Descritor global baseado em banco de filtros para reconhecimento de ações humanas

Dhiego O. Sad[1], **Marcelo Bernardes Vieira**

[1]Instituto de Ciências Exatas – Universidade Federal de Juiz de Fora
Programa de Pós Graduação em Ciência da Computação (PGCC)
(UFJF) Caixa Postal 36.036-900 - Juiz de Fora - MG - Brasil
Área: Computação Gráfica
Ano de Ingresso: 2011
Previsão de Conclusão: 2012/2013

dhiegosad@ice.com.br,bernardes@ice.com.br

Resumo. *Movimento é uma das características fundamentais que refletem a informação semântica em vídeos. Uma das técnicas de estimativa do movimento é através de histogramas de gradientes orientados. O mesmo é formado à partir da extração das características encontradas resultantes da aplicação de um banco de filtros em cada imagem do vídeo. É proposto o uso de operadores derivativos para extração de atributos locais de cada pixel. Estes operadores representam a máxima variação da intensidade de briho em um ponto da imagem. O descritor criado é avaliado classificando-se a base de vídeos KTH com um classificador SVM (máquina de vetor de suporte). Resultados experimentais apresentam altas taxas de reconhecimento se comparados aos descritores globais referenciados na literatura.*

Palavras-chave*: Descritor de movimento. Tensor de orientação. SVM. Histograma de gradiantes. Banco de filtros.*

1. Introdução

No final da década de 1970 surgiram as primeiras pesquisas voltadas para a área da visão computacional, sendo definida como um conjunto de métodos e técnicas através dos quais sistemas artificiais são capazes de obterem informações de imagens ou quaisquer dados multi-dimensionais. Um sistema de visão completo pode ser dividido da seguinte forma [Marr et al. 2010]:

- Aquisição de Imagem: consiste em obter uma sequência de imagens digitais através de sensores geralmente contidos em câmeras digitais, como por exemplo, webcam. Dependendo to tipo de sensor o resultado da captação pode variar entre uma imagem bidimensional ou em uma sequência de imagens. O pixels indicam em cada coordenada valores de intensidade de luz em uma cor.
- Pré-processamento: consiste em aplicar métodos de processamento de imagem, por exemplo, filtros de suavização, para reduzir os ruídos gerados pela aquisição da imagem antes de extrair informações.
- Extração de características: consiste em capturar informações de uma imagem. Uma imagem é formada por modelos matemáticos, como por exemplo matrizes, estas contêm características que podem matematicamente ser identificadas como: textura, bordas e etc.
- Detecção e segmentação: consiste em destacar uma determinada região de uma imagem e segmentá-la, com a finalidade de guardar essa informação para processamento posterior.
- Pós-processamento: consiste na verificação dos dados, a estimativa de parâmetros sobre a imagem e a classificação dos objetos detectados em diferentes categorias.

O foco de estudo deste trabalho, que se insere na área de visão computacional, está no reconhecimento de movimentos em vídeos. Movimento é a principal característica que representa a informação semântica em vídeos. Detectar um objeto ou uma pessoa e rastreá-lo é de grande interesse em diversas aplicações de segurança, como por exemplo rastreamento de mísseis e detecção de movimento em sistemas de vigilância.

Este trabalho utiliza banco de filtros para extrair informações de movimento dos vídeos (Extração de características). É proposto o uso de operadores derivativos para extração de atributos locais de cada pixel. Estes operadores representam a máxima variação da intensidade de briho em um ponto da imagem.

Um dos principais problemas deste trabalho está na redução de dimensionalidade, ou seja, conseguir de forma condensada representar toda informação de movimento extraída dos vídeos. Uma das formas de representar essa informação é utilizando histogramas de gradientes orientados [Zelnik-manor and Irani 2001] e tensores de orientação [Mota 2011] (Pós-processamento). Com isso, é possível armazenar essas informações em descritores, que é um par - vetor de características extraídas e função de distância - usado para indexação por similaridade de vídeos e/ou imagens, para que seja possível realizar uma comparação entre os vídeos. Os vídeos utilizados neste trabalho são oriundos das bases KTH e Hollywood2 [Schuldt et al. 2004] (Aquisição de imagem).

1.1. Definição do problema

Dada uma sequência de imagens $S(x, y, t)$,

$$S : [U \ \subset \Re^2] \times \Re \rightarrow \Re^n,$$

onde U é um conjunto suporte, $t \in \Re$ é o determinado tempo em que a imagem foi obtida e $\Re^n$ é o espaço de cores associado a cada imagem, o problema deste trabalho é propor uma função $w(g(S)) \rightarrow \Re^k$ que representa o movimento como um vetor k-dimensional, onde g é o resultado, no domínio do espaço, da aplicação de um banco de filtros em S capaz de salientar o movimento aparente.

1.2. Objetivos

O objetivo primário deste trabalho é fornecer um banco de filtros no domínio da frequência que vai culminar numa distribuição espectral que carrega a informação de movimento contida no vídeo.

Como objetivo secundário, deve-se obter um descritor que represente de forma compacta toda informação capturada após a ánalise de cada frame ou do vídeo como um todo.

2. Fundamentação teórica

Neste capítulo são apresentados os fundamentos para que seja possível extrair a informação de movimento do vídeo e armazená-la em descritores.

2.1. Banco de filtros

Um banco de filtros é uma combinação de filtros passa-faixa que decompõe o sinal em diversas componentes, cada uma contendo apenas uma sub-banda de freqüência do sinal original. É desejável que o projeto do banco de filtros seja tal que permita a recombinação das sub-bandas para se recuperar o sinal original. O primeiro processo (separação das sub-bandas) é chamado de análise, e o segundo de síntese. A saída da análise é chamada de sinal de sub-bandas, com tantas sub-bandas quantos forem os filtros.

O banco de filtros serve para isolar as diferentes componentes de freqüência de um sinal. Isto é importante pois, para muitas aplicações determinadas freqüências são mais importantes do que outras (Fig 1). Por exemplo, as frequências mais importantes podem ser codificadas com uma resolução maior. Em geral, pequenas diferenças no sinal dessas freqüências são muito significativas e um esquema de codificação que preserva essas diferenças precisa ser usado. Por outro lado, as frequências menos importantes não precisam de reconstrução exata e uma codificação mais esparsa poderia ser usada, mesmo que se percam detalhes na codificação [Boashash 2003].

2.2. 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

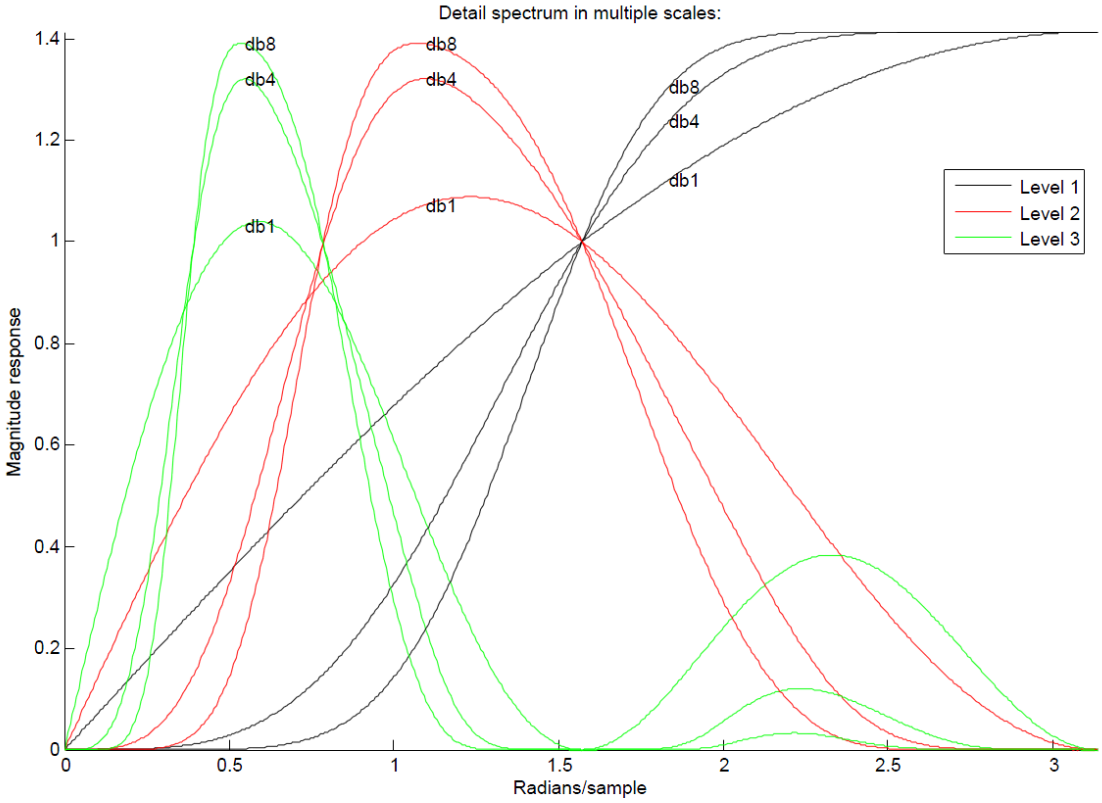

Figura 1. Resposta dos filtros Daubechies 1,4,8.

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.

2.3. Tensor de orientação

Tensores estendem o conceito de vetores e matrizes para ordens maiores. Na terminologia tensorial, vetores são tensores de primeira ordem e matrizes são tensores de segunda ordem. Um tensor pode ser definido matematicamente como [Westin 1994]:

$$\mathbf{T} = \sum_{i=1}^{n} \lambda_i e_i {e_i}^T, \tag{4}$$

onde λ_i são os autovalores e e_i os respectivos autovetores.

2.4. Descritor global de movimento

Descritor global de movimento é um par - vetor de características extraídas e função de distância - usado para indexação por similaridade de vídeos e/ou imagens. O vetor de características contém as propriedades da imagem ou do vídeo e a função de distancia mede a similaridade entre duas imagens ou dois vídeos.

Na maioria das vezes, a similaridade é definida como inversa à função de distância (por exemplo, distância Euclidiana), assim, quanto menor a distância entre as imagens ou vídeos, maior é a similaridade entre eles.

2.5. Máquina vetor suporte

Uma máquina vetor suporte (SVM) é uma técnica de aprendizado supervisionado que utiliza algoritmos de aprendizado para analisar dados e reconhecer padrões.

Basicamente, o SVM pega um conjunto de dados de entrada e prevê a qual de duas possíveis classes cada um deles pertence. Isto é feito através da criação de um hiperplano separador para dados linearmente separáveis (Fig 2).

A partir de um conjunto de treino, onde um dado é marcado como pertencente a uma de duas categorias distintas, a etapa de aprendizado do SVM constrói um modelo que associa cada dado a uma ou outra categoria.[Vapnik 1995].

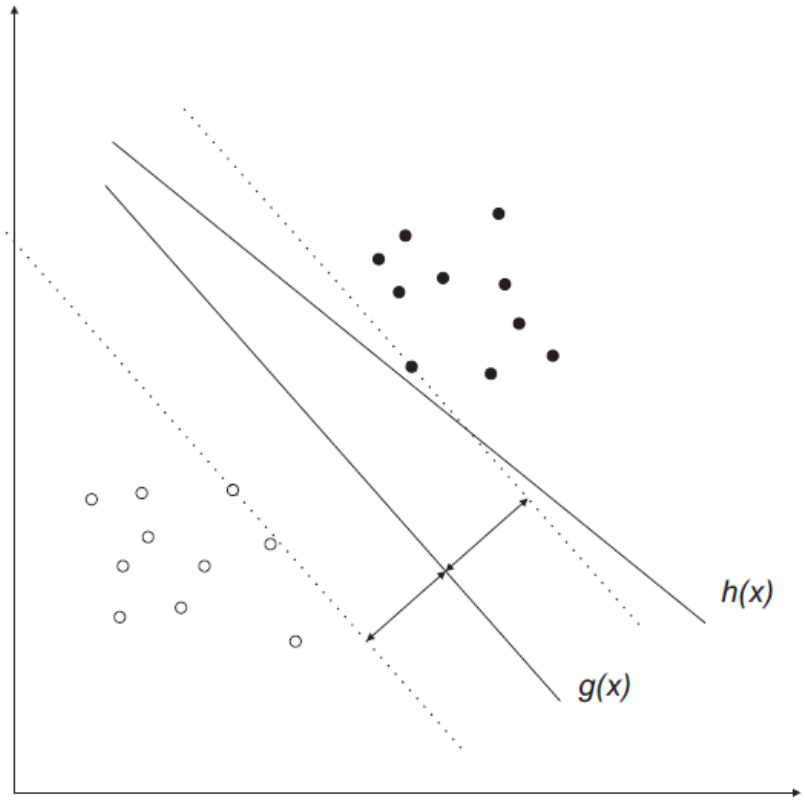

Figura 2. Exemplo de duas classes separáveis linearmente e os hiperplanos g(x) e h(x) que as separam.

3. Trabalhos relacionados e caracterização da contribuição

Este trabalho é continuação de uma dissertação de mestrado e de um artigo publicado [Perez et al. 2012], cujo objetivo é incrementar o trabalho anterior, visando um resultado melhor no que diz respeito à precisão no reconhecimento de ações em vídeos.

Em [Mota 2011] propõe-se um descritor global de movimento baseado em um tensor de orientação. Este tensor, assim como em [Kihl et al. 2010], também é extraído da projeção do fluxo óptico em uma base ortogonal de polinômios.

No trabalho de [Perez et al. 2012] é realizada uma combinação entre tensores de segunda ordem e histogramas de gradientes na geração dos descritores utilizando informação de todo quadro, sendo mais simples e menos custoso computacionalmente.

A principal contribuição deste trabalho está na utilização de uma outra técnica para calcular os histogramas de gradientes. O histograma agora é calculado através do uso de um banco de filtros que são operadores derivativos que fazem a extração de atributos locais de cada pixel. Estes operadores representam a máxima variação da intensidade de briho em um ponto da imagem.

A idéia de utilizar banco de filtros para extrair as informações de movimento em cada vídeo é que, com a possibilidade de usar uma quantidade maior de filtros para extrair características de movimento em cada vídeo, nos permite fazer uma análise melhor do espectro de cada imagem do vídeo, podendo assim optar pelo filtro que nos apresente uma

melhor resposta no que se refere à separação das frequências que representam movimento.

4. Estado atual do trabalho

Atualmente está sendo estudado algumas possibilidades de combinar os diferentes filtros usados para extrair informações de cada vídeo. Isto é feito para que seja possível conseguir um resultado melhor que os encontrados na literatura. A tabela 1 mostra os melhores resultados conseguidos até então:

Método	Taxa de reconhecimento
DAUB 1	83.20%
DAUB 1+2+3+4	84.01%
DAUB 1+2+3+4+5	85.17%
DAUB 1+2+3+4+5+6+7	86.79%
DAUB 1+2+3+4+5+6+7+8	**87,25%**

Tabela 1. Taxa de reconhecimento para base KTH.

Os melhores resultados foram conseguidos através da concatenação dos descritores, aumentando a quantidade de informação e consequentemente a dimensão do tensor. Esses descritores foram obtidos através da extração de informação de movimento em cada vídeo usando diferentes tipos de filtros.

Os filtros utilizados foram de uma família de filtros wavelets conhecidos como Daubechies.

Além da concatenação, outros métodos foram utilizados para combinar esses diferentes descritores, como por exemplo calcular o autovalor de cada descritor e usar aquele que apresentar o maior valor. Porém, tanto esse teste como os outros realizados até agora não apresentaram bons resultados.

Atualmente alguns métodos chegaram em resultados expressíveis como mostra a tabela 2:

Método	Taxa de reconhecimento
HOG pyramids [Laptev et al. 2007]	72%
Harris3D + HOG3D [Kläser et al. 2008]	91.4%
Harris3D + HOG/HOF [Laptev et al. 2008]	91.8%
HOG3D + Tensor [Perez et al. 2012]	**92.12%**
ISA [Le et al. 2011]	93.9%
TCCA [Kim et al. 2007]	95.33%

Tabela 2. Taxa de reconhecimento para base KTH.

Este trabalho apresenta um ótimo potencial para alcançar e ultrapassar o resultado obtido pelo trabalho de [Perez et al. 2012], possibilitando assim uma publicação numa renomada conferência internacional.

Para validar nosso resultado, usamos as bases de dados KTH [Schuldt et al. 2004] e futuramente usaremos a base Hollywood2 [Marszałek et al. 2009]. Na base KTH, usamos um classificador multiclasse adotando a estratégia utilizada por [Wang et al. 2009].

Para o conjunto de dados, cada classe é modelada utilizando um classificador SVM com uma função de kernel linear, com distância euclidiana.

Referências

Boashash, B. (2003). *Time Frequency Analysis: A Comprehensive Reference*. Elsevier Science Limited.

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.

Kihl, O., Tremblais, B., Augereau, B., and Khoudeir, M. (2010). Human activities discrimination with motion approximation in polynomial bases. In *IEEE International Conference on Image Processing*, pages 2469–2472, Hong-Kong.

Kim, T., Wong, S., and Cipolla, R. (2007). R.: Tensor canonical correlation analysis for action classification. In *In: CVPR 2007*.

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., Caputo, B., Schuldt, C., and Lindeberg, T. (2007). Local velocity-adapted motion events for spatio-temporal recognition. *Comput. Vis. Image Underst.*, 108:207–229.

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

Le, Q. V., Zou, W. Y., Yeung, S. Y., and Ng, A. Y. (2011). Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In *CVPR*, pages 3361–3368.

Marr, D., Poggio, T., and Ullman, S. (2010). *Vision: A Computational Investigation Into the Human Representation and Processing of Visual Information*. Mit Press.

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

Mota, V. F. (2011). Tensor baseado em fluxo óptico para descrição global de movimento em vídeos. *Periódico Desconhecido*.

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*.

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

Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. Springer.

Wang, H., Ullah, M. M., Kläser, A., Laptev, I., and Schmid, C. (2009). Evaluation of local spatio-temporal features for action recognition. In *BMVC*.

Westin, C.-F. (1994). *A Tensor Framework for Multidimensional Signal Processing*. PhD thesis, Linköping University, Sweden. N. 348.

Zelnik-manor, L. and Irani, M. (2001). Event-based analysis of video. In *In Proc. CVPR*, pages 123–130.