Aplicação do descritor HOG e classificador SVM no reconhecimento de poses humanas em imagens de profundidade

A. Márcio A. Almeida [1], C. David B. Borges [2], Iális C. de Paula Júnior [3]

[1 2 3]Engenharia de Computação – Universidade Federal do Ceará (UFC)
Campus Mucambinho – Sobral – CE – Brasil

[1] marcio.albu@alu.ufc.br, [2] david.borges@protonmail.com,
[3] ialis@sobral.ufc.br

Abstract. *Automatic gesture and pose recognition by machine learning is a significant challenge. This work proposes a simple method for human pose recognition using exclusively depth images. The method consists in background removal, feature extraction via histograms of oriented gradients (HOG) and classification through support vector machines (SVM). The method is validated in a test with five human poses, in which a 95.6% success rate is obtained.*

Resumo. *O reconhecimento automático de gestos e poses humanas por meio de aprendizado de máquina é um desafio significativo. O presente trabalho propõe uma metologia simples para reconhecimento de poses humanas exclusivamente através de imagens de profundidade. O método consiste de remoção de plano de fundo, extração de características de poses via Histogramas de Gradientes Orientados (HOG) e classificação via Máquina de Vetores de Suporte (SVM). A sequência de processamento proposta foi validada em um teste com cinco classes de poses humanas, no qual foi obtida uma taxa de acerto de 95.6%.*

1. Introdução

A visão e a percepção humana tem função no processo de reconhecimento dos objetos que constituintes um cenário [Lima et al. 2014]. Essa capacidade também se aplica ao reconhecimento visual de poses e expressões humanas. Infelizmente é complexo traduzir essa funcionalidade para sistemas digitais, visto que a ocorrência desse processo no cérebro não é totalmente compreendida. Para esse fim, as técnicas mais recentes de compreensão visual de gestos e poses fazem uso de processamento de imagens e visão computacional de forma restrita a solucionar problemas específicos. Por exemplo, limitando a quantidade de gestos ou poses humanas reconhecíveis pelo sistema. Isso ocorre pois, até o momento, não existem sistemas de análise de imagens complexas que funcionem de maneira generalizada [Szeliski 2010, Gonzalez and Woods 2010].

O presente trabalho estuda a aplicação de descritores de forma no processo de classificação de poses humanas em imagens de profundidade, onde essa imagens contém as referências das distâncias dos objetos em relação a câmera. Mais especificamente, a adequação do descritor denominado Histograma de Gradientes Orientados (HOG, do inglês *Histogram of Oriented Gradients*) é verificada em conjunto com um classificador baseado em Máquinas de Vetores de Suporte (SVM, do inglês *Support Vector Machine*).

A combinação dessas técnicas é testada sobre uma base de dados de imagens de profundidade capturadas a partir do sensor *Kinect*. Adicionalmente, este trabalho descreve uma sequência de pré-processamento para extração de regiões de interesse e remoção de ruído de fundo em imagens de profundidade.

A seção 2 menciona alguns trabalhos já realizados acerca do tema de reconhecimento de gestos e poses humanas em imagens de profundidade. Os métodos de pré-processamento e detalhes sobre o descritor Histogramas de Gradientes Orientados e classificador Máquinas de Vetores de Suporte são explanados na seção 3. Detalhes sobre o experimento realizado e resultados são expostos na seção 4.

2. Trabalhos relacionados

A ideia de explorar as aplicações de imagens de profundidade para o reconhecimento de padrões de poses e comportamentos humanos, utilizando métodos do aprendizado supervisionado, são apresentadas nas obras [Biswas and Basu 2011], [Boutella et al. 2015] e [Amaral et al. 2017]. No trabalho desenvolvido por [Biswas and Basu 2011], são utilizados um descritor de histogramas e um classificador Máquinas de Vetores de Suporte para distinção entre 8 gestos. A principal conclusão do trabalho é que a utilização de imagens de profundidade acelera o processo de extração de características em relação a métodos que usam imagens RGB.

O trabalho de [Boutella et al. 2015] trata da classificação de identidade, gênero e etnia utilizando imagens de profundidade de faces humanas. Os autores apresentam quatro métodos de extração de características a partir de descritores diversos. São eles: *Local Binary Patterns* (LBP) [Huang et al. 2011], *Local Phase Quantization* (LPQ) [Nanni et al. 2012], Histograma de Gradientes Orientados (HOG) e *Binarized Statistical Image Features* (BSIF) [Kannala and Rahtu 2012]. O trabalho sumariza os resultados de reconhecimento facial e de gênero e faz uma comparação entre técnicas baseadas em imagens D, RGB e RGB-D. Neste caso, D são imagens de profundidade, enquanto RGB-D são imagens com informações de cor e profundidade simultaneamente.

A pesquisa de [Amaral et al. 2017] aplica o reconhecimento de gestos em imagens de profundidade à compreensão da Linguagem Brasileira de Sinais (Libras). [Rios 2012] O descritor utilizado é a Transformada de Distância [Szeliski 2010], uma medida que determina a distância de cada pixel em relação à borda da região a que pertence. Uma rede neural convolucional é usada para classificação. Os autores obtém uma taxa de acerto de 96.42% para 14 diferentes configurações de gestos [Russell and Norvig 2010].

3. Método

Neste trabalho, as imagens de profundidade obtidas do *Kinect* passaram por etapas de segmentação e filtragem antes da extração de características e posterior classificação. Esses procedimentos tornam mais prático analisar uma imagem e reconhecer todos os objetos constituintes [Szeliski 2010]. Esta seção trata das técnicas de pré-processamento aplicada à imagem, extração do descritor Histogramas de Gradientes Orientados e treinamento e classificação usando Máquinas de Vetores de Suporte.

3.1. Pré-processamento

Muitas técnicas utilizadas em imagens RGB podem ser aplicadas em imagens de profundidade. Como pode ser observado na Figura 1, um exemplo de imagem de entrada,

que contém uma cadeira e objetos de fundo, passa pela sequência de pré-processamento utilizada neste trabalho. Em primeiro lugar, uma segmentação é realizada através de um processo de limiarização para remover os objetos de fundo, ou seja, aqueles cuja distância em relação ao sensor é maior que um determinado limiar. Em seguida, a imagem segmentada é submetida a uma suavização através do filtro mediana [Morais and Vieira 2013].

(a) Imagem original. (b) Segmentação. (c) Filtragem Mediana (5x5).

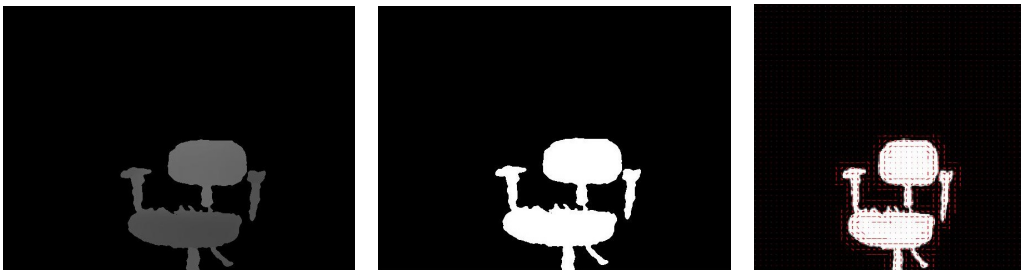
Figura 1. Pré-processamento da imagem de profundidade.

3.2. Histograma de Gradientes Orientados (HOG)

O Histograma de Gradientes Orientados (HOG, do inglês *Histogram of Oriented Gradients*) teve como sua primeira aplicação a detecção de pedestres em imagens [Triggs and Dalal 2005]. Este método tem como objetivo extrair informações referentes à orientação das arestas existentes em uma imagem, sendo estas arestas calculadas através de métodos de detecção de bordas.

A ideia essencial é que a aparência e forma do objeto local dentro de uma imagem pode ser descrita pela distribuição de gradientes de intensidade ou direções de suas bordas. A imagem é dividida em pequenas regiões conectadas chamadas células. Para os *pixels* dentro de cada célula, um histograma de direções de gradiente é compilado. O descritor é, então, a concatenação desses histogramas, o que efetivamente traduz a imagem em um vetor de dados.

Na Figura 2 são exibidos exemplos da etapa de extração do descritor Histogramas de Gradientes Orientados a partir da imagem pré-processada (Figura 2(a)). O momento da Figura 2(b) é caracterizado para definir as regiões da imagem em pretos e brancos, neste caso, para destacar apenas o objeto de interesse em branco. A Figura 2(c) é uma ilustração das orientações do gradiente sobre o contorno do objeto em questão.



(a) Imagem após filtro mediana. (b) Binarização. (c) Fluxo do Histograma.

Figura 2. Obtenção do descritor HOG.

3.3. Máquina de Vetores de Suporte (SVM)

O descritor Histogramas de Gradientes Orientados é capaz de representar numericamente a pose de um humano em frente a uma câmera. Faz-se necessário, então, um método de discriminar as características extraídas e determinar a que classe pertencem. Neste trabalho, utiliza-se o classificador denominado Máquina de Vetores de Suporte (SVM, do inglês *Support Vector Machine*).

A Máquina de Vetores de Suporte transforma o problema de classificação em múltiplas classes em diversos problemas de classificação binária. Para isso múltiplos classificadores Máquinas de Vetores de Suporte são treinados para diferenciar entre uma classe e todas as outras (abordagem *one versus all*). A técnica obtém resultados ótimos caso as classes sejam separáveis [Russell and Norvig 2010, Ng 2016].

A aplicação de Máquinas de Vetores de Suporte é robusta diante de dados de grande dimensão, como fornecidos pelos descritores. Outra característica atrativa é o processo de classificação rápido, no ponto de vista computacional [Bragatto et al. 2016].

A configuração de treinamento da Máquinas de Vetores de Suporte é composta por duas partes importantes: (1) os descritores Histogramas de Gradientes Orientados que serão usados no treinamento são extraídos a partir de suas respectivas imagens; (2) seus rótulos (*label*), que têm como função ”nomear”o gesto, são atribuídos a cada vetor de características. Como demonstrado na figura 3, cada gesto foi relacionado com seu rótulo específico. O parâmetro de custo utilizado foi 10, com gamma igual a 1 [Kittipat's 2013].

(a) Mão esquerda. (b) Mãos para frente. (c) Olá.

(d) Palmas. (e) Rock.

Figura 3. Exemplos de gestos para treinamento da SVM.

4. Experimentos e resultados

Os experimentos realizados neste trabalho utilizaram uma base de dados própria contendo 1000 imagens, todas com a mesma dimensão de 200x200 *pixels*, separadas em 5 classes de poses, exibidas na Figura 3. Para cada pose, 200 imagens de profundidade foram capturas utilizando o sensor *Kinect*. Foi feita uma divisão treino/teste de 50%, ou seja, 500 imagens usadas para treino da Máquinas de Vetores de Suporte e 500 para teste. O vetor de característica do descritor tem um tamanho de 86436 elementos e dimensão da célula para os cálculos foi 4x4 *pixels*.

Como pode-se observar na matriz de confusão exibida na Tabela 1, temos uma média de acerto de 95.6% utilizando o classificador Máquinas de Vetores de Suporte, isso para 500 imagens de 5 gestos treinados. A maior parte dos erros neste teste foram entre gestos semelhantes, por exemplo mão esquerda e mãos para a frente.

Tabela 1. Resultado da aplicação do descritor HOG mais classificador SVM.

	Mão esquerda	Mãos para frente	Olá	Palmas	Rock
Mão esquerda	97	0	0	3	0
Mãos para frente	10	90	0	0	0
Olá	2	0	98	0	0
Palmas	6	0	0	94	0
Rock	1	0	0	0	99

5. Conclusão

Apesar do desafio que envolve o reconhecimento automático de gestos e poses humanas através de aprendizado de máquina, a popularização de novas tecnologias, como sensores para aquisição de imagens de profundidade, fornece novas possibilidades e soluções. Este trabalho apresentou uma sequência de processos para reconhecimento de poses humanas através de imagens de profundidade. Foram apresentadas técnicas de remoção de plano de fundo, extração de características de poses via Histogramas de Gradientes Orientados e classificação via Máquina de Vetores de Suporte. A sequência de processamento proposta foi validada em um teste com cinco classes de poses humanas em uma base de dados com 1000 imagens, no qual foi obtida uma taxa de acerto de 95.6%.

A perspectiva do projeto é a melhoria dos métodos de processamento apresentados para testes de reconhecimento de um maior número de classes. Através dessas otimizações, pretende-se que seja possível realizar classificação de gestos com acurácia suficiente para implementação de sistemas de tradução da Linguagem Brasileira de Sinais.

Referências

Amaral, L., Lima, G., Vieira, T., and Vieira, T. (2017). Reconhecimento de gestos estáticos da mão usando a transformada de distância e aplicações em libras. Universidade Federal de Alagoas.

Biswas, K. and Basu, S. K. (2011). Gesture recognition using microsoft kinect.

Boutella, E., Hadid, A., Bengherabi, M., and Ait-Aoudia, S. (2015). On the use of kinect depth data for identity, gender and ethnicity classification from facil images. *ELSEVIER*.

Bragatto, T. A. C., Ruas, G., and Lamar, M. (2016). Uma comparação entre redes neurais artificiais e máquinas de vetores de suporte para reconhecimento de posturas manuais em tempo-real. pages 1–6.

Gonzalez, R. and Woods, R. (2010). ***Processamento Digital de Imagens***. Pearson Prentice Hall, São Paulo, 3 edition.

Huang, d., Shan, C., Ardabilian, M., Wang, Y., and Chen, L. (2011). Local binary patterns and its application to facial image analysis: A survey. 41:765–781.

Kannala, J. and Rahtu, E. (2012). Bsif: Binarized statistical image features. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, pages 1363–1366.

Kittipat's (2013). libsvm for matlab. Disponível em: `https://sites.google.com/site/kittipat/libsvm_matlab` Acessado: 09 de Maio de 2017.

Lima, V., Branco, K., and Colturato, A. (2014). *Reconhecimento De Padrões Em Imagens De Plantas De Eucalipto Obtida Por Um Veículo Aéreo Não Tripulado (VANT)*, volume 22. Simpósio Internacional de Iniciação Científica e Tecnológica da USP.

Morais, V. P. and Vieira, C. (2013). *MATLAB: curso completo*. FCA.

Nanni, L., Brahnam, S., and Lumini, A. (2012). Local phase quantization descriptor for improving shape retrieval/classification. 33:2254–2260.

Ng, R. (2016). Support vector machines - svms. Disponível em: http://www.ritchieng.com/machine-learning-svms-support-vector-machines/ Acessado: 15 de Novembro de 2017.

Rios, A. (2012). Libras - alfabeto e números. Disponível em: `http://www.ebah.com.br/content/ABAAAA9skAJ/libras-alfabeto-numeros` Acessado: 25 de Outubro de 2017.

Russell, S. and Norvig, P. (2010). *Artificial Intelligence: A Modern Approach*. Prentice Hall Series in Artifi. Prentice Hall.

Szeliski, R. (2010). *Computer Vision: Algorithms and Applications*. Texts in Computer Science. Springer London.

Triggs, B. and Dalal, N. (2005). *Histograms of Oriented Gradients for Human Detection.* Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05).