MOSIFT PARA O RECONHECIMENTO DE AÇÕES HUMANAS.

Marlon Ramos Avalos

Instituto de Ciências Exatas e Biológicas
Departamento de Computação
Universidade Federal de Ouro Preto (UFOP) – Ouro Preto, MG – Brasil leyenda887@gmail.com

ABSTRACT

In human action detection in videos, systems use different types of spatio-temporal descriptors to characterize the activity that define an action. This paper will show the MoSIFT descriptor, based on SIFT and Optical Flow methods, as an efficient alternative for feature extraction in videos that represent the action performed by a person.

Index Terms— MoSIFT, Bag-of-Word, Human Action

1. INTRODUÇÃO

O reconhecimento automático de diferentes comportamentos humanos em vídeos, é um dos objetivos que tem a visão computacional. Existem diversas aplicações,entre elas temos os sistemas de vigilâcia, onde as câmeras geram um grande volume de dados as 24 horas. A análise desses dados se torna em um grande desafio quando realizada manualmente, já que a maior parte dos videos não apresentam eventos interessantes.

Diversos descritores têm sido usados na literatura, entre os mais populares temos os descritores locais, já que eles possuem características que os tornam invariantes á transformações geométricas, tais como a rotação, escala, translação, etc. [1].

Um método de detecção de características utilizado atualmente é Scale Invariant Feature Transform (SIFT), que detecta pontos de interesse em um imagem estática [2]. No entanto, com objetivo de aumentar a robustez de um ponto de interesse, é utilizado o histograma de fluxo ptico (HOF, do inglês Histogram of Optical Flow) para acrescentar informação de movimento aos pontos detectados. É desta forma que o descritor MoSIFT criado, através da adição de informação de movimento a cada um dos descritores dos pontos de interesse detectados pelo SIFT [1].

Neste artigo, é apresentado o procedimento de extração de características através do descritor MoSIFT para obter os pontos de interesse espaço-tempo de um vídeo, utilizando os dois processos anteriormente mencionados. Para testar o descritor será usada a base de ações humana KTH, logo será aplicado a técnica Bag-of-Words como modelo de aprendizagem para uma posterior classificação da ação detectada.

2. MOSIFT

O algoritmo MoSIFT é usado para detectar e descrever pontos de interesse espaço-temporais de um objeto ao longo do tempo em uma sequência de vídeo. A detecção destes pontos de interesse torna-se um conjunto de dados descritivos, então precisa-se de uma certa quantidade de pontos de interesse para poder detectar uma açõ humana. [1].

Aplica-se o algoritmo SIFT para encontrar pontos de interesse na imagem estática e logo para cada punto de interesse são adicionados informação de movimento através do HOF, como é mostrado na Figura 1:

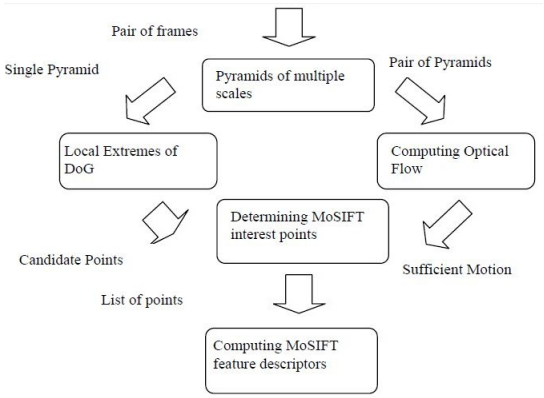

Fig. 1. Fluxograma de algoritmo MoSIFT

2.1. Detecção de Pontos de Interesse

O algoritmo utiliza um par de italico do vídeo para encontrar pontos de interesse espaço-temporal em múltiplas escalas. Os pontos de interesse so detectados por meio do SIFT. Essa detecção é realizada da seguinte forma: primeiro a imagem é escalada em diferentes tamanhos, sempre em potências de 2. Logo, para cada escala são geradas várias imagens suavizadas usando máscaras Gaussianas com parmetros diferentes, obtendo desta forma uma série de imagens suavizados com um diferente fator. O próximo passo consiste em calcular a diferena entre pares de imagens suavizadas. (Diferença de Gaussianas - DoG) (Figura 2).

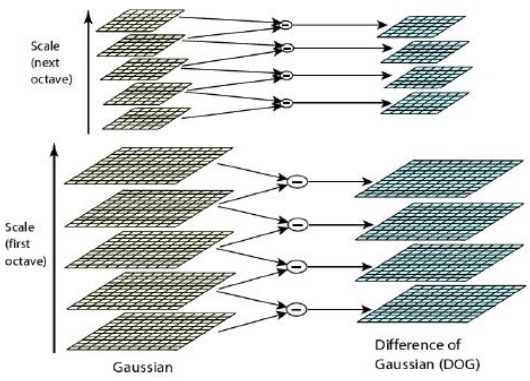

Fig. 2. Diferenças de Gauss em múltiplas escalas

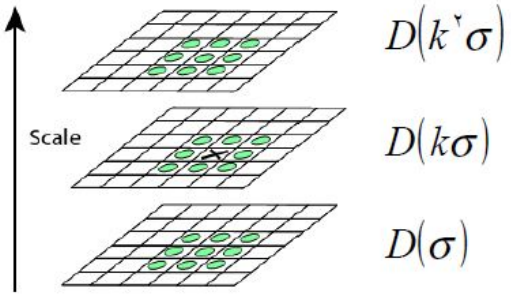

Fig. 3. A detecção de um ponto de interesse

Uma vez obtida a pirâmide de imagens através do DoG, para detectar os pontos de interesse candidatos so utilizados trłs imagens de DoGs. O ponto de interesse encontrado avaliando a imagem central, se o pixel avaliado é maior ou menor que seus oito vizinhos imediatos e seus nove vizinhos na camada superior e inferior (Figura 3), então temos a um candidato a ponto de interesse [5].

O O algoritmo SIFT esta desenhado para detectar pontos de interesse distintivos em imagens estáticas. Por exemplo, em imagens com fundo complexo, várias ponto de intesse são detectados, muitos deles não estão relacionados com a ação humana. Claramente, alguns pontos de interesse com o suficiente movimento proveram informação necessária para reconhecer a ao. É neste contexto que é usado o algoritmo HOF, ele detecta o movimento de uma região calculando para onde dita região se movimenta no espaço da imagem por meio de diferenas temporais. O HOF captura a magnitude e a direção do movimento.

As escalas múltiplas dos fluxos opticos são calculadas de acordo com as escalas do SIFT. Um extremo local a partir de pirâmides DoG só pode se tornar um ponto de interesse se existir suficiente movimento. Portanto, o algoritmo MoSIFT só considerará aqueles pontos de interesse que tenham suficiente movimento.

2.2. O descritor MoSIFT

Para encontrar os pontos de interesse, é usado histogramas de gradientes e histogramas de fluxo óptico para aumentar o desempenho.

Como a detecção de pontos MoSIFT é baseado em SIFT e fluxo optico, é natural que o descritor aproveite essas duas características. Em vez de combinar um classificador de histogramas de gradientes orientados (HOG, do inglês Histograms of Oriented Gradients) completo e com um classificador HoF completo, é construído um único descritor, que concatena os descritores HoG e HoF em um vetor [1].

Quando é detectado um ponto de interesse, é calculada a magnitude e direção para o gradiente em cada pixel de uma região ao redor do ponto de interesse na imagem. Um histograma de orientação é formado por 8 bins, com cada bin cobrindo um intervalo de 45 graus. Os elementos são agrupados em uma grade de 4×4 células ao redor do ponto de interesse, onde cada célula de 4×4 pixels ao redor do ponto de interesse (figura 4). Para cada célula é calculado um histograma de 8 bins, Assim SIFT gera um vetor com 128 dimensões (4x4x8 = 128), cada vetor é normalizado para tornar o descritor invariante á iluminação.

MoSIFT adapta a ideia da rede de agregação SIFT para descrever também o movimento, o fluxo óptico detecta a magnitude e a direção de um movimento. Por tanto, o fluxo óptico tem as mesmas características que os gradientes de aparência. A mesma agregação pode ser aplicado ao fluxo óptico na área dos pontos de interesse para aumentar a robustez. Em seguida, os dois vetores de histograma agregados (SIFT e Fluxo Optico) sâo concatenados para criar o descritor MoSIFT, que agora tem 256 dimenses.

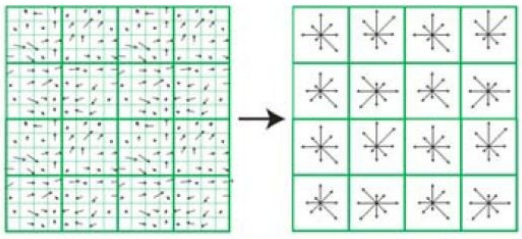

Fig. 4. Formação de histograma através da agregação células em regiões de 4x4 para SIFT e fluxo óptico dando assim 256 dimensões que fazem MoSIFT.

3. VISUAL BAG-OF-WORD

Tomando da recuperação de informação textual, o italico tem bons resultados quando aplicado no campo do processamento de imagem. A técnica BoW é uma representação de caracterísicas usualmente usada para representar um evento de movimento usando ponto de interesse espaç-temporais [3]. Um diccionario de palavras visuales é construído agrupando

pontos de interesse espaço-temporais. Cada ponto de interesse é atribuído é atribudo á palavra do vocabulario mais próxima e o histograma de palavras visuais é calculado sobre um volume espaço-temporal para descrever uma ação. Uma palavra visual é um conjunto de vetores de características que contêm finformações semelhantes.

Para encontrar as palavras visuais, é seguido o seguinte conjunto de passos [4]:

1. Dado um conjunto de amostra de dados, aplica-se um algoritmo de agrupamento.
2. A comparação dos vetores característicos com os grupos formado é realizado através de uma função de distância, formando assim um padrão único de dados com um centro denominado de palavra visual (visual codeword).
3. O conjunto de palavras visuais forma o dicionanário, que é usado para calcular os histogramas de palavras visuais.

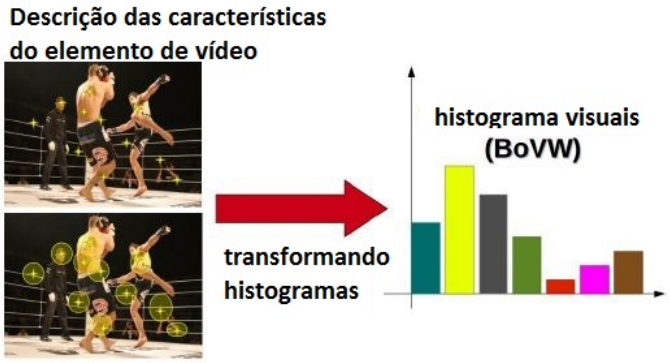

Fig. 5. Exemplo de Histogramas visuais para representar os elementos de vídeo de acordo a ocorrência de palavras visuais no espaço de característica

4. PROCEDIMENTO

O banco de dados KTH Human motion tem seis classes ou tipos de ações exclusivas de pessoas em diferentes cenários (Figura 7), com diferente iluminação e movimento da câmera, cada classe tem 100 vídeos diferentes (600 vídeos no total).

Na Figura 8 é mostrada um processo [6] semelhante ao procedimento neste artigo, nessa seção se detalha tal procedimento, tendo como entrada um vídeo para a obtenção de vetores do histograma de cada um de seus frames.

Para calcular o descritor MoSIFT de um vídeo, cada frame do vídeo é convertido a escala de cinza (Figura 6(a). Logo, o frame $f_t$deve ser segmentado, para isso é calculada a diferena entre o frame anterior f_{t-1} e o frame posterior f_{t+1}. Logo, é calculado o bouding box ao redor da área com maior número

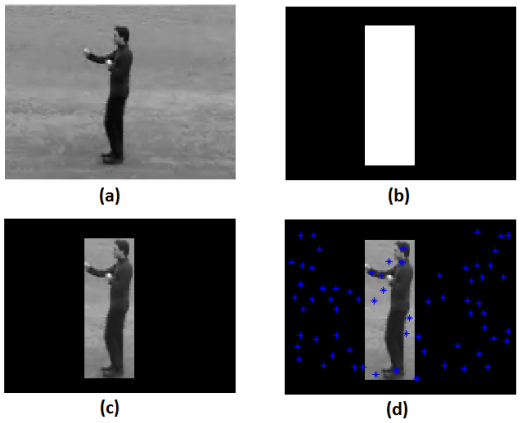

Fig. 6. (a) Frame de uma ação de uma pessoa. (b) A máscara (c) fixar a máscara sobre o frame (d) Os pontos de interesse válidos em torno da pessoa.

de pixels em movimento, como mostrado na Figura 6(b). Como o ator é o objeto em movimento, o bounding box sera gerado ao redor da pessoa; como mostrado na Figura 6(c). Quando detectados os pontos de interesse no frame f_t, devido ao ruido da imagem, vários pontos de interesse não relevantes são gerados, ver Figura 6(d). É através do bounding box que pontos de interesse causado pelo ruído são eleminados. O discritor MoSIFT é calculada na região segmentada, gerando um vetor de 256 elementos.

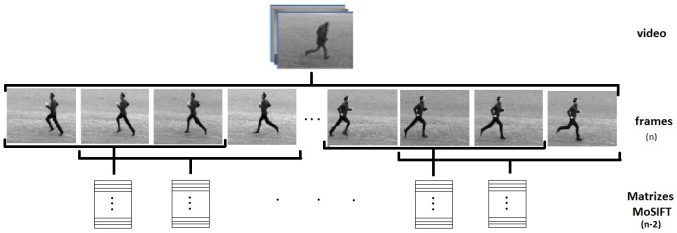

Fig. 9. Estrutura de obtenção das matrizes MoSIFT em um vídeo

Depois da extração de características é aplicada a técnica Bag-of-Words A partir dos vetores de características calculados pelo MoSIFT da base de vdeos, é extraída uma amostra para poder gerar o dicionário de palavras visuais, ver Figura (figura 10).

Logo, as amostras são agrupadas em K grupos através do algoritmo de clusterização K-means. Uma vez gerado o dicionário, é calculado o histograma de palavras visuais para cada frame. O histograma contabiliza o número de ocorrências de cada palavra.

Depois de calculados os histogramas de palavra visuais é realizada a etapa de classificação. Para isso, são concatenados os histogramas dos vídeos, gerando uma matriz de $n \times 256$, onde n é o nmero total de frames na base. Logo, a matriz é

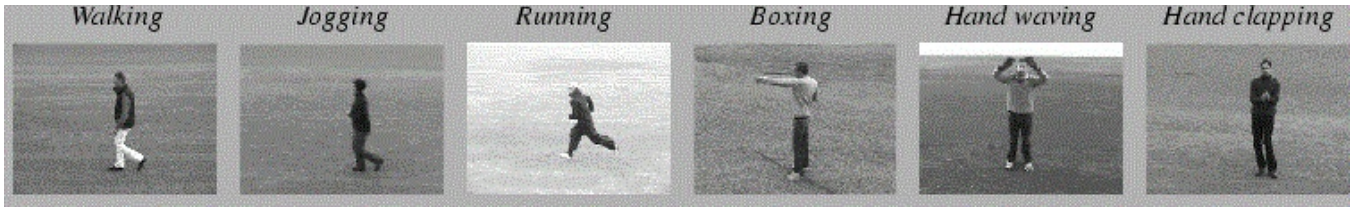

Fig. 7. Exemplos de Base de Dados KTH Human motion

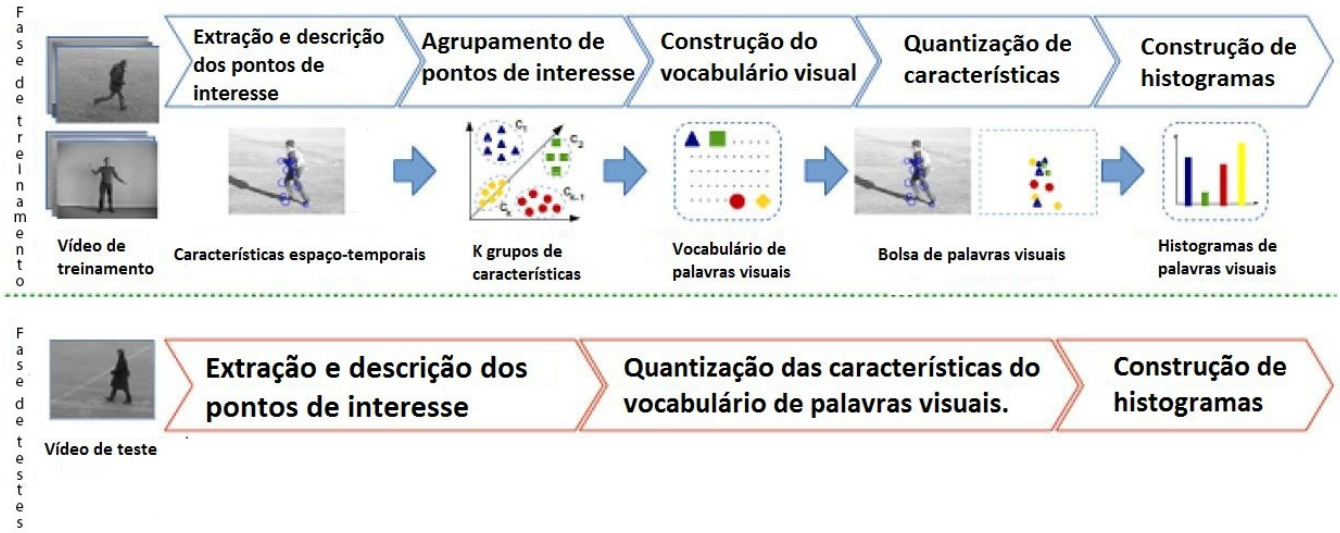

Fig. 8. Procedimento de extração e descrição usando MoSIFT e o Modelo Bag of Visual Word

normalizada e finalmente classificada por uma SVM.

5. RESULTADOS E ANÁLISES

Da base de vdeos KTH, foram utilizados 30 vídeos de cada classe (15 para treinamento e 15 para o teste), cada classe que compõe o banco de dados foi nomeado por rótulos como se segue:

- boxing = 1
- handclapping = 2
- handwaving = 3
- jogging = 4
- running = 5
- walking = 6

Para a extracção da amostra, foi utilizado 15% dos pontos de interesse das matrizes obtidas a partir do descritor em cada vídeo. Para o agrupamento de características foram utilizados valores K =100, 200, 400, 500, 600 y 1000 palavras visuais. A classificação foi realizada usando o classificador SVM e

Dicionrio	SIFT	Optical Flow	MoSIFT
100	52.22%	41.11%	55.56%
200	61.11%	42.23%	57.78%
400	64.45%	40%	61.11%
500	64.44%	40%	67.77%
600	68.89%	36.67%	67.78%
1000	73.33%	36.67%	73.33%

Table 1. Desempenho de detecção de Ações Humana com os descritores: SIFT, Optical Flow é MoSIFT.

foi conduzida da seguiente forma. A etapa de treinamento foi realizada através de um aprendizado por frame, i.e., os histogramas de cada frame e suas respectivas etiquetas são usadas para alimentar o classificador, para finalmente ter um modelo de aprendizagem. Para a etapa de teste, cada frame é classificado de forma independente, mas a etiqueta final do vídeo é gerada por votação, i.e., se a maioria dos frames de um vd́ieos indicam que ação X está sendo executada, então X é a etiqueta do vídeo. A Tabela 1 mostra os resultados da análise feita com os descritores SIFT, fluxo óptico e MoSIFT. Como pode ser visto o fluxo óptico tem um resultado inferior comparados com SIFT e MoSIFT,Como os resultados com SIFT

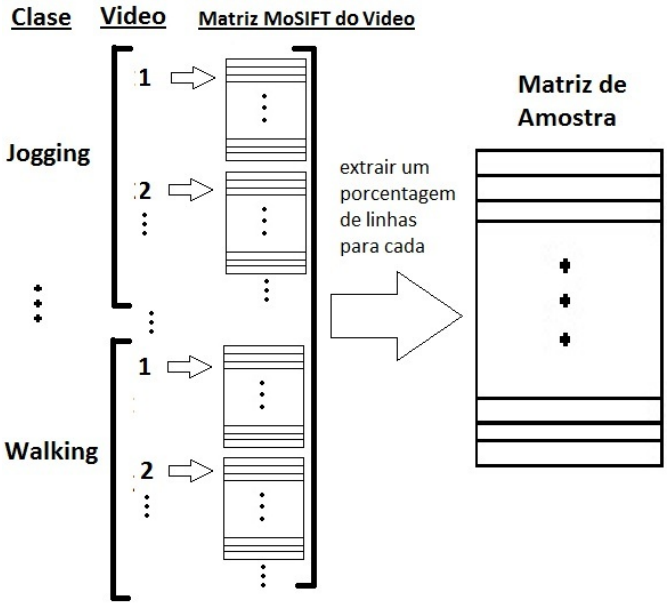

Fig. 10. Gerando matriz de amostras

esparso foram muito baixos, propomos neste trabolho usar um SIFT denso. Como pode ser observado na Tabela 1,assim que aumenta o número de palavras visuais, as acurácias dos descritores SIFT e MoSIFT também aumentam. A acurácia do HOF termina caindo quando o número de palavras visuais aumenta. Um motivo da baixa taxa de acurcia é devido a que existem vários pontos interese gerados por causo da base ruidosa, i.e., vários pontos de interesse fazem parte do fundo da imagem da parte segmentada.

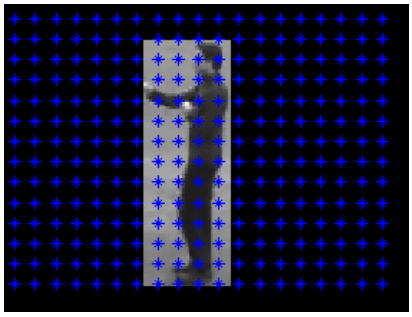

Fig. 11. Exemplo de SIFT mais denso

A melhor acurácia foi conseguida usando o MoSIFT, mas os resultados são baixos devido a que poucos pontos de interesse são gerados nas bordas do corpo da pessoa que executa o movimento, sendo muitas vezes maior o numero de pontos de interesse do fundo. Apesar do uso do SIFT denso o resultado é ainda baixa. Acreditamos que o problema reside no baixo número de pontos de interesse dentro do corpo, segundo a ação, a maioria de pontos está localizado no funda da imagem segmentada.

	Boxing	Clapping	Waving	Jogging	Running	Walking
Boxing	15	0	0	0	0	0
Clapping	0	6	9	0	0	0
Waving	0	1	12	0	2	0
Jogging	0	0	0	6	4	5
Running	0	0	0	2	12	1
Walking	0	0	0	0	0	15

Fig. 12. Matriz de confusão MoSIFT com 1000 palabras para dados KTH. As classes com mais problemas sao Clapping e Jogging que terminam sendo confundidas com Waving e Walking, respectivamente.

Na matriz de confusão mostrada na figura 12 é o resultado de MoSIFT com um diccionário de 1000 palavras, percebemos que a classe handclapping é confundida com handwaving e a classe jogging com a walking.

6. CONCLUSÕES

Nós mostramos que o algoritmo MoSIFT é eficiente para detectar pontos de interesse espácio-temporal de um vídeo, que pode usar no campo de detecção e reconhecimento.

O descritor MoSIFT como mostrado nos resultados, acreditamos que terá uma melhor acurácia quando o processo de segmentação melhore, reduzindo desta forma o número de pontos de interesse não descriminantes.

Como mostrado na Tabela 1, quando aumenta o tamanho do dicionário também aumenta a acuracia do modelo usado neste trabalho. A maior quantidade de palavras, o acoracidad é maior, porque há um agrupamento melhor por semenjanza das caracterśticas que melhor difiere dos outros grupos formados.

7. REFERENCES

[1] M.-Y. Chen and A. Hauptmann, Mosift: Recognizing human actions in surveillance videos CMU-CS-09-161. Carnegie Mellon University,Pittsburgh PA 15213,September 2009.

[2] D.G. Lowe. Distinctive image features from Scale-Invariant Keypoints, In IJCV, International Journal of Computer Vision, January 2004.

[3] T. Deselaers, L. Pimenidis, and H. Ney, Bag-of-visual-words models for adult image classification and filtering in ICPR, Deutsche Forschungsgemeinschaft (DFG)-NE-572/6, 2008, pp. 1-4.

[4] F. D. M. de Souza, G. Ca. Chávez, E. A. do Valle, and A. de A Araujo, Violence detection in video using spatio-temporal features, in Proceedings of the 23rd SIBGRAPI Conference on Graphics, Patterns and Images. IEEE, 2010, pp. 224-230.

[5] D. G. Lowe, Distinctive image features from scaleinvariant keypoints, International Journal of Computer Vision, vol. 60, no. 2, pp. 91-110, 2004.

[6] Ruben Hernndez Garca, Edel Garca Reyes, Julan Ramos Czar, Nicols Guil Mata, Modelos de representacin de caractersticas para la clasificacin de aciones humanas en video: Estado del arte, Revista Cubana de Ciencias Informtica Vol. 8 No 4, Octubre-Diciembre, 2014, Pag. 21-51.