Representação de Gestos em Análise Sintática: uma Revisão Sistemática da Literatura

Alternative Title: Representation of Gestures in Syntactic Analysis: a Systematic Literature Review

Ricardo A. Feitosa
Universidade de São Paulo
Arlindo Bettio, 1000
03828-000, São Paulo – SP
ricardoasfa@usp.br

Sarajane M. Peres
Universidade de São Paulo
Arlindo Bettio, 1000
03828-000, São Paulo – SP
sarajane@usp.br

Clodoaldo A. M. Lima
Universidade de São Paulo
Arlindo Bettio, 1000
03828-000, São Paulo – SP
c.lima@usp.br

RESUMO

Sistemas de análise de gestos vêm ganhando força por sua capacidade de contribuir com a interação entre humanos, humanos e máquinas, e humanos e ambiente. Nesses sistemas, o estabelecimento de uma representação de dados eficiente para os gestos é uma tarefa crítica. A representação escolhida e sua associação a técnicas de análise podem ou não favorecer a solução sob implementação. Nesta Revisão Sistemática são identificadas e discutidas as estratégias de representação de gestos usadas em 21 estudos publicados nos últimos 5 anos, no contexto da análise sintática de gestos.

Palavras-Chave

Análise Sintática de Gestos, Representação de Gestos, Reconhecimento de Padrões Sintático.

ABSTRACT

Gestures analysis systems have been getting attention for their ability to contribute to the interaction between humans, humans and machines, and humans and environments. In such systems, the establishment of an efficient data representation for gestures is a critical task. The chosen representation as well as its combination with techniques for analysis can or can not favor the solution being developed. In this systematic review we identify and discuss the strategies of representation of gestures used in 21 studies published in the last five years in the context of syntactic gesture analysis.

Categories and Subject Descriptors

H.4.m [**Information Systems Applications**]: Miscellaneous; I.5.1 [**Pattern Recognition**]: Models—*Structural*

SBSI 2016, May 17th-20th, 2016, Florianópolis, Santa Catarina, Brazil

General Terms

Documentation

Keywords

Syntactic Gesture Analysis, Representation of Gestures, Syntactic Pattern Recognition.

1. INTRODUÇÃO

Análise de gestos pode ser vista como uma aplicação da área de reconhecimento de padrões, e a depender do contexto de análise, os gestos podem assumir o papel de um elemento de sistema linguístico. Considerando o uso de gestos para comunicação ou interação, essa relação se estabelece de forma natural. Dentro de um sistema linguístico a execução e interpretação de um gesto pode assumir caráter estrutural e pode embutir informação relacional. O reconhecimento de padrões considerando estrutura e informação relacional pode ser mais facilmente implementado a partir de abordagens que são capazes de processar informação estruturada, com descrições hierárquicas, ou ainda a partir de abordagens que sejam capazes de implementar análise estruturada ou sequencial. Assim, o contexto de reconhecimento de padrões sintático, e de análise estruturada [28] ou de informação sequencial [8], aplicados à análise de gestos, delimita a área tratada neste artigo como "análise sintática" ou "análise de gestos considerando aspectos sintáticos".

O desenvolvimento de soluções com reconhecimento de padrões, depara-se com a necessidade da escolha de uma representação computacional para os dados sob análise. Essa é uma escolha crítica, juntamente com a escolha das técnicas de análise de dados, pois elas influenciam diretamente o desempenho da solução. Para diferentes domínios de aplicação, diferentes representações de dados podem ser usadas, e quando um pesquisador se debruça sob uma tarefa deste tipo, uma de suas primeiras ações é levantar as possibilidades de representação e quais são as chances dessas representações levarem a um bom desempenho de análise de dados.

Este artigo tem o objetivo de apresentar uma revisão sistemática da literatura (RS) voltada para a área de análise sintática de gestos, com foco em identificar e discutir formas de representação de gestos. Uma RS é um método de pesquisa científica estruturado para coletar e analisar a pesquisa relevante sobre um determinado problema [2]. Estudos indi-

viduais selecionados e analisados em uma RS são chamados de estudos primários; uma RS é chamada de estudo secundário. Uma RS é preparada para ser reprodutível e para minimizar possíveis vieses de análises comumente presentes em revisões e *surveys* não sistemáticas, nas quais as decisões advém unicamente da experiência de especialistas. Até onde foi possível identificar por meio da condução desta RS e também por meio de análises exploratórias, não foram encontrados estudos secundários que apresentem uma RS, uma revisão simples de literatura ou uma *survey* sobre o tema tratado nesta RS. Entretanto, dentro da área de análise de gestos há um grande esforço em produzir estudos secundários, e alguns destes estudos são mencionados neste artigo como uma forma de ilustrar a área em que esta RS se aplica.

A fim de documentar os procedimentos executados e os resultados obtidos nesta RS, este artigo está organizado da seguinte forma: na seção 2 são mencionadas revisões de literatura, sistemáticas ou não, que ajudam a definir a área de análise de gestos; o método aplicado nesta RS é apresentado na seção 3, com informações sobre as etapas de planejamento, condução e análise de resultados; os resultados são discutidos na seção 4; e finalmente, na seção 5 as conclusões são apresentadas seguidas das referências bibliográficas.

2. TRABALHOS CORRELATOS

A área de análise de gestos pode ser estudada sob, pelo menos, duas perspectivas. A primeira considera um vocabulário finito de gestos com significados pré-definidos. Nesse contexto, desenvolve-se aplicações, por exemplo, para interação humano-máquinas e humano-ambientes, usando gestos executados no espaço tridimensional ou em recursos com superfícies sensíveis ao toque, ou para interpretação de línguas de sinais considerando suas gesticulações primitivas e as combinações delas. A segunda perspectiva considera a gesticulação natural, na qual os gestos não têm significado pré-definido e podem ser dependentes de aspectos culturais e regionais. Nesse contexto, desenvolve-se aplicações que estão diretamente ligadas à análise do comportamento humano, à análise do discurso e de aspectos complexos de sistemas linguísticos. Para conhecer em mais detalhes esses e outros aspectos de análise de gestos, faz-se interessante a consulta a estudos secundários desenvolvidos na área. A tabela 1 lista alguns desses estudos, organizados por ano de publicação, junto com uma pequena descrição sobre o tema abordado em cada um.

Tabela 1: Revisões e *surveys* sobre análise de gestos

Estudos	Ano	Tema
[25]	2005	Interpretação de línguas de sinais
[23]	2007	Desenvolvimento de soluções que necessitam de reconhecimento de gestos
[11]	2013	Aplicações de visão computacional com apoio do sensor MS Kinect
[22]	2013	Aspectos temporais na análise de gestos no discurso e na conversação natural
[5]	2015	Análise de ações humanas a partir de dados de profundidade e de inércia
[20]	2015	Análise de ações humanas com apoio sensor MS Kinect
[26]	2015	Métodos para análise de gestos a partir do uso de visão computacional

3. MÉTODO

Essa RS foi conduzida seguindo as etapas [12]: planejamento, condução e análise de resultados, as quais são relatadas nas próximas seções deste artigo. Esse processo foi apoiado pela ferramenta StArt[1]. Dentro do problema delineado na introdução desse trabalho, essa RS é guiada pelo objetivo de identificar estratégias de representação de dados que vêm sendo usadas na tarefa de análise de gestos, considerando aspectos de caráter sintático. Estratégias aplicadas em estudos que analisaram gestos, manuais ou não-manuais, com diferentes objetivos, são de interesse nesta revisão, visto que as representações de dados podem ser transferidas para diferentes contextos e, portanto, o relato de diferentes tipos de experiências pode ser de alto valor agregado para o desenvolvimento de pesquisas na área.

3.1 Planejamento

O planejamento desta RS abrange a definição da questão de pesquisa acompanhada das justificativas para sua escolha, e a definição do protocolo de execução, o qual inclui definição das fontes de dados, da estratégia de busca, de critérios de seleção de estudos primários e de regras para avaliação da qualidade desses estudos.

Questão de pesquisa:

A presente RS responde a seguinte questão de pesquisa:

- Quais são as formas de representação de dados que vêm sendo usadas nas iniciativas de pesquisa que propõem a resolução da tarefas de análise de gestos considerando aspectos sintáticos?

O estudo focado em formas de representação de dados é principalmente motivado pela dificuldade inerente à tarefa de escolha da representação a ser usada em uma iniciativa de análise de gestos. Embora a presente RS faça menções ao desempenho obtido na análise de gestos usando uma ou outra representação de dados, não é pretendido fornecer uma resposta referente ao melhor tipo de representação uma vez que o sucesso de uma abordagem de análise de gestos depende ainda de outros fatores como a complexidade do problema sob resolução, as técnicas usadas na tomada de decisão e a metodologia usada para chegar aos resultados e conclusões apresentadas nas pesquisas. Entretanto, pretende-se fazer desta RS um instrumento para que pesquisadores tenham um panorama geral das possibilidades de representação de dados diante de um recorte de experiências na área, suportando ainda que parcialmente a etapa de levantamento bibliográfico de uma pesquisa.

A escolha para o recorte no tipo de análise (sintática) explorado na presente RS é principalmente motivado pela hipótese de que, em termos de automação e desenvolvimento de sistemas, a análise de gestos é frequentemente relacionada ao problema no contexto de comunicação, seja esta uma forma de interação entre humanos, entre o humano e a máquina ou entre o humano e o ambiente; e tomando a comunicação como motivação, há uma correspondência natural que leva à interpretação de que o uso dos gestos constituem-se como uma linguagem ou como elementos de uma língua, em algum nível, com o uso de regras sintáticas explícitas ou implícitas que estabelecem algum tipo de sequência ou conjunto de pré e pós condições que devem ser aplicadas.

[1] `http://lapes.dc.ufscar.br/tools/start_tool`

Protocolo:

As **fontes de dados** e a **estratégia de busca** foram determinadas por meio de uma análise exploratória. Para as fontes de dados verificou-se sua representatividade dentro das áreas nas quais as pesquisas de análise de gestos automática são geralmente desenvolvidas (Computação e Engenharias) e áreas que analisam gestos dentro de um sistema linguístico (Linguística e Psicolinguística), e também o quão eficiente são seus mecanismos de busca. Para a estratégia de busca, a análise exploratória permitiu executar refinamentos nas palavras chaves usadas, incluindo as principais variações e sinônimos dos termos comumente usados na área.

- Fontes de dados – *Scopus* e *Web of Science*. Essas bases indexam diferentes bibliotecas digitais de artigos científicos, permitindo uma cobertura abrangente do universo de interesse desta RS. Duplicidade de indexação foram tratadas.
- Estratégia de busca – Foram definidas *strings* de buscas específicas para cada base, compostas pelas palavras-chave: reconhecimento, segmentação, análise, gestos, gesticulação, sintático, parser e gramática. A tabela 2 contém a expressão regular genérica que representa a *string* de busca utilizada para pesquisa. Essa expressão regular foi aplicada sobre os campos de indexação: título do estudo, resumo e palavras-chave.

Tabela 2: Expressão regular para a *string* de busca.

(recogni* **OR** segment* **OR** analys*)
AND
(gesture **OR** gestures **OR** gesticulation)
AND
(syntacti* **OR** parse* **OR** parsing **OR** gramma*)

Para a **seleção de estudos** foram definidos critérios de inclusão (CI). Tais critérios foram aplicados na análise do título e do resumo dos estudos retornados na busca. Os CIs têm por objetivo garantir critérios mínimos de qualidade, removendo estudos que não interessam ao escopo da RS.

- **CI-1**: o estudo emprega *parsers*/análise sintática como estratégia, ou parte dela, para a análise de gestos;
- **CI-2**: o estudo está disponível na íntegra na *web*;
- **CI-3**: o estudo é escrito na íntegra em língua inglesa;
- **CI-4**: o estudo está publicado em um periódico científico a partir de uma revisão por pares[2];
- **CI-5**: trata-se de estudo primário.

Para **avaliar a qualidade** dos estudos foram aplicados dois critérios:

- o estudo descreve detalhadamente a metodologia de representação de dados (sim/não);
- o estudo aplica a representação de dados em algum processo para de interpretação ou descrição de padrões, automatizado ou não (classificação, agrupamento, ambos, outros, não aplica).

[2]Considerar apenas estudos publicados em periódicos foi um critério escolhido para classificar estudos de acordo com a sua qualidade, já que espera-se que os periódicos publiquem resultados de pesquisas mais sedimentadas.

3.2 Condução

A condução dessa RS foi realizada em três etapas: identificação de estudos, seleção dos estudos e extração de dados, conforme descrito a seguir.

Para **identificação de estudos** a *string* de busca foi adaptada e aplicada a cada uma das bases selecionadas. A pesquisa foi realizada em março de 2015, e reexecutada em fevereiro 2016 para atualização dos estudos recuperados. Foi considerado como período de publicação o ano atual (janeiro de 2016) e os últimos cinco anos (2011–2015). Considerar esse período de publicação objetivou apresentar uma análise das formas de representação de dados usadas mais recentemente. Nessa primeira etapa foram recuperados 253 registros de estudos; sendo 154 da base *Scopus* e 99 da base *Web of Science*. Com o tratamento das duplicidades, o conjunto de registros recuperados indicou 184 estudos.

Na **seleção dos estudos**, os critérios CI-2 a CI-5 foram aplicados para remover os estudos que não interessam ao escopo desta RS em termos de forma, ou que não são acessíveis para leitura pelos autores desta RS. Estudos que não atenderam aos quatro critérios não foram incluídos, o que resultou em 83 remoções. Então, o título e o resumo dos 101 estudos resultantes foram submetidos ao critério CI-1, de forma que ele deveria ser atendido para que o estudo permanecesse no conjunto de estudos selecionados. Nessa etapa, 75 estudos foram rejeitados e, portanto, 26 estudos compuseram o conjunto de estudos incluídos nesta RS, representando 14% do conjunto de estudos recuperados na estratégia de busca.

Para a **extração de dados**, os estudos selecionados foram lidos na íntegra e analisados quanto aos critérios de qualidade. O primeiro critério de qualidade, que analisa se o estudo descreve detalhadamente a representação de dados, foi utilizado para excluir alguns estudos, uma vez que a descrição da representação é o foco desta RS. O segundo critério foi aplicado para indicar a profundidade com que a análise do gesto foi tratada no estudo e também para categorizá-los em termos de objetivo de análise (Tabela 3. Como resultado dessa etapa, 21 estudos permaneceram no conjunto de estudos discutidos na revisão (11% do total de estudos identificados e 81% do total de estudos selecionados).

Tabela 3: Categorização *vs.* Quantidade de estudos discutidos (#). Total: 21 estudos.

	Categorização	**#**
Classificação	[31, 19, 9, 18, 16, 30, 27, 14, 21, 17, 24, 3, 32, 4, 29]	15
Agrupamento	[13, 10]	2
Ambos	[1, 7, 15, 6]	4
Outro	-	0
Não aplica	-	0

4. RESULTADOS

Conforme ilustrado na figura 1, os estudos relacionados à análise de gestos que consideram aspectos sintáticos, e também descrevem a forma de representação de dados com um nível de detalhe que permite a compreensão da mesma, vêm recebendo mais atenção nos últimos anos, com um destaque para o ano de 2014. Ainda, observa-se que esses estudos aplicam, principalmente, algoritmos de classificação de dados para realização da análise.

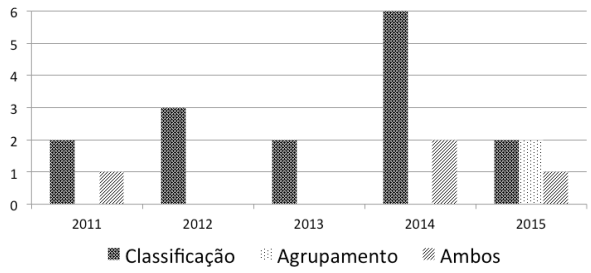

Figura 1: Distribuição dos estudos ao longo dos anos de publicação organizados por tipos de processo usados na análise dos gestos.

A análise detalhada dos 21 estudos resultantes da condução da RS permitiu compreender a representação de dados usada na área, respondendo a questão de pesquisa elaborada para esta RS (Subseção 4.1), e ainda permitiu conhecer as técnicas e estratégias de análise de dados utilizadas com essas representações a fim de elucidar como o aspecto sintático é incorporado à análise, e qual foi a eficiência de tais combinações considerando os resultados obtidos (Subseção 4.2).

4.1 Formas de Representação de Dados

A tabela 4 resume as características utilizadas por cada estudo para construir a representação de dados a ser utilizada na análise de gestos, e responde, de forma objetiva, a questão de pesquisa desta RS. Essa tabela também mostra a distribuição desses estudos ao longo dos anos, de forma que é possível observar como a prática usada na representação de dados tem evoluído no tempo. Informações de contorno e de trajetória já vêm sendo utilizadas desde os primeiros estudos em 2011 e 2012. A identificação de objetos em cena tem gerado o interesse de estudos a partir de 2013 e está relacionada a estudos que propõem análise de gestos como suporte à interpretação do significado semântico da cena. A correspondência de pixels vem sendo explorada mais recentemente, sem associação com informação espacial extraída diretamente da segmentação de membros do corpo ou objetos na cena. Correspondência de pixels foi um termo utilizado por Liu et al. [19] para representar uma técnica que analisa as imagens à procura de pixels com características semelhantes quanto à coordenadas espaciais e tonalidade da cor do pixel. Esse termo será utilizado para representar trabalhos que aplicam a mesma técnica. Também é recente o uso de mais de 15 pontos do corpo humano para representação de gestos, estratégia associada a estudos de análise de gestos em tempo real.

Em sua maioria, os estudos descrevem as representações de dados considerando que eles são derivados de um pré-processamento realizado sobre um vídeo (sequência de *frames* ou imagens dinâmicas). Tais vídeos são obtidos na internet ou são capturados com o uso de câmeras do tipo RGB ou sensores que são capazes de obter as coordenadas espaciais de pontos de interesse em um ambiente (como o MS Kinect por exemplo). Em alguns estudos a fonte de informação não é um vídeo, e sim diretamente uma imagem estática [18, 19, 30, 31]. O uso de imagens estáticas ocorre quando o problema sob estudo tem o objetivo de realizar uma análise semântica da cena (neste caso, é comum a análise de poses, consideradas como gestos estáticos ou configuração assumida por partes do corpo).

Os estudos aplicam diferentes abordagens para a extração de caraterísticas necessárias para compor a representação de dados desejada, produzindo uma representação numérica (maioria dos casos) ou uma representação simbólica cite-Rosani2014. Independente do uso de imagens estáticas ou dinâmicas, as características mais utilizadas são as coordenadas espaciais de partes do corpo humano (cabeça, pescoço, mãos, pulsos, ombros e demais junções do corpo), ou do objeto que se pretende representar. Porém, utilizar apenas esse tipo informação, como em [30] e [31], não é comum, indicando, possivelmente, que apenas a localização espacial não é informação suficientemente discriminante no contexto da análise sintática de gestos.

As representações de dados mais empregadas geralmente combinam coordenadas espaciais com outras informações. A angulação em relação ao tronco é usada por Kyan et al. [13] na representação do posicionamento de mãos, pulsos, cotovelos, ombros, centro dos ombros, quadris, centro dos quadris, joelhos, pés, tornozelos, espinha e cabeça. Esses autores combinam posição com angulações entre os 20 pontos do corpo, o tronco e os membros superiores e inferiores. Já Hachal e Ogiela [9] adicionam à representação de pontos (neste caso 15 pontos representando mãos, cotovelos, ombros, centro dos ombros, quadris, joelhos, pés, espinha e cabeça) e angulações, a informação temporal (um selo de tempo associado a cada *frame*). Posteriormente, os mesmos autores incorporam mais 5 pontos adicionais em sua representação [10], seguindo o que foi proposto em [13]. Em [3] apenas a posição das mãos e braços é combinada com a angulação da trajetória dos membros inferiores e superiores.

A informação da trajetória das mãos e da posição de partes do corpo é explorada por Abid et al. [1], na tarefa de reconhecimento e classificação de gestos em tempo real, e por Liu et al. [17], no reconhecimento de atividades humanas complexas. A trajetória junto com o posicionamento de objetos com o qual as pessoas estão interagindo nas cenas capturadas compõem a representação de dados usada por Lücking et al. [21] e Rosani et al. [27] também na tarefa de reconhecimento e classificação de gestos em tempo real.

Outra informação que vem sendo utilizada para estabelecer representações para os dados é a matriz de cores das imagens. Essa característica é usada em estudos que objetivam resolver tarefas que exigem a identificação do posicionamento de partes do corpo, como em [18], e para resolver tarefas de reconhecimento de gestos, como em [7, 6, 29]. Esses últimos estudos usam matriz de cores para encontrar o contorno das mãos. O contorno de membros e objetos também pode ser representado por meio de informações extraídas da correspondência de pixels, como feito por Liang et al. [15] e Liu et al. [19], que utilizam apenas essa informação para realizar o reconhecimento de gestos e a identificação do corpo, respectivamente.

A capacidade de determinar o contorno dos membros e objetos tem sido explorada também para utilizar a informação do formato desses itens na construção de representações de dados. Em [32] a informação de contorno é utilizada para selecionar pontos ao redor do corpo ou do objeto sob análise e esses pontos são utilizados na representação dos dados a serem usados na análise de gestos.

Enfim, considerando o contexto de gestos em termos de expressões faciais, Liu et al. [16] utilizam 79 pontos do rosto junto com o formato do nariz e formato e movimento dos

Tabela 4: Representações de dados usadas nos estudos, organizados por ano de publicação

Estudo	Ano	Características utilizadas na Representação
[6]	2011	Contorno da mão determinado por informação de cores e representados por pontos chave
[29]	2011	Contorno da mão determinado por informação de cores e posições* do centro da mão ao longo do tempo (trajetória)
[32]	2011	Contorno do corpo e posição associada ao contorno
[3]	2012	Posição das mãos e braços, angulação e sequências de angulações (variações nas angulações)
[17]	2012	Trajetória de pontos centrais de cada parte do corpo humano
[24]	2012	Posicionamento e formato das sobrancelhas, olhos, nariz e boca
[14]	2013	Formato, orientação, movimento e posição da mão
[21]	2013	Posição e contorno da mão, interação da mão com objetos, o contorno e a trajetória do objeto
[4]	2014	Posição e formato dos das sobrancelhas, olhos, nariz, boca e posição da cabeça
[7]	2014	Contorno da mão determinado por informação de cores e representados por pontos chave
[9]	2014	Coordenadas (3D) de 15 pontos do corpo, angulações entre esses pontos, o tronco e os membros superiores e inferiores; *timestamp* dos *frames* usados para estabelecer a ordem dos *frames*
[15]	2014	Correspondência de pixels
[16]	2014	79 pontos da face, sobre os quais são estimados os ângulos de rotação do rosto, o formato do nariz e o formato e movimento dos olhos
[18]	2014	Posição e distância entre 26 pontos do corpo, cor associada a imagem na região de cada parte do corpo
[27]	2014	Pontos de interesse definidos empiricamente e a trajetória da pessoa pelo espaço considerando a aproximação a esses pontos de interesse, representados por meio de uma sequência de símbolos
[30]	2014	Posição do tronco, cabeça, braços, coxas, antebraços e canelas
[1]	2015	Posição e angulação do centro da mão ao longo do tempo (trajetória)
[10]	2015	Coordenadas (3D) de 20 pontos do corpo e o *timestamp* dos *frames*
[13]	2015	Coordenadas (3D) de 20 pontos do corpo e angulações entre o tronco e os membros superiores e inferiores
[19]	2015	Correspondência de pixels
[31]	2015	Posição das juntas dos membros, pulsos, cotovelos, ombros, pescoço e cabeça

* O termo "posição" é usado seguindo a nomenclatura no estudo original. Entretanto, ela diz respeito a coordenadas espaciais.

olhos para reconhecer as expressões. Sendo que o formato é derivado desses 79 pontos iniciais. A mesma tarefa é realizada pelos autores de [24] e [4] porém, nesses casos, ainda combinando com o formato de sobrancelhas e boca e suas coordenadas espaciais.

4.2 Discussão

Como apresentado na subseção 4, dentre as propostas de representação de dados referente a gestos levantadas nessa RS, foi identificada uma tendência no uso de coordenadas espaciais de partes do corpo humano combinadas com outras características como contorno, angulação dos membros em relação ao tronco e trajetória dos membros ao longo do tempo. Em menor quantidade, foram exploradas características como correspondência de pixels e objetos na cena.

Todos os trabalhos empregam a representação dos dados em um procedimento de análise de gestos, dinâmicos ou estáticos, automatizado ou não, seja com o uso de procedimentos referentes a classificação, agrupamento de dados ou ambos. É comum os trabalhos analisarem o resultado de suas abordagens em termos de acurácia ou erro em relação a um resultado esperado. A tabela 5 traz um resumo sobre as técnicas utilizadas para análise dos gestos, o melhor desempenho obtido em testes em termos acurácia (%) e o aspecto sintático tratado pela abordagem apresentada.

Em [29], por exemplo, é proposta uma metodologia para reconhecimento e classificação de gestos manuais. A identificação das mãos é feita com o algoritmo *Adaptive Skin Threshold*, que detecta o contorno da mão com base nas cores da imagem, usando o espaço de cores YCbCr. A informação da cor, junto com uma série temporal contendo as posições do centro das mãos são passadas aos *Hidden Markov Models* (HMM) para determinação do gesto. Esse estudo obteve acurácia máxima de 93,86%.

Os autores em [32] apresentam uma abordagem para reconhecimento de poses humanas e não-humanas (poses de cavalos). Nesse trabalho dezenas de pontos no contorno do corpo sob análise são identificados e organizados hierarquicamente, de forma que a cabeça é o primeiro nível hierárquico e as extremidades são os últimos, construindo o chamado *Hierarchical Configurable Deformable Template*. As angulações e relacionamentos entre os pontos são apresentados a um algoritmo probabilístico de aprendizado supervisionado proposto para segmentar as poses. Todas essas informações ajudam na estruturação da hierarquia. A escolha dos parâmetros e o treinamento do algoritmo é feito usando a estratégia *Max Margin*. Essa abordagem alcançou acurácia de 94,5%. Quando a informação de nível hierárquico não é passada ao algoritmo, usando apenas um grafo AND/OR, a acurácia de segmentação chega a 94,8%, mas com custo computacional maior.

Dardas e Georganas [6] implementam um sistema para reconhecimento de gestos manuais em tempo real, utilizado para controlar aplicações computacionais. O sistema capta as imagens das mãos com o uso de uma *webcam*. O *Hue Saturation Value Color Model* é usado para detectar a cor da pele e definir o contorno da mão. Em seguida é aplicada a *Scale Invariant Feature Transform* para extrair pontos chave que representem a imagem. Cada ponto chave é organizado em um histograma e mapeado para um vetor de 128 dimensões. Esses vetores são apresentados ao algoritmo *K-means* para determinação de centróides que representam conjuntos de característica. Os grupos de dados

Tabela 5: Resumo (organizado por ano): técnicas utilizadas, desempenho obtido e aspecto sintático analisado

Estudo	Acurácia	Técnicas	Aspecto(s) Sintático(s)
[6]	96,23	SVM e Gramáticas	Uso de gramáticas para tomada de decisão
[29]	93,86	HMM	Estados e transições; predição estruturada
[32]	94,5	Modelo probabilístico e Grafos AND/OR	Hierarquia na representação das poses
[3]	95	Segmental HMM	Estados e transições; predição estruturada
[17]	93,4	Markov Semantic Model	Estados e transições; predição estruturada
[24]	91,76	HMM, SVM	Estados e transições; predição estruturada
[14]	–	Inspeção visual (humana)	Analisa características linguísticas
[21]	–	Gramáticas	Uso de gramáticas para tomada de decisão
[4]	100	Rede neural artificial recorrente	Análise de sequências (temporais)
[7]	96,23	K-means, SVM e Gramáticas	Uso de gramáticas para tomada de decisão
[9]	98,5	Classificador GDL e Gramáticas	Uso de gramáticas para tomada de decisão
[15]	89,29	Random Decision Forest, Super Markov Random Fields e Simple Linear Iterative Clustering	Estados e transições; predição estruturada
[16]	97,6	Ranking SVM e Conditional Random Fields	Predição estruturada
[18]	–	Simple Linear Iterative Clustering, Markov Random Fields e SVM	Estados e transições; predição estruturada
[27]	85,5	Gramática Livre de Contexto	Uso de gramáticas para tomada de decisão
[30]	–	Markov Chain Monte Carlo e Gramáticas	Uso de gramáticas para tomada de decisão
[1]	98,5	Gramática Linear Estocástica	Uso de gramáticas para tomada de decisão
[10]	100	Gramática e K-means	Uso de gramáticas para tomada de decisão
[13]	100	Self Organizing Maps	Predição estruturada
[19]	–	Gramática	Uso de gramáticas para tomada de decisão
[31]	–	HMM e Structural Tree	Estados e transições; predição estruturada

associados a cada centróide compõem a *bag-of-words* (BOW) que representa a imagem, sendo que o conjunto das BOWs representa o modelo *bag-of-features* (BOF). Cada BOW é passada à uma *Support Vector Machine* (SVM) para determinação da classe da configuração da mão. Por fim, cada sequência de dois *frames* é submetido à uma gramática para determinação do gesto, que é traduzido em um comando para a aplicação. Esta abordagem alcançou uma acurácia de 96,23%. Também com o objetivo de controlar aplicações, no estudo realizado por Abid et al. [1] é utilizada a mesma estratégia de representação de dados do estudo de Dardas e Georganas [6], com o diferencial que para determinação do comando, Abid et al. implementam uma Gramática Linear Estocástica, conseguindo acurácia de 98,5%.

O estudo de Caramiaux et al. [3] apresenta uma abordagem para reconhecimento de gestos feitos por músicos ao manipular instrumentos, mais precisamente um clarinete. Os autores usam a posição das mãos e braços, suas angulações e consideraram sequências de angulações para determinar a variação do gesto. Essa representação de dados é usada em combinação com um *Segmental Hidden Markov Model*, modelado a partir de gestos referentes a movimentos básicos do contexto sob análise. A identificação de descansos é caracterizada pela ausência de movimentos. Como avaliação, é utilizada a distância Euclidiana para medir a diferença (variação) entre o resultado obtido e o esperado, e a quantificação do desempenho se dá em termos de porcentagens de divergência (5% de divergência no pior caso).

Uma abordagem para implementação de um sistema de reconhecimento de expressões faciais usadas na comunicação em língua se sinais é apresentada por Nguyen e Ranganath [24]. A identificação das características descritivas da expressão facial é realizada com a combinação de um sistema de reconhecimento, chamado de *KLT tracker*, com um *Probabilistic Principal Component Analysis* (PPCA). Com essa estratégia são obtidas informações sobre o posicionamento e formato das sobrancelhas, olhos, nariz e boca. Os dados coletados são apresentados para HMMs com o objetivo de avaliar a probabilidade de cada tipo de expressão facial. Essas probabilidades e as demais características da face compõem uma nova representação de dados que é submetida à uma SVM para determinação da expressão facial dentre seis expressões possíveis. A abordagem alcançou uma acurácia de 91,76%.

Dentre o conjunto de estudos analisados, o primeiro a considerar a trajetória na modelagem da forma de representação do gesto foi [17]. Nele é descrito um *framework* para reconhecimento de atividades e comportamentos humanos complexos. A partir de uma sequência de *frames* extrai-se a trajetória de pontos centrais de cada parte do corpo humano com um sistema conhecido como Affine SIFT. A partir desses pontos de interesse são extraídos três tipos de informação: o sinal das trajetórias usando descritores de Fourier; as distâncias entre os pontos iniciais e finais das trajetórias de cada parte do corpo; a média de todos os valores de todos os pontos de interesse como uma medida de "Aparência". Também são coletados pontos de interesse do cenário que ajudaram a eliminar a influência da movimentação da câmera. Então é feita uma seleção de características com um *Markov Semantic Model*. O resultado desta seleção é então submetido à uma SVM para fazer a classificação das atividades. Essa abordagem alcançou 93,4% de acurácia.

Em [21] é apresentado um sistema para reconhecimento do discurso e dos gestos de usuários, mas não são apresentados testes que permitam a análise de desempenho da abordagem. O estudo está baseado no algoritmo *Bielefeld Speech and Gesture Alignment*, que faz uso de dados parcialmente anotados (por especialistas). Os autores usam câmeras para

coletar imagens e dessas imagens extraem informações sobre as mãos: a posição, se está interagindo com algum objeto, o contorno do objeto, o contorno da mão e a trajetória do objeto. O sistema realiza a segmentação do gesto em termos de fases do gesto (preparação, stroke e retração). Essas informações são também combinadas com dados sobre o discurso para o reconhecimento do gesto.

O estudo realizado por Rosani et al. [27] é um dos primeiros a utilizar a representação simbólica na construção da representação dos gestos. São selecionados pontos de interesse definidos empiricamente de acordo com as preferências dos usuários. Esses pontos de interesse são convertidos em símbolos e, de acordo com a trajetória da pessoa para perto dos pontos de interesse são criados vetores contendo a sequência de símbolos visitados. Essas sequências são informadas para uma Gramática Livre de Contexto responsável por aplicar regras e definir a atividade realizada pela pessoa. A acurácia obtida nessa abordagem foi de 85,5%.

O estudo de Liu et al. [16] se destaca por utilizar 79 pontos do rosto, sobre os quais são estimados os ângulos de rotação do rosto, o formato do nariz e o formato e movimento dos olhos. É usado um *Ranking SVM* para escolha das características que serão utilizadas no reconhecimento dos gestos. Então, para reconhecimento do gesto da cabeça, essas informações são passadas para um *Conditional Random Fields* que alcançou uma acurácia de reconhecimento de 97,6%.

Os trabalhos [15] e [19], que utilizam correspondência de pixels em seu modelo não obtiveram bons resultados. Em [15] a informação de cada pixel é passada à uma *Random Decision Forest* (RDF) para classificação. Em seguida é usada uma *Super Markov Random Fields* para refinar essa classificação usando a informação dos pixels adjacentes. O resultado é passado ao algoritmo *Simple Linear Iterative Clustering* para definição das partições da imagem da mão e identificação da pose da mão. Para identificação de um gesto, é proposto que as posições e profundidades de pixels anteriores sejam adicionados aos dados passados para a RDF, alcançando acurácia máxima de 89,29%.

Os estudos [18], [9], [10] e [13] apresentam abordagens que usam 15 ou mais pontos do corpo na representação dos dados. Dentre eles, o que obteve resultados mais expressivos foi Kyan et al. [13] ao propor um *framework* para captar passos de dança de *ballet* e categorizá-los, possibilitando que os dançarinos melhorem seu treinamento de acordo com o *feedback* do sistema provido por uma ferramenta do tipo CAVE (espaço com 4 projetores em 4 paredes de uma sala mostrando o passo realizado e o passo que deveria ser realizado). O sensor MS Kinect é usado para obter a posição tridimensional dos pontos e, visto que o sensor capta coordenadas tridimensionais, a representação vetorial final contou com 60 características e mais aquelas referentes angulações (veja subseção 4.1). Essas características são submetidas à uma variação do algoritmo Self Organizing Maps chamada *Spherical SOM*. O melhor resultado obtido alcançou acurácia de 100%. Outro estudo que reportou o alcance de uma acurácia de 100% foi o de Caridakis et al. [4]. Esses autores propõem uma abordagem para reconhecimento de expressões faciais usadas na comunicação por língua de sinais. Eles utilizam apenas características de posição e de formato de elementos da face combinados a um algoritmo de rede neural artificial recorrente com 1 atraso no tempo.

Como é possível observar, formas de representação de dados similares são combinadas a diferentes técnicas de análise de gestos, e a depender do problema sob resolução, estratégias similares obtém resultados bem diferentes em termos de acurácia. É importante ressaltar que não há intenção de induzir o leitor a uma comparação entre os estudos em termos de acurácia, visto que cada iniciativa resolve problemas de complexidades diferentes, realiza experimentos sobre conjuntos de dados distintos e aplica a avaliação de desempenho a partir de estratégias mais ou menos otimistas ou exigentes. O intuito de apresentar o desempenho das abordagens é meramente descritivo. A observação das mais diferentes estratégias de soluções de problemas ligados a um mesmo tipo de dado (neste caso dados gestuais) é útil para elucidar a área e motivar o desenvolvimento de novas iniciativas.

5. CONCLUSÃO

Neste artigo foi apresentada uma RS na área de análise sintática de gestos, com foco na identificação e discussão de formas de representação de gestos. Foram selecionados 21 estudos primários, todos publicados em periódicos, no período de janeiro de 2011 a janeiro de 2016.

Na análise dos estudos, foi observada a tendência para uso de informações espaciais na representação de gestos, porém de forma combinada a uma ou mais características que informam angulações, contornos e trajetórias de partes do corpo. Informações temporais são também usadas, porém para construção de sequências de dados (como no caso da trajetória) ou de forma implícita nos algoritmos de análise. Foi também observada uma forte tendência no uso de algoritmos, ou procedimentos não automatizados, referentes à solução de tarefas de classificação como forma de tomada de decisão na análise dos gestos. Em relação às técnicas de análise aplicadas, predomina o uso de HMMs e gramáticas.

A apresentação dos estudos incluiu menção ao desempenho obtido em cada um, porém alguns deles não apresentaram avaliação de resultados de forma quantitativa. Em relação àqueles que apresentaram seus resultados quantitativamente, predomina o uso da medida de acurácia (taxa de acerto nos experimentos realizados). No entanto, devido à natureza diversa dos problemas resolvidos em cada estudo e das diferentes estratégias para aferição dos resultados, não foi possível realizar uma comparação que pudesse indicar as melhores formas de representação, ou as melhores combinações de representação e técnicas de análise sintática de gestos.

Na avaliação de uma RS é preciso considerar que há algumas ameaças à validade do estudo. No caso da presente RS, são duas as principais ameaças:

- (i) não foi realizado um teste do protocolo no que diz respeito à recuperação de artigos "esperados";
- (ii) há um caráter de subjetividade inerente ao processo de extração de dados, pois nem todos os estudos apresentam de forma clara e objetiva o processo de condução de seus experimentos, ficando a encargo dos autores da RS tomar algumas decisões com base em seu conhecimento de especialista.

A RS apresentada neste artigo representa o mapeamento de um recorte da pesquisa documentada na área de análise sintática de gestos. Considerando a *string* de busca utilizada, ainda há um potencial de análise de 14 estudos primários referentes à publicações em conferências da área. Os autores têm a intenção de realizar a análise desses estudos, de forma sistemática, e também documentar os resultados obtidos.

6. REFERÊNCIAS

[1] M. Abid, E. Petriu, and E. Amjadian. Dynamic sign language recognition for smart home interactive application using stochastic linear formal grammar. *IEEE Trans. on Inst. and Measurement*, 64(3):596–605, mar. 2015.

[2] J. Biolchini, P. G. Mian, A. C. C. Natali, and G. H. Travassos. Systematic rev. in soft. eng. Technical report, System Engineering and Computer Science Department COPPE/UFRJ, Technical Report ES–679–05, 2005.

[3] B. Caramiaux, M. M. Wanderley, and F. Bevilacqua. Segmenting and parsing instrumentalists' gestures. *J. of New Music Research*, 41(1):13–29, 2012.

[4] G. Caridakis, S. Asteriadis, and K. Karpouzis. Non-manual cues in automatic sign language recognition. *Personal and Ubiquitous Comput.*, 18(1):37–46, 2014.

[5] C. Chen, R. Jafari, and N. Kehtarbavaz. A survey of depth and inertial sensor fusion for human action recognition. *Multimedia Tools App.*, 2015. to appear.

[6] N. H. Dardas and N. D. Georganas. Real-time hand gesture detection and recognition using bag-of-features and support vector machine techniques. *IEEE Trans. Inst. and Measurement*, 60(11):3592–3607, 2011.

[7] N. H. Dardas, J. M. Silva, and A. El-Saddik. Target-shooting exergame with a hand gesture control. *Multimedia Tools Appl.*, 70(3):2211–2233, 2014.

[8] T. G. Dietterich. Machine learning for sequential data: A review". In *Proc. of Joint Structural, Syntactic, and Statistical Pattern Recogn. International Workshops*, pages 15–30. Springer Berlin Heidelberg, 2002.

[9] T. Hachaj and M. R. Ogiela. Rule-based approach to recognizing human body poses and gestures in real time. *Multimedia Syst.*, 20(1):81–99, 2014.

[10] T. Hachaj and M. R. Ogiela. Full body movements recognition – unsupervised learning approach with heuristic r-gdl method. *Digital Signal Proces.*, 46:239–252, 2015.

[11] J. Han, L. Shao, D. Xu, and J. Shotton. Enhanced computer vision with microsoft kinect sensor: A review. *IEEE Trans. on Cyber.*, 43(5):1318–1334, 2013.

[12] B. A. Kitchenham. Systematic review in software engineering: Where we are and where we should be going. In *Proc. of the 2Nd Int. Workshop on Evidential Assessment of Soft. Tech.*, EAST '12, pages 1–2, New York, NY, USA, 2012. ACM.

[13] M. Kyan, G. Sun, H. Li, L. Zhong, P. Muneesawang, N. Dong, B. Elder, and L. Guan. An approach to ballet dance training through ms kinect and visualization in a cave virtual reality environment. *ACM Trans. Intell. Syst. Technol.*, 6(2):23:1–23:37, mar. 2015.

[14] S. H. Ladewig and J. Bressem. New insights into the medium hand: Discovering recurrent structures in gestures. *Semiotica*, 2013(197):203–231, Oct. 2013.

[15] H. Liang, J. Yuan, and D. Thalmann. Parsing the hand in depth images. *IEEE Trans. on Multimedia*, 16(5):1241–1253, aug. 2014.

[16] J. Liu, B. Liu, S. Zhang, F. Yang, P. Yang, D. N. Metaxas, and C. Neidle. Non-manual grammatical marker recognition based on multi-scale, spatio-temporal analysis of head pose and facial expressions. *Image Vision Comput.*, 32(10):671–681, 2014.

[17] J. Liu, X. Wang, T. Li, and J. Yang. Spatio-temporal semantic features for human action recognition. *KSII Trans. on Internet and Inf. Syst.*, 6(10):2632–2649, 2012.

[18] S. Liu, J. Feng, C. Domokos, H. Xu, J. Huang, Z. Hu, and S. Yan. Fashion parsing with weak color-category labels. *IEEE Trans. on Multimedia*, 16(1):253–265, jan. 2014.

[19] S. Liu, X. Liang, L. Liu, K. Lu, L. Lin, X. Cao, and S. Yan. Fashion parsing with video context. *IEEE Trans. on Multimedia*, 17(8):1347–1358, aug. 2015.

[20] R. Lun and W. Zhao. A survey of applications and human motion recognition with microsof kinect. *Int. J. of Pattern Recogn. and Art. Intel.*, 29(5):1–48, 2015.

[21] A. Lücking, K. Bergmann, F. Hahn, S. Kopp, and H. Rieser. Data-based analysis of speech and gesture: the Bielefeld Speech and Gesture Alignment corpus (SaGA) and its applications. *Journal on Multimodal User Interfaces*, 7(1-2), 2013.

[22] R. C. B. Madeo, P. K. Wagner, and S. M. Peres. A review of temporal aspects of hand gesture analysis applied to discourse analysis and natural conversation. *Int. J. of C. Sci. & Inf. Tech.*, 5(4), dec. 2013.

[23] S. Mitra and T. Acharya. Gesture recognition: A survey. *IEEE Trans. on Syst., Man, and Cyber., Part C: App. and Rev.*, 37(3):311–324, May 2007.

[24] T. D. Nguyen and S. Ranganath. Facial expressions in american sign language: Tracking and recognition. *Pattern Recogn.*, 45(5):1877–1891, 2012.

[25] S. C. W. Ong and S. Ranganath. Automatic sign language analysis: A survey and the future beyond lecical meaning. *IEEE Trans. on Pattern Analysis and Machine Intel.*, 27(6):873–891, jun. 2005.

[26] P. K. Pisharady and M. Saerbeck. Recent methods and databases in vision-based hand gesture recognition: A review. *Computer Vision and Image Underst.*, 141:152–165, 2015.

[27] A. Rosani, N. Conci, and F. G. B. D. Natale. Human behavior recognition using a context-free grammar. *J. Electronic Imaging*, 23(3):033016, 2014.

[28] N. A. Smith. *Linguistic Structure Prediction*. Morgan & Claypool, 2011.

[29] W. Xu and E.-J. Lee. Hand gesture recognition using improved hidden markov models. *J. of Korea Multimedia Soc.*, 14(7):866–871, 2011.

[30] X. Zhang, C. Li, W. Hu, X. Tong, S. J. Maybank, and Y. Zhang. Human pose estimation and tracking via parsing a tree structure based human model. *IEEE Trans. Syst., Man, and Cyber.: Systems*, 44(5):580–592, 2014.

[31] L. Zhao, X. Gao, D. Tao, and X. Li. Tracking human pose using max-margin markov models. *IEEE Trans. on Image Proces.*, 24(12):5274–5287, 2015.

[32] L. Zhu, Y. Chen, C. Lin, and A. L. Yuille. Max margin learning of hierarchical configural deformable templates (hcdts) for efficient object parsing and pose estimation. *Int. J. of Comp. Vision*, 93(1):1–21, 2011.