Action Recognition Applied to Monitor Domestic Animals

Ruth Keglevich de Buzin
Departamento de Ciência da Computação
Universidade Federal de Minas Gerais (UFMG)
Belo Horizonte - Brasil
ruth@ufmg.br

William Robson Schwartz
Departamento de Ciência da Computação
Universidade Federal de Minas Gerais (UFMG)
Belo Horizonte - Brasil
william@dcc.ufmg.br

Abstract—Action Recognition is a problem of visual computing consisting in categorizing sequences of images according to action labels. Robust solutions to this problem have various applications with large impact on improving the quality of life. In this work, we opted for the application of action recognition to the problem of identifying unwanted actions of pets. Our main contributions are two: (1) present an innovative and important application of the problem of action recognition, (2) build "space-time blocks" that simplify the problem by making possible the analysis of sequences of images as a single static image, allowing the use of robust methods for recognizing objects / actions in video sequences.

***Resumo*—Reconhecimento de ações é um problema da computação visual que consiste em categorizar sequencias de imagens em rótulos de ações. Soluções robustas para este problema apresentam diversas aplicações de grande impacto na melhoria da qualidade de vida. Neste trabalho, optamos pela aplicação do reconhecimento de ações ao problema de identificação de ações indesejadas de animais de estimação. Nossas principais contribuições são duas: (1) apresentamos uma aplicação importante e inovadora do problema de reconhecimento de ações; (2) construímos "blocos espaço-temporais" que simplificam o problema ao tornar possível a análise de sequencias de imagens como uma única imagem estática, possibilitando a utilização de métodos robustos para reconhecimento de objetos/ações em sequências de vídeos estáticas.**

I. INTRODUÇÃO

A principal motivação deste trabalho aliou a busca por soluções mais simples e viáveis para o problema de reconhecimento de ações ao problema da disciplina de animais de estimação. Animais domésticos estão presentes nas vidas de muitas pessoas. Segundo a ABINPET – Associação Brasileira da Indústria de Produtos para Animais de Estimação, o Brasil possui a segunda maior população de animais de estimação do mundo, ultrapassando 100 milhões de indivíduos. Tendo em vista a frequência com a qual o animal apresenta dificuldades na adaptação ao novo meio, o que causa desconforto ao o dono e ao próprio animal, tornou-se relevante a utilização de computação visual para a monitoração de ações indesejadas com o propósito educativo. Este trabalho desenvolve uma metodologia de reconhecimento de ação que tem como objetivo futuro a identificação prévia e eficiente de ações indesejadas com o propósito de inibi-las.

O reconhecimento de ações tornou-se uma área de pesquisa bem ativa nos últimos anos [1, 8]. Abordagens usando *bag-of-words* (BoW) baseadas no método *k-means* têm sido extensivamente utilizadas para reconhecimento de ações [9, 10, 11, 12]. Métodos supervisionados que proporcionam uma maior discriminação também são explorados, tais como *codebooks* [13], florestas aleatórias [14] e o aprendizado de matrizes de distância [15]. Apesar dos esforços, métodos baseados em BoW geralmente ignoram a distribuição espaço-temporal dos descritores, o que tende a reduzir significativamente a discriminação obtida pelos descritores usados para representar ações distintas.

Características visuais considerando informações espaço-temporais extraídas de vídeos são cruciais para o reconhecimento de ações. Idealmente, essas características devem generalizar variações na aparência, compensar certa quantidade de ruído presente no fundo e serem capazes de discriminar entre ações distintas. Tais características podem ser divididas em duas categorias: holísticas e locais [8]. A primeira captura a informação visual como um todo, ao longo do tempo, enquanto que a segunda concentra o foco em regiões de tempo e espaço que apresentam alto poder discriminativo entre as ações, sendo menos sensível a variações de ponto de vista e oclusões parciais.

Neste artigo serão descritas e analisadas soluções para o problema do reconhecimento de ação aplicado ao comportamento de um animal doméstico. A ação visada para o trabalho, inicialmente, foi o a demarcação de território em si. Contudo, esta ação é muito sutil e parecida com outras ações como "sentar". Portanto, ao longo do desenvolvimento optou-se pela identificação da ação de "enterrar a areia", que apresenta peculiaridades muito mais significativas.

O restante deste artigo está dividido nas seguintes seções. A Seção II apresenta uma descrição dos descritores de características utilizados, a Seção III descreve os resultados experimentais obtidos com a aplicação da metodologia. Finalmente, as conclusões obtidas são apresentadas na Seção IV.

II. METODOLOGIA

Nesse trabalho as estratégias abordadas foram subtração de fundo [19], extração de características [16, 17, 18] e reconhecimento de ações [1,8]. Para a extração de características utilizamos o descritor SURF [2], que trabalha com características espaciais em imagens, independente de escala ou rotação. Extrapolamos o método para a obtenção de características espaço-temporais. Para a classificação, utilizamos duas abordagens: a primeira consiste na classificação apenas da ação "enterrar na areia" (considerando apenas uma classe) com a aplicação de um limiar fixo para a classificação e a segunda consiste na classificação de outras ações (andar, correr, sentar, comer), visando uma maior diferenciação das ações e, consequentemente, uma maior acurácia no reconhecimento. Os detalhes do método são descritos nas próximas seções.

A. Subtração de Fundo

A subtração de fundo foi obtida com o método de mistura de Gaussianas com base na segmentação de fundo [3, 4, 5, 6], utilizando o algoritmo proposto por P. KadewTraKuPong e R. Bowden [3], que utiliza gaussianas distintas para representar cada cor e atribuir pesos aos parâmetros com base no tempo de permanência de cada cor na cena, definindo o fundo como o conjunto de pixels com a maior probabilidade para um determinado conjunto de cores. O algoritmo apresenta uma melhora desempenho sobre o modelo de Grimson [4, 5, 6]. Neste trabalho aplicamos um filtro de cores no resultado da subtração de fundo, substituindo-as por escalas de cinza.

B. Blocos Espaço-Temporais

Para possibilitar a utilização do descritor SURF [2] em vídeos, foram construídos "blocos espaço-temporais". Os blocos são construídos a partir da segmentação dos vídeos. Essa segmentação depende do intervalo de tempo definido e ocorre de modo linear no vídeo sem nenhuma intervenção manual. Ao término da seleção dos frames relevantes para um determinado intervalo de tempo, os frames são somados, formando uma única imagem e concluindo a construção do bloco. Esse processo é utilizado tanto para o treinamento do classificador, que armazena todos os blocos referentes à ação de interesse do animal em uma galeria de imagens, como para o processo de classificação, o qual segmenta os vídeos de teste e compara cada bloco criado com os blocos da galeria de imagens formada pelo processo de treinamento.

A construção do bloco é variável e se baseia nos parâmetros de tempo (em segundos) e de quantidade de frames relevantes.

Foram utilizadas quatro configurações distintas de blocos:

1s2f: Um segundo com dois frames relevantes.

1s3f: Um segundo com três frames relevantes.

2s2f: Dois segundos com dois frames relevantes.

2s3f: Dois segundos com três frames relevantes.

A seleção do frame relevante é determinada com base na divisão de frames equidistantes dentro do intervalo de tempo definido para a execução de uma determinada ação. O vídeo é então segmentado de acordo com o tempo definido. Os frames selecionados são então somados utilizando um operador de mistura linear

$$g(x) = (1 - \alpha)f_0(x) + \alpha f_1(x),$$

onde α é a opacidade de cada frame, definida em 50%. A Figura 1 exemplifica a configuração 1s2f. Na Figura 1(a), dois frames são selecionados e somados para cada uma das imagens formando dois "blocos" de 1 segundo cada. Nessa configuração não utilizamos o recurso da subtração de fundo. O mesmo ocorre para a Figura 1(b), contudo utilizamos a subtração de fundo para reduzir a influência do ambiente no reconhecimento da ação. De acordo com a Figura 1(b), as regiões que apresentam mudanças são justamente aquelas presentes no corpo do gato, provendo informações cruciais para o reconhecimento das ações.

Fig. 1 (a). *Frames relevantes (2) sobrepostos para o tempo de 1 segundo. Modelo de bloco 1s2f.*

Fig. 1 (b). *Frames relevantes (2) sobrepostos para o tempo de 1 segundo com subtração de fundo utilizando mistura de guassianas (cores originais). Modelo de bloco 1s2f.*

C. Descritores de Características

A extração de características [16, 17, 2] apresenta-se como uma etapa fundamental na resolução dos problemas na área de visão computacional, devido, principalmente, à importância de se obter uma boa representação das informações visuais contidas na imagem ou vídeo para que não haja propagação de erros.

O descritor de características utilizado no trabalho é o *Speeded Up Robust Features* (SURF) devido a sua robustez e desempenho eficiente. SURF [2] é um descritor local de regiões, apresentado pela primeira vez por Herbert Bay *et al.* [2], normalmente aplicado em reconhecimento de objetos, reconstrução tridimensional, dentre outras aplicações. Inspirada pelo descritor *Scale-Invariant Feature Transform* (SIFT) [7], sua versão padrão é várias vezes mais rápida.

1) SURF: Aplicação no Reconhecimento de Ações de Animais Domésticos

Como as atividades a serem reconhecidas são diferenciáveis apenas com a presença de informações temporais, blocos espaço-temporais são capturados a partir da extração do descritor SURF para múltiplos quadros.

O descritor SURF [2] passou a ser aplicado sobre esses blocos de imagens e não mais sobre imagens estáticas. A sequência de imagens passou a ser processada como uma única imagem e as ações de interesse passaram a ser processadas como um "objeto" de interesse. A partir desse ponto, o descritor funciona da mesma forma que na abordagem para reconhecimento de objetos em imagens de duas dimensões.

Nesse trabalho utilizamos as implementações da biblioteca OpenCV [20] para o método SURF, tanto para a descrição de características.

D. Classificadores

O processo de classificação consiste na comparação dos vetores de características extraídos dos blocos referentes à ação de interesse do animal e dos blocos referentes aos vídeos de teste. Os pontos de interesse comparados que apresentarem vetores com as menores distâncias são considerados "bons casamentos".

O treinamento do classificador constrói os blocos da galeria a serem posteriormente comparados com os vídeos de teste. O processo de treinamento necessita da compilação de dois vídeos. O primeiro deles é formado exclusivamente a partir da edição das ações de interesse identificadas nos outros vídeos obtidos do animal, e o segundo vídeo foi obtido a partir da edição de falsos positivos para a ação de interesse. Desta maneira, há exemplos e contraexemplos da ação a ser reconhecida.

A construção dos blocos de interesse depende das configurações de blocos selecionadas. Para as configurações com intervalo de tempo de 1 segundo, um bloco é criado para cada segundo de vídeo. No processo de treinamento do classificador, quanto menor o número de blocos armazenados, melhor o desempenho para o processo de classificação. Com o propósito de diminuir o custo computacional e de evitar a criação de blocos que possam gerar muitos erros de identificação, o processo de treinamento efetua uma validação dos blocos criados. Essa validação consiste na aplicação do processo de classificação sobre o vídeo que originou os blocos, ou seja, o vídeo editado para conter apenas as ações de interesse do animal. Um bloco é considerado válido, ou relevante, se obtiver mais de 100 "bons casamentos" com pelo menos um bloco do vídeo de teste. O número mínimo de acertos foi definido como padrão em 1.

Por fim, os blocos são testados em um vídeo de falsos positivos, sendo, ao contrário do primeiro caso, eliminados os blocos que obtiverem mais de 100 "bons casamentos" com pelo menos um bloco do vídeo de teste. O número máximo de erros foi definido como padrão em 0.

1) Abordagem Utilizando Limiar

Na primeira abordagem, o classificador foi treinado apenas para a ação de interesse em questão. Definimos, então, um limiar fixo para o número de "bons casamentos" originados da comparação entre um bloco treinado e um bloco do vídeo de testes. Dessa maneira, a identificação positiva ocorre quando a função de casamento retorna um número superior ao limite pré-definido de "bons casamentos" para a ação de interesse treinada (no caso, "enterrar na areia"). Esse limiar foi definido com base na média de "bons casamentos" observada experimentalmente.

De acordo com os experimentos mostrados na seção III, essa abordagem apresentou uma média de erros elevada, por isso foi necessário revisá-la conforme descrito a seguir. Acreditamos que o número elevado de erros tenha ocorrido em função de o limiar ser muito alto, visto que a maioria dos erros nessa abordagem foram falhas na identificação positiva, ou seja, blocos contendo a ação de interesse nos vídeos de teste não foram devidamente identificados.

2) Múltiplas Classes

Na segunda abordagem, mais de uma classe de ações foi considerada. Uma classe de ações é o conjunto de movimentos diferenciados que caracterizam uma única ação, como, por exemplo, todas as formas diferentes de "sentar" caracterizam a ação de "sentar", logo, pertencem à classe de ações "sentar".

Utilizamos múltiplas classes com o propósito de não mais necessitar de um limiar fixo de "bons casamentos". Devido ao alto índice de erros de classificação no método de classificação baseada na utilização de limiar, tornou-se necessária uma aproximação mais eficiente do número de "bons casamentos", visto que, dependendo da qualidade do vídeo ou da iluminação, esse limiar poderia ser muito mais baixo, ocasionando muitos erros com a utilização de um limiar fixo. Na abordagem para múltiplas classes, o processo de classificação passou a considerar os casamentos obtidos com cada classe distinta de ações, contabilizando o maior número de "bons casamentos" obtidos em cada classe, de forma a associar o bloco testado à classe com maior pontuação (maior número de "bons casamentos").

Para efetuar o treinamento dessas classes, vídeos contendo outras ações (que não apenas a de interesse) e seus respectivos falsos positivos foram considerados. Os blocos criados para cada classe foram então armazenados em amostras distintas. No processo de classificação, os blocos contendo amostras de todas as classes são comparados com os blocos do vídeo de teste. Aquela classe que obtiver o maior número de "bons casamentos" será considerada como será atribuída à amostra de teste.

Como será apresentado na próxima seção, essa abordagem apresentou um resultado significativamente melhor em comparação com a abordagem baseada em limiar.

III. RESULTADOS EXPERIMENTAIS

Avaliamos o desempenho das abordagens para cada configuração de "blocos espaço-temporal", para as abordagens de classificação baseada em limiares e múltiplas classes, e para os métodos de subtração de fundo utilizados. Os resultados foram obtidos com a utilização de blocos formados a partir da classificação dos vídeos de testes. Os blocos identificados como contendo a ação de interesse foram salvos

com uma marcação identificando a ação, enquanto os blocos sem identificação positiva foram salvos sem alteração. A verificação do resultado foi realizada manualmente, verificando em cada um dos blocos salvos se a ação havia sido corretamente identificada.

As próximas seções mostram resultados experimentais utilizando a abordagem baseada na aplicação de um limiar para classificação e na abordagem utilizando múltiplas classes, a qual não necessita de limiar, apenas atribui a amostra à classe mais próxima.

A. Abordagem Baseada em Limiar

A tabela 1 mostra o resultado experimental para o reconhecimento de ações. Para este experimento, utiliza-se a abordagem baseada em limiar de classificação, exibindo o número de erros e acertos para cada método de subtração de fundo e para cada método de agrupamento de frames relevantes (bloco). Os erros são divididos entre *falsos positivos* (ação reconhecida indevidamente) e *falsos negativos* (deixou de reconhecer a ação quando deveria). Os acertos são divididos entre *verdadeiros positivos* (identificou corretamente a ação) e *verdadeiros negativos* (não identificou indevidamente uma ação irrelevante). A sigla SF significa Subtração de Fundo e S/F indica a não utilização da subtração de fundo.

Bloco	SF	FP	FN	VN	VP	Acurácia %
1s2f	S/ SF	18	40	77	96	74,89
1s2f	SF	21	46	81	83	71,00
1s3f	S/ SF	20	51	66	94	69,26
2s2f	S/ SF	8	80	38	106	62,07
2s3f	S/ SF	8	78	36	110	62,93

Tabela 1. Resultados obtidos com a aplicação do método limiar. Onde, S/SF indica a não utilização de subtração de fundo, SF indica subtração de fundo com filtro de cores, FP indica o número de falsos positivos FN o número de falsos negativos, VN indica o número de verdadeiros positivos e VP o número de verdadeiros positivos.

De acordo com os resultados, a configuração considerando um segundo com dois frames relevantes (1s2f) obteve os melhores resultados comparando com as configurações 1s3f, 2s2f e 2s3f. Adicionalmente, a versão do 1s2f que obteve a maior acuraria (74,89%) foi aquela que não aplica a subtração de fundo. Percebe-se que a aplicação da subtração de fundo obteve resultados menos acurados que aqueles obtidos sem a aplicação da técnica.

B. Múltiplas Classes

A tabela 2 mostra o resultado experimental para a identificação exibindo o número de erros e acertos para para cada método de agrupamento de frames relevantes (bloco).

Bloco	SF	FP	FN	VN	VP	Acurácia %
1s2f	S/ SF	12	3	118	36	91,1
1s2f	SF	12	51	70	36	62,7
1s3f	S/ SF	5	8	113	43	92,3
2s2f	S/ SF	12	14	106	38	85,2
2s3f	S/ SF	10	8	100	44	85,3

Tabela2. Método múltiplas classes para teste de base Verdadeiro Positiva. S/SF indica a não utilização de subtração de fundo, SF indica subtração de fundo com filtro de cores, FP indica o número de falsos positivos FN o número de falsos negativos, VN indica o número de verdadeiros positivos e VP o número de verdadeiros positivos.

O resultado chegou a uma média de 92,3% de acertos para as configurações sem subtração de fundo, sendo a melhor configuração a 1s3f.

C. Discussão

O método que utiliza múltiplas classes apresentou uma acurácia maior em comparação com o método limiar. Contudo, o custo computacional para o método que considera múltiplas classes ficou muito elevado, chegando a 400% do tempo de processamento do método baseado em limiar devido à consideração das várias classes de ações, o que faz da implementação inviável para o uso prático em tempo real.

No gráfico da Figura 2 é possível observar o ganho que cada método de configuração de bloco obteve com o método múltiplas classes em comparação com o método baseado em um limiar.

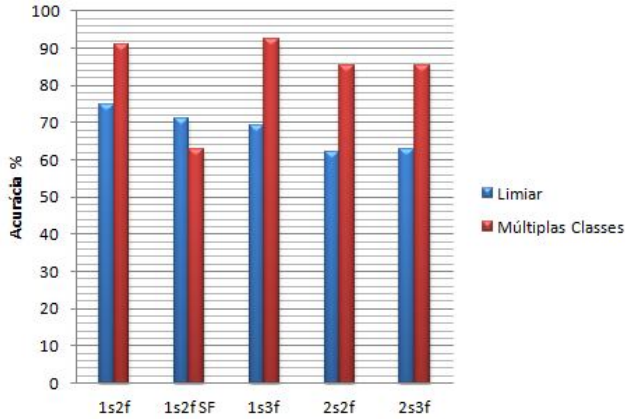

Figura 2. *Acurácia de cada configuração de blocos para os dois métodos de classificação. O conjunto de barras mais a esquerda (1s2f) foi obtida sem subtração de fundo enquanto que o segundo conjunto de barras (1s2f SF) foi obtido com a utilização de subtração de fundo, os demais conjuntos (1s3f, 2s2f, 2s3f) foram obtidos sem a subtração de fundo.*

No método de classificação baseada em limiar, a configuração de bloco que obteve maior acurácia foi a 1s2f S/SF (1 segundo, 2 frames relevantes, sem subtração de fundo), chegando quase a 75%. O segundo melhor resultado para a classificação baseada em limiar foi com a utilização de subtração de fundo, na configuração 1s2f SF, no entanto, a mesma configuração apresenta o pior resultado dentre todos obtidos para a classificação com múltiplas classes. Portanto, conclui-se que a aplicação de subtração de fundo não apresenta uma boa contribuição para a resolução do problema.

As configurações de bloco com mais de 2 segundos não apresentaram bons resultados em nenhum dos métodos, principalmente no método limiar. Isso se deve ao fato de que na maioria das ações de enterrar registradas, o gato concluía um ciclo de movimento em apenas 1 segundo. Mesmo para esse intervalo de tempo, o método de múltiplas classes apresentou bom resultados, o que torna essa configuração mais viável (visto que é processada na metade do tempo).

Apesar do método de múltiplas classes ter melhorado em muito a acurácia, é um método com elevado custo computacional, tornando-se inviável para o uso em tempo real.

Além do alto custo computacional, outro fator prejudicial ao método de múltiplas classes foi o resultado para a subração de fundo, que piorou 5 pontos percentuais.

O método também não mostrou uma diminuição significativa dos erros de classificação indevida (falso positivo). Sua principal melhora foi na diminuição dos erros de classificação positiva (falso negativo). Como objetivo do trabalho é a disciplina de animais domésticos, os erros não apresentam a mesma importância. Por exemplo, um erro de classificação indevida (falso positivo) é muito mais grave, visto que um ruído sonoro incômodo estaria sendo disparado indevidamente. Desta maneira, uma métrica de pesos poderia ser adotada de forma a penalizar mais os erros de falso positivo e menos os erros de falso negativo, aproximando os resultados dos métodos de múltiplas classes e limiar. Tornando, desta maneira, viável a aplicação dessa abordagem em um cenário de tempo real com baixo custo computacional e alta acurácia.

D. Resultados Qualitativos

As Figuras 3(a) e 3(b) apresentam resultados qualitativos para casos de classificação de falsos positivos e verdadeiros positivos.

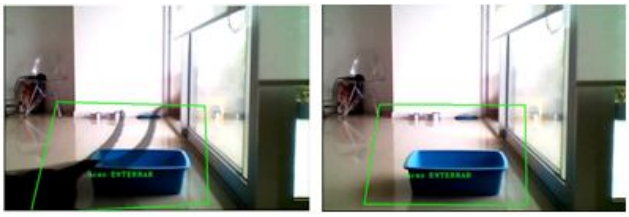

Fig. 3a. *Falso positivos identificados.*

Fig. 3b. *Verdadeiro positivos identificados.*

O falso positivo que classifica a caixa de areia vazia como "ação enterrar", identificado na Figura 3(a), foi um erro comum de classificação do método limiar sobre as filmagens realizadas no mesmo horário do dia, o que demonstra a influência da iluminação e da formação de sombras na classificação limiar da ação. Esse erro foi corrigido no método de múltiplas classes devido a classificação da amostra em relação a outras classes e não de acordo com um limiar.

IV. CONCLUSÕES

Neste trabalho foram desenvolvidos e testados métodos distintos de reconhecimento de ações utilizando o descritor SURF [2]. O método de classificação que utiliza múltiplas classes apresentou um ganho surpreendente em comparação com o método limiar, sendo mais significativo para as configurações de cubo 1s3f (1 segundo, 3 frames relevantes). A implementação em tempo real seria viável para o método limiar, com a configuração de cubo 1s2f S/SF, que apresenta 75% de acurácia e um baixo custo computacional. Entretanto 75% ainda é um valor muito baixo na prática, tendo em vista que a maior parte dos erros é de classificação indevida (falso positivo).

Como trabalho futuro, pretende-se avaliar a utilização de outras técnicas de subtração de fundo com intuito de se obter resultados mais acurados com a aplicação da abordagem baseada em limiares. Desta maneira, seria possível aplicar na prática o sistema de disciplina de animais domésticos com base em ruídos sonoros.

AGRADECIMENTOS

Os autores gostariam de agradecer a FAPEMIG, o CNPq e a CAPES pelo apoio financeiro.

REFERÊNCIAS

[1] Weinland, Daniel and Ronfard, Remi and Boyer, Edmond. "A survey of vision-based methods for action representation, segmentation and recognition." Journal Computer Vision and Image Understanding. vol. 115, n. 2, pp. 224--241, 2011.

[2] Herbert Bay, Tinne Tuytelaars and Luc Van Gool. "SURF: Speeded Up Robust Features". Journal Computer Vision and Image Understanding. 2006.

[3] P. KaewTraKulPong and R. Bowden. "An Improved Adaptive Background Mixture Model for Realtime Tracking with Shadow Detection", Proceedings of 2nd European Workshop on Advanced Video Based Surveillance Systems. 2001.

[4] Grimson Wel, Stauffer C. Romano R. Lee L. "Using adaptive tracking to classify and monitor activities in a site" Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1998.

[5] Stauffer C, Grimson W. E. L. "Adaptive background mixture models for real-time tracking. in Proceedings.", Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, 1999.

[6] Stauffer C, Grimson W. E. L. "Learning patterns of activity using real-time tracking.", Proceedings of the IEEE Computer Society Conference on Transactions on Pattern Analysis & Machine Intelligence, 22(8): p. 747-57, 2000.

[7] Ke, Y., Sukthankar, R. “PCA-SIFT: A more distinctive representation for local image descriptors.” Computer Vision and Pattern Recognition CVPR (2). pp 506 – 513, 2004.

[8] Moeslund, Thomas B. and Hilton, Adrian and Krüger, Volker. "A Survey of Advances in Vision-Based Human Motion Capture and Analysis" Journal Computer Vision and Image Understanding, vol 104, n. 2, pp. 90 – 126, 2006.

[9] Poppe, R. "A Survey on Vision-Based Human Action Recognition" Journal Image and Vision Computing, vol 28, n. 6, pp. 976 – 990, 2010.

[10] Ivan Laptev and Marcin Marszalek and Cordelia Schmid and Benjamin Rozenfeld. "Learning Realistic Human Actions from Movies" Computer Vision and Pattern Recognition CVPR, 2008.

[11] Jingen Liu and Mubarak Shah. "Learning human actions via information maximization" Computer Vision and Pattern Recognition CVPR, 2008

[12] Niebles, Juan Carlos and Wang, Hongcheng and Fei-Fei, Li. "Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words" International Journal of Computer Vision, vol 79, n. 3, pp. 299 - 318, 2008.

[13] Junsong Yuan and Zicheng Liu and Ying Wu. "Discriminative Subvolume Search for Efficient Action Detection" Computer Vision and Pattern Recognition CVPR pp. 2442 - 2449, 2009.

[14] Liu Yang and Rong Jin and Rahul Sukthankar and Frederic Jurie. "Unifying Discriminative Visual Codebook Generation with Classifier Training for Object Category Recognition" Computer Vision and Pattern Recognition CVPR, 2008.

[15] Frank Moosmann and Bill Triggs and Frèderic Jurie. "Fast Discriminative Visual Codebooks using Randomized Clustering Forests" Advances in Neural Information Processing Systems, 2006.

[16] Bilenko, Mikhail and Basu, Sugato and Mooney, Raymond J. "Integrating Constraints and Metric Learning in Semi-Supervised Clustering" International Conference on Machine Learning, 2004.

[17] Tuytelaars, Tinne and Mikolajczyk, Krystian. "Foundation and Trends in Computer Graphics and Vision" FTC, PP. 177 - 280, 2008.

[18] Zhang, J. and Marszalek, M. and Lazebnik, S. and Schmid, C. "Local Features and Kernels for Classification of Texture and Object Categories: A Comprehensive Study" International Journal of Computer Vision , vol 73, n. 2, pp. 213 - 238, 2007.

[19] Gauglitz, Steffen and Höllerer, Tobias and Turk, Matthew. "Evaluation of Interest Point Detectors and Feature Descriptors for Visual Tracking" International Journal of Computer Vision, vol 94, n. 3, pp. 335 - 360, 2011.

[20] OpenCV (Open Source Computer Vision) Library. www.opencv.org