Classificação de gestos da Libras utilizando redes neurais

Igor Leonardo Oliveira Bastos[1], Michele Fúlvia Angelo[2]

[1]Mestrando – MMCC - Mestrado em Ciência da Computação (UFBA/UEFS)

[2]Orientadora. Departamento de Tecnologia – Universidade Estadual de Feira de Santana (UEFS)

igorcrexito@gmail.com, mfangelo@ecomp.uefs.br

Semestre de ingresso: 2012.2

Conclusão prevista: 2014.2

Qualificação realizada em 2013.1 (21/10/2013)

Abstract.*Gesture recognition has become in recent years, an area which has attracted a great attention of researchers because of the good results obtained and the use, in most cases,of little or no additional hardware but a camera. Thus, based on digital images obtained with these cameras, methods are used in order to track regions of interest on these images, allowing a subsequent classification based in some aspect. In this context fits the present work, which aims to develop a system to classify gestures based on the use of artificial neural networks and that uses, as main features, shape descriptors such as Histogram of Oriented Gradients (HOG) and Zernike moments. As a final step, it is intended to perform the classification according to a practical field such as for recognizing gestures of the Brazilian Sign Language (Libras).*

Palavras-chave: Reconhecimento de gestos; Redes neurais artificiais; Libras

1. Problema de pesquisa e caracterização da contribuição

Sistemas que realizam o reconhecimento de gestos, baseando-se em técnicas de visão computacional, estão cada vez mais presentes no nosso dia-a-dia. Estes gestos incluem ações e movimentos de mãos, dedos, face, braços, entre outros [Mitra e Acharya 2007].

Tomando como base os sistemas de reconhecimento de gestos, é notório que estes se utilizam de diversas abordagens distintas [Pavlovic, Sharma e Huang 1997], as quais variam desde técnicas baseadas em modelos estatísticos, como o proposto por Yang e Xu(1994), até técnicas que se utilizam exclusivamente de processamento digital de imagens, como o trabalho desenvolvido por Yikai, Wang, Cheng e Lu (2007).

Um campo para a aplicação de sistemas para o reconhecimento de gestos é o campo das línguas de sinais. Dentre estas, pode-se citar as línguas *American Sign Language* (ASL) e a Língua Brasileira de Sinais (Libras), sendo a última, oficialmente, adotada pelo governo como uma língua falada no Brasil [Felipe 2007].

No entanto, apesar da existência de trabalhos na área e que englobam a Libras, como o proposto por Carneiro et al (2010), nota-se que há uma carência no tocante à construção de sistemas ou abordagens que objetivam realizar a interpretação (tradução) de Libras para o português, sendo raros os que realizam tal tarefa.

Neste âmbito enquadra-se o presente trabalho, o qual tem como objetivo desenvolver uma abordagem para a tradução de sinais da Libras para o português a partir de imagens obtidas com uma câmera, utilizando para tal, os descritores de forma: Histograma de Gradientes Orientados (HOG) e Momentos Invariantes de Zernike.

2. Fundamentação Teórica e Trabalhos Relacionados

Trabalhos que utilizam técnicas de visão computacional e classificação para o reconhecimento de sinais têm se tornado mais comuns nos últimos anos. Dentre estes, pode-se destacar alguns que utilizam abordagens similares à do presente trabalho.

O trabalho realizado por Carneiro, Cortez e Costa (2010) apresenta uma abordagem para a classificação de sinais da Libras utilizando a classificação através de redes neurais. Contudo, as características utilizadas por Carneiro, para a classificação dos sinais foram os Momentos Invariantes de Hu. Este projeto classificou 26 gestos da Libras e apresentou taxas de eficácia que variaram de 78% a 97% [Carneiro et al 2010].

Já o projeto proposto por Rodríguez, Chávez e Menotti [Rodríguez et al 2012] utilizou do classificador *Support Vector Machine* (SVM) para a realização da classificação. Neste trabalho, foram comparados resultados utilizando-se os descritores de momentos propostos por Hu e por Zernike para o reconhecimento de sinais da Libras, sendo que os últimos obtiveram resultados superiores.

Além de trabalhos relacionados à Libras, existem outros que utilizaram abordagens semelhantes para o reconhecimento de sinais de outras línguas. O sistema proposto por Bowden [Bowden et al 2004] opera com o reconhecimento de sinais da *British Sign Language* (BSL). Este utiliza, além de aspectos referentes à forma dos sinais, características relacionadas ao movimento e posicionamento das mãos para a composição de um vetor de características, permitindo o reconhecimento de 49 sinais.

3. Estado Atual do Trabalho

O presente trabalho teve a sua metodologia e desenvolvimento divididos em 5 etapas, como mostrado na Figura 1. Destas, a etapa 1 encontra-se parcialmente concluída, as etapas 2 e 3 já foram concluídas e as demais estão em desenvolvimento. Atualmente, o desenvolvimento do trabalho encontra-se na etapa 4 (Classificação/Treinamento) e na criação dos *datasets* da etapa 1 (Aquisição de Imagens).

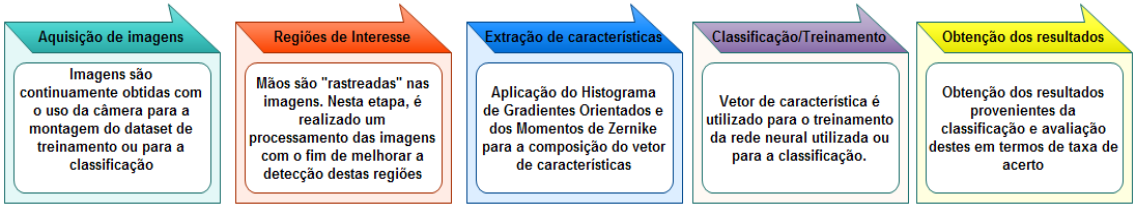

Figura1. Metodologia da abordagem proposta.

3.1. Etapas parcialmente concluídas

3.1.1 Aquisição de imagens

A etapa de aquisição de imagens representa a obtenção de imagens a partir de uma câmera. As rotinas necessárias para esta aquisição encontram-se concluídas e a câmera utilizada é do modelo *Microsoft Life Cam HD-3000*. As imagens adquiridas nesta etapa podem ser utilizadas para o treinamento e para o reconhecimento por parte do classificador que se pretende utilizar na etapa de classificação.

É importante ressaltar que para a realização do treinamento do classificador, é necessário que seja criado, baseando-se em imagens adquiridas nesta etapa, um *dataset* contendo imagens que correspondam aos sinais que se pretende reconhecer na etapa de classificação. Estes sinais ainda não foram definidos, porém, espera-se reconhecer cerca de 50 sinais distintos, utilizando 20 imagens para cada sinal. Estes sinais serão realizados por indivíduos distintos que apresentem diferentes tamanhos e configuração de mãos, além de estarem sujeitos à variadas condições de iluminação. Vale ressaltar que esta etapa, apesar de apresentar as suas rotinas já implementadas, ainda encontra-se em andamento, já que os *datasets* relacionados ao treinamento ainda não estão concluídos.

3.1.2 Regiões de interesse

A abordagem para o reconhecimento das regiões de interesse foi iniciada com a criação de um *dataset* contendo 243 imagens de pele. Destas imagens, 90 foram obtidas do *dataset Annotated Skin Database* [Annotated Skin Database 2013], 50 foram obtidas do *dataset Labelled Faces in the Wild* [Labelled Faces in the Wild 2013] e outras 103 foram adquiridas com o uso de uma câmera por parte dos próprios pesquisadores do projeto. Vale ressaltar que estas imagens foram escolhidas com condições variadas de iluminação, além de pessoas com diferentes tons de pele e com a presença de diferentes *backgrounds*.

As regiões de interesse para o presente projeto são as mãos do usuário. Apesar de aspectos como o posicionamento das mãos e o seu movimento corresponderem a parâmetros da Libras, os quais influenciam no significado dos sinais [Rodríguez et al 2012], apenas o parâmetro de forma (configuração) das mãos será utilizado.

Para o reconhecimento das regiões da imagem que correspondem às mãos do usuário, foram implementados algoritmos que realçam a pele. Estes algoritmos estipulam margens (limites) que definem se um dado pixel da imagem pertence (em termos de coloração) à pele ou não. Os algoritmos implementados foram os propostos por Peer [Peer et al 2003], Gomez [Gomez et al 2002] e Buhyiam [Buhyiam et al 2002]. Todos estes usam componentes de diferentes espaços de cores para a determinação dos limites que permitem a classificação dos pixels como pele.

Além dos algoritmos já citados, um classificador Perceptron Multicamada foi utilizado com o objetivo de identificar as regiões de pele. Para isso utilizou-se os componentes dos espaços de cores dos algoritmos de pele citados para a criação de um vetor de características. Para cada pixel de cada imagem do treinamento, um vetor de características foi gerado e estes usados para treinar o classificador. Este classificador teve o seu número de camadas escondidas e o número de neurônios nestas camadas determinado empiricamente, tomando como base, a realização de testes e a obtenção dos melhores resultados, os quais foram encontrados com o uso de 1 camada escondida e esta possuindo 35 ou 45 neurônios. A função de ativação usada foi a sigmóide. Com esta abordagem, obteve-se resultados, em termos de taxa de acerto, de 15% a 60% superiores aos algoritmos de pele mencionados. Vale lembrar que, assim como o treinamento, a classificação é realizada em cada pixel da imagem, sendo o resultado desta, a resposta se este pixel corresponde à pele. Na Figura 2, a arquitetura do classificador pode ser vista (a), além do vetor de características (b). Com esta classificação, pôde-se salientar as regiões de pele nas imagens obtidas com a câmera, fazendo-se necessário o rastreamento das regiões que correspondem às mãos.

Figura 2. a) Camadas do classificador utilizado. b) Vetor de características

3.1.3 Extração de características

Esta etapa corresponde à aplicação, sobre as imagens, das técnicas HOG [Cruz et al 2013] e Momentos de Zernike [Hse e Newton 2004]. O HOG é aplicado diretamente sobre a imagem original, usando a etapa anterior apenas para determinar sobre qual região da imagem ele deve atuar (coordenadas das mãos). Enquanto os momentos de Zernike são aplicados sobre a região das mãos na imagem binária resultante da classificação de pele.

3.2. Desenvolvimento Necessário para a conclusão

3.2.1 Classificação/Treinamento

A classificação/treinamento dos gestos corresponde à utilização do vetor de características criado na extração de características para treinar ou classificar as imagens. No caso do treinamento, pretende-se utilizar o *dataset* com imagens adquiridas na etapa de "Aquisição de Imagens". Apesar de ainda não definidos os sinais que serão reconhecidos, espera-se reconhecer 23 sinais do alfabeto de Libras (com exceção das letras não-estáticas H, J e Z), além de outros gestos estáticos que façam parte da Libras.

Neste projeto, pretende-se utilizar uma abordagem particionada para a classificação. Esta se dará com o agrupamento dos sinais que apresentam similaridades em termos das características extraídas. Como os descritores HOG e momentos de Zernike se baseiam na forma, é esperado que os sinais com configuração semelhante apresentem respostas mais parecidas do que aqueles que apresentam formas distintas.

Baseado nisso, será utilizada uma rede neural distinta para cada grupo de sinais criado, utilizando, primeiramente, uma rede genérica que tem como função conduzir a classificação para as redes menores e mais específicas. A opção por esta classificação segmentada foi feita a fim de reduzir as chances de problemas relacionados à utilização de uma única grande rede, tais como problemas de convergência e diminuição da eficácia da classificação realizada, os quais podem ser gerados quando se quer classificar grupos muito distintos. Além disso, para que se tenha uma rede capaz de reconhecer cada gesto, provavelmente, ter-se-ia que montar uma rede com muitas camadas escondidas e/ou muito neurônios, o que poderia ocasionar problemas de memorização (*overfitting*) [Lawrence, Giles e Tsoi 1996].

A arquitetura da etapa de classificação pode ser vista na Figura 3. A rede primária é treinada com todos os gestos e realiza uma classificação genérica, reconhecendo o grupo ao qual a entrada pertence. A classificação efetiva do sinal é feita nas redes secundárias, as quais recebem a mesma entrada da rede primária.

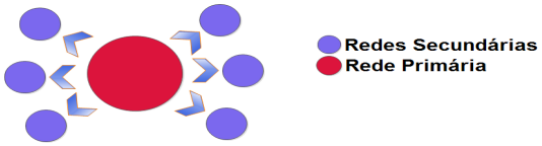

Figura 3. Arquitetura da etapa de classificação

3.2.2 Obtenção dos resultados

A etapa de obtenção dos resultados corresponde à avaliação da classificação dos gestos. Esta será feita variando-se parâmetros da classificação, como o número de neurônios e de camadas intermediárias, a função de ativação da rede, além de aspectos referentes à extração de características, como o número de *bins* usados e tamanhos de blocos e células do HOG. A representação dos resultados será feita através de curvas *Receiving Operating Characteristic* (ROC).

4. Avaliação dos Resultados

Até o presente momento alguns resultados já foram alcançados, como pode ser visto na Figura 4, que apresenta o realce dos pixels de pele, realizado na etapa de "Regiões de Interesse". Nota-se a pouca ocorrência de falsos positivos e negativos na detecção.

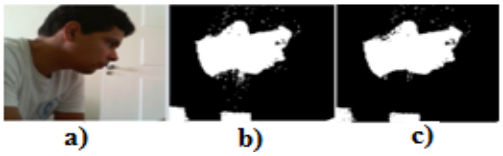

Figura 4. Classificação de pixels de pele usando rede neural. a) Imagem original. b) Classificação com 35 neurônios. c) Classificação com 45 neurônios

Com o realce das regiões de pele, utilizou-se, para o rastreamento das zonas de interesse, algoritmos de processamento de imagens, tais como operadores de mediana e fechamento para a remoção dos ruídos, além de um algoritmo para o reconhecimento de

regiões conexas. Este permitiu, como mostrado na Figura 5, a detecção das regiões dos braços e também da cabeça do usuário, sendo apenas necessária a definição de um *frame,* de tamanho fixo na extremidade das regiões conexas, para o rastreio das mãos.

Figura 5. Realce das regiões de interesse (braços).

Assim, espera-se, ao fim deste trabalho, que se obtenha resultados satisfatórios com a abordagem proposta. Ressalta-se que o atual projeto apresenta limitações quanto à dependência de fatores ambientais (iluminação e *background*), dependência da postura do usuário (deve-se vestir camisa e mostrar as regiões de braços, tórax e cabeça), sendo que se intenciona, ao fim do trabalho, avaliar os impactos destes fatores ambientais e da variação dos indivíduos envolvidos no reconhecimento de pele. Por fim, destaca-se a limitação quanto ao reconhecimento de apenas parte dos sinais que compõem a Libras.

Referências

Annotated Skin Database (2013), http://agami.die.uchile.cl/skindiff/, Dezembro.

Bowden, R., Windridge, D., Kadir, T., Zisserman, A. e Brady, M. (2004).“A Linguistic Feature Vector for the Visual Interpretation of Sign Language”. In: European Conference of Computer Vision, RepúblicaTcheca.

Buhyian, A., Ampornaramveth, V. e Ueno, S (2003). “Face detection and Facial Feature Localization for Human-machine interfaces”. In: *NII Journal*, n.5.

Carneiro, A., Cortez, P. e Costa, R. (2010). “Reconhecimento de gestos da Libras com classificadores neurais a partir dos momentos invariantes de Hu” .In: Interaction South America, Brasil.

Cruz, J.,Shiguemori, E. e Guimarães, L. (2013) “Comparação entre HOG+SVM e Haar-Like em cascata para a detecção de campos de futebol em imagens aéreas e orbitais”. In: XVI Simpósio Brasileiro de Sensoriamento Remoto.

Felipe, T (2007). A. Libras em Contexto. Walprint Gráfica e Editora, 8ª Edição. Brasília.

Gomez. G, Sanchez, M. e Sucar, L. (2002).“On selecting an appropriate colour space for skin detection”. In: Mexican International Conference on Artificial Intelligence, páginas 70-79, México.

Lawrence, S, Giles, C. e Tsoi, A. (1996). “What Size Neural Network Gives Optimal Generalization? Convergence Properties of Backpropagation”. In: Technical Report UMIACS-TR-96-22 e CS-TR-3617,Institute for Advanced Computer Studies, EstadosUnidos.

Labelled Faces in the Wild(2013), http://vis-www.cs.umass.edu/lfw/, Dezembro

Mitra, S. eAcharya, T. (2007).“Gesture Recognition: A Survey”. In: Systems, Man and Cybernetics, Part C: Applications and Reviews.

Pavilovic, V., Sharma, R. e Huang, T. (1997). “Visual Interpretation of Hand Gestures for Human-Computer Interaction: A Review”. In: IEEE Transactions on Pattern Analysis and Machine Intelligence.

Peer, P, Kovac, J. e Solina, F (2003).“Human skin colour clustering for face detection”. In: EUROCON - International Conference on Computer as a Tool, Eslovênia.

Rodríguez, K, Chávez, G. e Menotti, D (2012). “Hu and Zernike Moments for Sign Language Recognition”. In: International Conference on Image Processing, Computer Vision and Image Recognition, EstadosUnidos.

Yakai, F., Wang, K., Cheng, J. e Lu, H. (2007).“A real-time gesture recognition method”. In: Multimedia and Expo IEEE International Conference.

Yang, J. e Xu, Y. (1994).“Hidden Markov Model for Gesture Recognition”. Robotics Institute, Carnegie Mellon University.