Reconhecimento de ações humanas utilizando histogramas de gradiente e vetores de tensores localmente agregados

Luiz Maurílio da Silva Maciel[1], **Marcelo Bernardes Vieira**[1]

[1]Departamento de Ciência da Computação – Universidade Federal de Juiz de Fora

luiz.maurilio@ice.ufjf.br, marcelo.bernardes@ufjf.edu.br

Resumo. *Este trabalho apresenta um método para reconhecimento de ações humanas em vídeos utilizando histogramas de gradientes (HOG) e vetores de tensores localmente agregados (VLAT). Esse método consiste na geração de descritores utilizando HOG, gerar assinaturas VLAT a partir deles e então classificar utilizando um classificador Máquina Vetor Suporte (SVM). Para a realização dos testes utilizou-se a base de dados KTH. Foram obtidos resultados semelhantes aos encontrados na literatura, indicando que o método é promissor.*

Área de pesquisa: Computação Gráfica

Ano de ingresso: 2012

Palavras-chave: Reconhecimento de ações. Histograma de gradiente. Vetor de tensores localmente agregados.

1. Caracterização do Problema

Este trabalho trata do problema do reconhecimento e classificação de ações humanas em vídeos. Tal problema consiste em, dado um vídeo, determinar qual movimento está sendo realizado dentre um conjunto de movimentos possíveis. Esse tipo de reconhecimento tem diversas aplicações tais como em sistemas de segurança, indexação de vídeos, reconhecimento de gestos e entretenimento.

Para que se consiga determinar o movimento de uma sequência de imagens (vídeo) é necessário extrair características de cada imagem da sequência e representá-las em descritores. Os descritores devem ser capazes de extrair o máximo da informação de movimento de cada vídeo, devem ser semelhantes para um mesmo movimento e altamente discriminativos para movimentos distintos. Uma vez gerados, os descritores devem então ser classificados em ferramentas apropriadas.

O objetivo deste trabalho é a obtenção de descritores baseados em histogramas de gradientes [Dalal and Triggs 2005] que serão processados gerando vetores de tensores localmente agregados [Negrel et al. 2012] e então classificados. Esses descritores devem ser altamente discriminativos para que se obtenha altas taxas de reconhecimento.

2. Fundamentação Teórica

Esta seção trata das duas principais técnicas utilizadas neste trabalho: histogramas de gradientes e vetor de tensores localmente agregados.

2.1. Histogramas de Gradientes

Histogramas de gradientes (HOG - Histograms of Oriented Gradients) são histogramas gerados a partir dos gradientes de imagens. Proposto inicialmente por [Dalal and Triggs 2005] para a detecção humana em imagens foi posteriormente, estendida para o reconhecimento de ações em vídeos. [Kläser et al. 2008] propôs um descritor baseado em HOG em três dimensões (HOG3D) utilizando também a informação temporal do vídeo, além da informação espacial de cada quadro.

Neste trabalho os descritores serão calculados de modo semelhante a [Perez et al. 2012]. O gradiente do j-ésimo quadro de um vídeo em um ponto p é dado por:

$$\vec{g}_t = [dx\ dy\ dz] = \left[\frac{\partial I_j(p)}{\partial x}\ \frac{\partial I_j(p)}{\partial y}\ \frac{\partial I_j(p)}{\partial t}\right], \tag{1}$$

ou em coordenadas esféricas:

$$\vec{s}_t = [\rho_p\ \theta_p\ \psi_p], \tag{2}$$

com $\theta \in [0, \pi]$, $\psi \in [0, 2\pi)$ e $\rho = ||\vec{g}_t||$. Esse vetor indica a direção de maior variação de brilho que pode ser resultado de movimento local.

O gradiente dos n pontos de uma imagem I_j pode ser representado por um histograma tridimensional de gradientes $\vec{h}_j = \{h_{l,k}\}$, $k \in [1, b_\theta]$ e $l \in [1, b_\psi]$, onde b_θ e

b_ψ são o número de células para as coordenadas θ e ψ respectivamente. O histograma é calculado da seguinte forma:

$$h_{l,k} = \sum_p \rho_p, \tag{3}$$

onde $\{p \in I_j | k = 1 + \lfloor \frac{b_\theta \cdot \theta_p}{\pi} \rfloor, l = 1 + \lfloor \frac{b_\psi \cdot \psi_p}{2\pi} \rfloor\}$ são todos os pontos cujos ângulos são mapeados no intervalo da célula (k, l). O gradiente é então representado por um vetor de $b_\theta \cdot b_\psi$ elementos.

Para adicionar uma maior correlação espacial e aumentar a taxa de reconhecimento, cada quadro do vídeo é particionado em subjanelas e é calculado um histograma de gradientes para cada uma delas em separado. Assim, cada quadro é dividido em $n_x \times n_y$ partições não sobrepostas e para cada partição é calculado um histograma $\vec{h}_j^{a,b}$, $a \in [1, n_x]$ e $b \in [1, n_y]$. Pode-se ainda fazer uma reflexão horizontal do quadro a fim de reforçar simetrias horizontais do gradiente.

2.2. Vetor de Tensores Localmente Agregados

Vetor de tensores localmente agregados (VLAT - Vector of Locally Aggregated Tensors) é uma assinatura compacta para busca de similaridade. Foi proposto inicialmente por [Picard and Gosselin 2011] e recentemente melhorado por [Negrel et al. 2012]. O método propõe agregar produtos tensoriais de descritores locais para produzir uma assinatura única.

O VLAT proposto inicialmente dá bons resultados na busca de similaridade, porém os vetores são muito grandes. A fim de reduzir o tamanho do VLAT e manter seu poder discriminativo foi proposto o VLAT compacto [Negrel et al. 2012].

O VLAT compacto consiste em preprocessar o VLAT com um passo de normalização. Em seguida, computa-se a matriz de Gram do VLAT normalizado para um conjunto de treinamento. Encontra-se, então, uma aproximação da matriz de Gram calculada para o conjunto de treinamento utilizando os maiores autovalores da matriz. Por fim, computa-se a projeção dos vetores associados com o subespaço gerado. Essas projeções são o VLAT compacto.

O VLAT compacto mostrou-se eficiente na classificação de imagens. [Negrel et al. 2012] utilizou VLAT combinado com descritores SIFT para a classificação da base de imagens Holidays. Obtiveram resultados muito bons utilizando assinaturas várias ordens de magnitude mais compactas que métodos semelhantes.

3. Caracterização da Contribuição

Este trabalho busca unir os dois conceitos apresentados anteriormente, HOG e VLAT, de modo a se extrair informação de vídeos e obter boas taxas de reconhecimento de movimento.

A proposta é gerar descritores utilizando HOG para que tais descritores possam ser agregados utilizando VLAT gerando vetores altamente discriminativos. Parâmetros como número de subdivisisões de cada quadro, número de células do histograma, normalização

dos histogramas, utilização de reflexão horizontal são considerados para a geração dos HOG.

Os HOG gerados são então processados para a criação do VLAT e classificado por alguma ferramenta apropriada como máquina vetor suporte (SVM - support vector machine). Espera-se melhorar as taxas de reconhecimento atuais através desse processo.

4. Estado Atual do Trabalho

Alguns testes tem sido realizados utilizando-se a base de vídeos KTH [Schüldt et al. 2004]. Foram gerados descritores utilizando HOG para cada vídeo da base. Utilizou-se uma divisão dos quadros em 8×8 subjanelas, 16 células para o ângulo ψ e 8 para o ângulo θ. Esses valores foram escolhidos porque tiveram o melhor resultado em [Perez et al. 2012].

Inicialmente realizou-se duas normalizações no HOG. Uma normalização utilizando potência 0.72 e uma normalização L^2. O VLAT gerado a partir desses descritores obteve uma taxa de reconhecimento de 87.7%.

A seguir foi incluída a informação da reflexão horizontal de cada quadro e retiradas as normalizações uma vez que o VLAT também realiza normalizações ao processar os descritores. Esses novos descritores obtiveram uma taxa de reconhecimento de 89.9%.

A proposta de continuação do trabalho é realizar novas manipulações no HOG para que o VLAT gerado a partir deles possa obter maiores taxas de reconhecimento. Outra proposta é testar em uma base de dados mais difícil como a Hollywood2 [Marszałek et al. 2009].

5. Comparação com Trabalhos Relacionados

Comparando os resultados obtidos pelo método apresentado neste trabalho, HOG combinado com VLAT, com os resultados obtidos por outras técnicas utilizando HOG, encontrou-se resultados bastante próximos conforme mostra a Tabela 1. A expectativa é que com alguns ajustes na geração dos descritores com HOG os resultados possam ser melhorados.

Tabela 1. Comparação com outras técnicas utilizando HOG

Técnica	Taxa de reconhecimento
Harris3D + HOG3D [Kläser et al. 2008]	91.4%
Harris3D + HOG/HOF [Laptev et al. 2008]	91.8%
HOG3D + Tensor [Perez et al. 2012]	92.01%
HOG3D + VLAT	**89.9%**

Atualmente, o estado da arte para a base KTH foi obtido por [Gilbert et al. 2011] utilizando *Data Mining*. Esse trabalho obteve 95.7%.

Referências

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In Schmid, C., Soatto, S., and Tomasi, C., editors, *International Conference on Computer Vision & Pattern Recognition*, volume 2, pages 886–893, INRIA Rhône-Alpes, ZIRST-655, av. de l'Europe, Montbonnot-38334.

Gilbert, A., Illingworth, J., and Bowden, R. (2011). Action recognition using mined hierarchical compound features. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(5):883–897.

Kläser, A., Marszałek, M., and Schmid, C. (2008). A spatio-temporal descriptor based on 3d-gradients. In *British Machine Vision Conference*, pages 995–1004.

Laptev, I., Marszałek, M., Schmid, C., and Rozenfeld, B. (2008). Learning realistic human actions from movies. In *Conference on Computer Vision & Pattern Recognition*.

Marszałek, M., Laptev, I., and Schmid, C. (2009). Actions in context. In *Conference on Computer Vision & Pattern Recognition*.

Negrel, R., Picard, D., and Gosselin, P.-H. (2012). Compact Tensor Based Image Representation for Similarity Search. In *IEEE International Conference on Image Processing*, pages –, Orlando, États-Unis.

Perez, E. A., Mota, V. F., Maciel, L. M., Sad, D., and Vieira, M. B. (2012). Combining gradient histograms using orientation tensors for human action recognition. In *International Conference on Pattern Recognition*.

Picard, D. and Gosselin, P. H. (2011). Improving image similarity with vectors of locally aggregated tensors. In Macq, B. and Schelkens, P., editors, *ICIP*, pages 669–672. IEEE.

Schüldt, C., Laptev, I., and Caputo, B. (2004). Recognizing human actions: A local svm approach. In *In Proc. ICPR*, pages 32–36.