meio de gráficos. Xiong e Quek [2006] analisam gráficos com os resultados da aplicação de CWT e WFT nos sinais de localização da mão, visando detectar movimentos oscilatórios, e usa análises gráficas para argumentar que o CWT é mais apropriado para a tarefa. De forma similar, Swaminathan et al. [2008] utilizam gráficos para apresentar os resultados da inferência do tipo de articulação – *legato* ou *staccato* – em determinado trecho de uma música.

4. Considerações Finais

Este documento apresentou alguns conceitos básicos sobre análise de gestos e uma revisão sistemática sobre a pesquisa em análise de gestos manuais, focando em aspectos temporais da análise e trabalhos aplicados à conversação natural e análise do discurso.

O processo de revisão sistemática é documentado neste documento através do protocolo de revisão (Apêndice A) e da descrição do processo de condução. A análise dos resultados cobre os principais aspectos dos trabalhos, incluindo as aplicações de análise de gesto, os métodos para aquisição, extração e representação de dados, os tipos de análise realizadas, e a forma de análise dos resultados. Este documento também apresenta uma sistematização para classificar os trabalhos de acordo com cada um desses aspectos.

Apesar de existirem outras revisões sistemáticas sobre análise de gestos, esta revisão apresenta alguns diferenciais: seu foco em aspectos temporais; seu escopo, visando cobrir várias aplicações, incluindo análise psicolinguística; e a sistematização proposta para cada tópico analisado, como tipo de aplicação, estratégias para aquisição, extração e representação dos dados, tipo de análise realizada e métodos para análise dos resultados.

Referências

Bryll, R., Quek, F., and Esposito, A. (2001). Automatic Hand Hold Detection in Natural Conversation. In *IEEE Workshop on Cues in Communication, Kauai, Hawaii*, pages

1–6.

Chen, L., Harper, M., and Quek, F. (2002). Gesture patterns during speech repairs. In *Proceedings. Fourth IEEE International Conference on Multimodal Interfaces*, pages 155–160. IEEE Comput. Soc.

Chen, L., Liu, Y., Harper, M. P., and Shriberg, E. (2004). Multimodal model integration for sentence unit detection. In *Proc. of the 6th international conference on Multimodal interfaces - ICMI '04*, pages 121–128, New York, New York, USA. ACM Press.

Chen, Y., Liu, M., Liu, J., Shen, Z., and Pan, W. (2011). Slideshow: Gesture-aware ppt presentation. In *Multimedia and Expo (ICME), 2011 IEEE International Conference on*, pages 1 –4.

Choi, C., Ahn, J., and Byun, H. (2008). Visual recognition of aircraft marshalling signals using gesture phase analysis. In *Intelligent Vehicles Symposium, 2008 IEEE*, pages 853 –858.

Corradini, A. (2002). Real-time gesture recognition by means of hybrid recognizers. In *Gesture and sign language in human-computer interaction*, pages 157–189. Springer.

Dollar, P., Rabaud, V., Cottrell, G., and Belongie, S. (2005). Behavior recognition via sparse spatio-temporal features. In *Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on*, pages 65 – 72.

Eisenstein, J. (2008). *Gesture in Automatic Discourse Processing*. PhD thesis, Massachussets Institute of Technology, Cambridge, MA, USA.

Eisenstein, J., Barzilay, R., and Davis, R. (2008a). Gesture Salience as a Hidden Variable for Coreference Resolution and Keyframe Extraction. *Artificial Intelligence*, 31(1):353–398.

Eisenstein, J., Barzilay, R., and Randall, D. (2008b). Discourse topic and gestural form. In *Proc. of AAAI*, pages 836–841.

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

Kettebekov, S. (2004). Exploiting prosodic structuring of coverbal gesticulation. In *Proc. of the 6th international conference on Multimodal interfaces - ICMI '04*, pages 105–112, New York, New York, USA. ACM Press.

Kettebekov, S., Yeasin, M., Krahnstoever, N., and Sharma, R. (2002). Prosody based co-analysis of deictic gestures and speech in weather narration broadcast. In *Workshop on Multimodal Resources and Multimodal System Evaluation.(LREC 2002), Las Palmas, Spain*. Citeseer.

Kettebekov, S., Yeasin, M., and Sharma, R. (2005). Prosody based audiovisual coanalysis for coverbal gesture recognition. *Multimedia, IEEE Trans. on*, 7(2):234 – 242.

Kim, D., Song, J., and Kim, D. (2007). Simultaneous gesture segmentation and recognition based on forward spotting accumulative HMMs. *Pattern Recognition*, 40(11):3012–3026.

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

Lamar, M., Shoaib Bhuiyan, M., and Iwata, A. (1999). Temporal series recognition using a new neural network structure t-combnet. In *Neural Information Processing, 1999. Proceedings. ICONIP '99. 6th International Conference on*, volume 3, pages 1112 – 1117 vol.3.

Li, H. and Greenspan, M. (2007). Segmentation and recognition of continuous gestures. In *Image Processing, 2007. ICIP 2007. IEEE International Conference on*, volume 1, pages I –365 –I –368.

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

McNeill, D. (2005). *Gesture and Thought*. University of Chicago Press.

Mitra, S. and Acharya, T. (2007). Gesture recognition: A survey. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Trans. on*, 37(3):311 –324.

Nguyen, N. T. and Bui, T. D. (2010). Automated posture segmentation in continuous finger spelling recognition. In *Human-Centric Computing (HumanCom), 2010 3rd International Conference on*, pages 1 –5.

Quek, F. (2004). The Catchment Feature Model: A Device for Multimodal Fusion and a Bridge between Signal and Sense. *EURASIP Journal on Advances in Signal Processing*, 2004(11):1619–1636.

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

Swaminathan, D., Thornburg, H., Ingalls, T., Rajko, S., James, J., Campana, E., Afanador, K., and Leistikow, R. (2008). Capturing expressive and indicative qualities of conducting gesture: An application of temporal expectancy models. In Kronland-Martinet, R., Ystad, S., and Jensen, K., editors, *Computer Music Modeling and Retrieval. Sense of Sounds*, volume 4969 of *Lecture Notes in Computer Science*, pages 34–55. Springer Berlin / Heidelberg.

Werapan, W. and Chotikakamthorn, N. (2004). Improved dynamic gesture segmentation for thai sign language translation. In *Signal Processing, 2004. Proceedings. ICSP '04. 2004 7th International Conference on*, volume 2, pages 1463 – 1466 vol.2.

Wilson, A. and Bobick, A. (2000). Realtime online adaptive gesture recognition. In *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000*, volume 1, pages 270–275. IEEE Comput. Soc.

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

Wong, S. and Cipolla, R. (2006). Continuous Gesture Recognition using a Sparse Bayesian Classifier. In *18th International Conference on Pattern Recognition (ICPR'06)*, pages 1084–1087. IEEE.

Xiong, Y. and Quek, F. (2006). Hand Motion Gesture Frequency Properties and Multimodal Discourse Analysis. *International Journal of Computer Vision*, 69(3):353–371.

Xiong, Y., Quek, F., and McNeill, D. (2002). Hand gesture symmetric behavior detection and analysis in natural conversation. In *Proceedings. Fourth IEEE International Conference on Multimodal Interfaces*, pages 179–184. IEEE Comput. Soc.

Xiong, Y., Quek, F., and McNeill, D. (2003). Hand Motion Gestural Oscillations and Multimodal Discourse. In *Proc. of the 5th international conference on Multimodal interfaces. ICMI '03*, pages 132–139.

Yuan, Y. (2008). *Image-based gesture recognition with support vector machines*. PhD thesis, University of Delaware, Newark, DE, USA.

Agradecimentos

As autoras agradecem à Fundação de Amparo à Pesquisa do Estado de São Paulo (Fapesp) pelo suporte através do processo número 2011/04608-8 e à professora Dra. Fátima de Lourdes dos Santos Nunes Marques por seu apoio durante o processo de realização desta revisão sistemática.

Apêndice A. Protocolo da Revisão Sistemática sobre Análise de Gestos

Este apêndice apresenta o protocolo da revisão sistemática cujos resultados são apresentados neste documento.

A.1. Objetivos

Identificar e analisar métodos e técnicas utilizados na análise de gestos manuais aplicada a problemas relacionados à conversação natural e à análise do discurso.

A.2. Questões de Pesquisa

1. Quais são os tipos de análises realizadas em gestos manuais?
2. Quais métodos e técnicas são aplicados a problemas relacionados à análise da conversação natural e do discurso?
3. Quais são os métodos e técnicas utilizados na análise temporal de gestos manuais?
4. Quais destes métodos e técnicas utilizados para detecção e/ou identificação dos componentes elementares dos gestos, como, por exemplo, detecção de suas fases?

- **Intervenção:** Sistematização dos métodos e técnicas utilizados na análise de gestos manuais aplicada ao processamento automático da conversação natural e do discurso.
- **Controle:** Artigos e livros referentes a teoria lingüística sobre gestos indicados pelo Prof. Dr. Leland McCleary, artigos de survey sobre análise de comportamento humano e análise de gestos, e artigos obtidos na Internet através da análise exploratória.
- **População:** Artigos que contenham estratégias, métodos e técnicas utilizados na análise de gestos manuais considerando características temporais ou aplicados a conversação natural e análise do discurso.
- **Resultados:** Métodos e técnicas aplicáveis (ou aplicados) a análise de gestos manuais.
- **Aplicação:** Projetos em computação relacionados à análise automatizada de gestos manuais, projeto relacionados à utilização de gestos manuais em Interação Humano-Computador e pesquisas na área de lingüística sobre análise de gestos manuais.

A.3. Seleção de Fontes

A.3.1. Lista de fontes iniciais

- Biblioteca Digital do IEEE (http://ieeexplore.ieee.org/Xplore/dynhome.jsp)
- Biblioteca Digital da ACM (http://portal.acm.org/dl.cfm)
- Biblioteca Digital CiteSeerX (http://http://citeseerx.ist.psu.edu)
- Biblioteca Digital SpringerLink (http://www.springerlink.com)
- Biblioteca de Teses da USP (http://www.teses.usp.br)

A.3.2. Critério de definição de fontes

Os trabalhos devem, preferencialmente, estar disponíveis na Internet, em bases de dados científicas.

A.3.3. Idiomas dos artigos

Inglês e português.

A.3.4. Palavras-Chave

1. Os termos *gesture* e *hand* relacionados a *discourse analysis* ou *natural conversation* e a *hand motion* ou *hand movement*.
2. O termo *gesture* relacionado a *temporal structure* ou *gesture phases* ou *movement phases* ou *temporal analysis*.
3. O termo *gesture* relacionado a *continuous gesture* e *segmentation*.

A.3.5. Seleção das fontes após avaliação

- Biblioteca Digital do IEEE (http://ieeexplore.ieee.org/Xplore/dynhome.jsp)
- Biblioteca Digital da ACM (http://portal.acm.org/dl.cfm)
- Biblioteca Digital CiteSeerX (http://http://citeseerx.ist.psu.edu)
- Biblioteca Digital SpringerLink (http://www.springerlink.com)
  Obs.: A Biblioteca de Teses da USP foi excluída após pesquisas iniciais que indicaram que não existem resultados para as palavras-chave pesquisadas.

A.4. Critérios de inclusão e exclusão considerados

Critérios de Inclusão:

- I1: Artigos que abordem estratégias, métodos e técnicas utilizados na análise automatizada de gestos aplicada à conversação natural e à análise do discurso; ou
- I2: Artigos que abordem estratégias, métodos e técnicas utilizados na análise temporal de gestos visando à detecção e/ou identificação de componentes elementares, como detecção de fases ou segmentação de gestos; ou
- I3: Trabalhos que discutam interação multimodal, com pelo menos uma seção dedicada à análise de gestos de mão.

Critérios de Exclusão:

- E1: Artigos que não discutam os métodos e técnicas utilizados na análise automatizada, como artigos com foco na segmentação de gestos e construção de conjuntos de dados;
- E2: Artigos com foco em gestos que não sejam realizados com as mãos, como gestos de cabeça ou olhar;
- E3: Artigos que considerem gestos objetivando a síntese ou reconstrução de gestos através de avatares ou que analisem dados sintéticos (gerados a partir de animações);

- E4: Trabalhos sobre a construção de ferramentas multimodais, sem ênfase na análise de gestos;
- E5: Artigo mais antigo no caso de artigos do mesmo autor que abordem o mesmo tema, com diferenças pouco significativas entre si;
- E6: Trabalhos que não tenham sido publicados em conferências ou periódicos (como relatórios técnicos) ou que não tenham sido aprovados (para teses e dissertações);
- E7: Trabalhos recuperados pelas pesquisas que não se encaixem em nenhum critério de inclusão ou exclusão serão excluídos.

A.5. Critério de qualidade dos estudos

Para avaliar os artigos, serão considerados: tipo de análise realizada e resultados obtidos, de acordo com a forma de avaliação dos resultados. Serão considerados apenas artigos publicados em periódicos e conferências com revisão por pares e teses ou dissertações aprovadas por banca examinadora.

A.6. Definição da estratégia de seleção de dados

Serão construídas strings com as palavras-chave e seus sinônimos. As strings serão submetidas às máquinas de busca. Após a leitura do resumo, títulos das seções e aplicação dos critérios de inclusão e exclusão, o trabalho será selecionado se confirmada a sua relevância pelo principal revisor. Se houver dúvida da relevância os demais revisores serão consultados.

Após definidos os trabalhos definitivamente incluídos, estes deverão ser lidos na íntegra. O revisor fará um resumo de cada um deles, destacando os métodos utilizados para a análise de gestos, parâmetros considerados, quando for o caso, o tipo de análise executada e um resumo dos resultados obtidos.

A.7. Definição da Síntese dos dados extraídos

Após a leitura e o resumo dos trabalhos selecionados, será elaborado um relatório técnico com uma análise quantitativa dos trabalhos. Também será elaborada uma análise qualitativa a fim de definir as vantagens, desvantagens e aplicação de cada método. Para auxiliar na análise qualitativa será elaborado um checklist com itens importantes a serem observados em cada método apresentado. Será composta uma tabela identificando para cada trabalho, incluindo: tipo de análise de gestos realizada, métodos e técnicas utilizados para a análise, se o trabalho utiliza análise temporal, se o trabalho utiliza as fases dos gestos para realizar a análise ou realiza segmentação de gestos contínuos, resultados obtidos e forma de avaliação dos resultados. Posteriormente, uma análise quantitativa será realizada utilizando os itens levantados para a análise quantitativa.