Classificação não-supervisionada de eventos em imagens de videomonitoramento baseada em altas frequências do fluxo ótico diferencial

Ana Paula Gonçalves S. de Almeida
Programa de Pós-Graduação em Sistemas Mecatrônicos
Universidade de Brasília
Email: anapaula.gsa@gmail.com

Flávio de Barros Vidal
Departamento de Ciência da Computação
Universidade de Brasília
Email: fbvidal@unb.br

***Abstract*—Events related to vandalism and violence often occur in crowded environments and mostly in unstructured environments (dynamic). The use of security cameras in crowd monitoring analysis for anomaly detection and alarms could be an efficient and inexpensive method. The goal of this paper is to build an unsupervised classification method to abnormal events that is robust and stable. The main idea is based on analysis of Fourier Transform's high-frequency spatial components features of optical flow, allowing the detection of abnormal acts in surveillance videos. Preliminary results show that the proposed methodology is capable of successfully execute the process of detection, permitting the development of an efficient recognition stage for future works.**

I. Introdução

Eventos no mundo real ocorrem frequentemente em ambientes não estruturados (dinâmicos) e com grande aglomeração de pessoas. Aplicações para câmeras de segurança podem ser utilizadas na análise de multidões para detecção de anomalias e situações de risco [Ke et al., 2007]. Dessa forma, o comportamento de multidões atrai o interesse de muitos pesquisadores devido à sua complexidade e possibilidades de abstrações nas aplicações [Husni and Suryana, 2010]. Existem diversos obstáculos que dificultam este processo como: oclusão, mudanças de iluminação e outros fatores que possam influenciar o processo de detecção. Atualmente, o uso de câmeras de videomonitoramento é maior em ambientes externos do que internos, sejam estas instaladas em prédios públicos ou privados, para manter a segurança dos usuários. Entretanto, considerando este aumento de equipamento, o acompanhamento de todas as imagens capturadas por várias câmeras distintas é fatigante e deve ser feito com atenção, além de requerer uma maior quantidade de pessoas para monitorar todas as imagens. Ademais, há também a propensão ao fator erro humano. Portanto, técnicas que desenvolvam uma aplicação automática para detectar e classificar eventos em multidões que sejam precisas e eficientes são muito importantes para este mercado com alto crescimento no país.

De acordo com os argumentos apresentados, o desenvolvimento inicial de um sistema robusto que é capaz de detectar eventos em multidões baseado em análises das características dos componentes de alta frequência do fluxo ótico é proposto. A Seção II descreve os principais trabalhos relacionados à detecção de eventos em multidões. Nas Seções III e IV, a metodologia proposta e os resultados são apresentados, respectivamente. Conclusões e trabalhos futuros são discutidos na Seção V.

II. Revisão Bibliográfica

De acordo com [Liao et al., 2011] e [Ke et al., 2007], a detecção de eventos em multidões é uma tarefa fundamental para a segurança pública. [Li et al., 2012] descreve que na análise de vídeos de segurança em ambientes com multidões, a classificação de eventos é um ponto crítico, requerendo maior atenção.

Para a detecção de eventos em multidões existem alguns trabalhos recentes que descrevem técnicas automáticas para detectar o tempo exato no qual um evento (como ações violentas) ocorre. Em [Xu et al., 2014] um modelo de classificação *Bag-of-Words* (*BoW*) e uma fusão entre *BoW* e o algoritmo de *Motion SIFT* é proposto.

Em [Esen et al., 2013], um novo modelo de características de movimento é criado, chamado *Motion Co-Occurrence Feature* (*MCF*), além de um modelo baseado em energia, que utiliza informações da velocidade de movimento do alvo estimado por fluxo ótico e uma abordagem entrópica para características de desordem.

III. Metodologia

Para o processo de detecção de eventos anormais em câmeras de videomonitoramento, realiza-se em cada *frame* do vídeo os seguintes passos, descritos na Figura 1 e nas subseções seguintes.

A. Sequência de Imagens de Entrada

Nessa abordagem, apenas imagens de vídeos de monitoramento são utilizadas. Um vídeo de monitoramento é capturado a partir de uma câmera fixa encontradas em locais com grande concentração de pessoas ou com elevados índices de violência [Kruegle, 2011]. Nenhuma outra informação, *a priori*, sobre as cenas é usada e toda a informação do processamento de imagem é feita diretamente no *frame* atual do vídeo, sem nenhum pré-processamento, sendo este *frame* definido por um índice temporal $k + 1$ e comparado com o *frame* anterior no tempo k.

B. Fluxo Ótico

O fluxo ótico aproxima o campo de movimento da imagem através da representação do movimento aparente

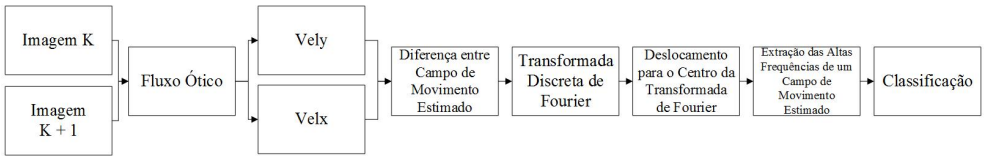


Figura 1: Fluxograma da metodologia proposta.

do padrão de brilho da imagem no plano da imagem [Horn and Schunck, 1981]. Para determinar o fluxo ótico, dois aspectos devem ser levados em conta: o nível de acurácia dos dados relativos à direção de movimento e intensidade e as propriedades relacionadas à carga computacional requerida para determinar o fluxo ótico sob condições mínimas de precisão. O compromisso entre esses aspectos dependem da situação analisada e dos resultados esperados [Liu et al., 1998]. Nessa abordagem, apenas as técnicas diferenciais são consideradas.

De acordo com [Horn and Schunck, 1981], o fluxo ótico não pode ser calculado em um ponto da imagem independentemente dos pontos vizinhos sem que haja limitações adicionais. Isto ocorre porque o campo de velocidade em cada ponto da imagem possui dois componentes, enquanto a mudança no brilho naquele ponto possui apenas uma restrição, devido ao movimento. Antes de descrever o método, algumas condições devem ser satisfeitas. Assume-se que a velocidade aparente dos padrões de brilho pode ser diretamente identificada com o movimento das superfícies na cena. Isso implica, de acordo com a superfície do objeto que se move, que não há, ou existe pouca, variação de brilho. Denota-se $I(x, y, t)$ como o brilho da imagem no instante de tempo t no ponto da imagem (x, y). Durante o movimento, supõe-se que o brilho de um ponto particular é constante, significando que

$$\frac{dI\,(x, y, t)}{dt} = 0 \tag{1}$$

Expandindo e reescrevendo a Equação 1

$$I_x u + I_y v + I_t = 0 \tag{2}$$

I_x, I_y e I_t representam as derivadas parciais de brilho em x, y e t, respectivamente; u e v são as componentes da velocidade x e y. Assim, o padrão de brilho pode se mover independentemente do resto da cena e há possibilidade de recuperação da informação da velocidade.

O cálculo da restrição adicional da velocidade é resultado a partir da hipótese de que o campo de velocidade é suave. O fator de ponderação α^2 é introduzido para associar a magnitude do erro com a quantização de erros e ruído. Os valores estimados para as componentes da velocidade de u_{k+1} e v_{k+1} são obtidos de

$$u^{k+1} = \overline{u}^k - \frac{I_x \left[I_x \overline{u}^k + I_y v^k + I_t\right]}{\left(\alpha^2 + I_x^2 + I_y^2\right)} \tag{3}$$

$$v^{k+1} = \overline{v}^k - \frac{I_y \left[I_x \overline{u}^k + I_y v^k + I_t\right]}{\left(\alpha^2 + I_x^2 + I_y^2\right)} \tag{4}$$

Nas Equações 3 e 4, $\overline{u}^k$ e $\overline{v}^k$ são as velocidades médias aproximadas das derivadas parciais dos padrões de brilho na iteração k, na qual os valores de pixels vizinhos são ponderados com a máscara mostrada em [Horn and Schunck, 1981].

C. Diferença de Campo de Movimento Estimado

Após a obtenção do fluxo ótico de Horn e Schunck, duas componentes de velocidade são estimadas nas direções horizontais (u) e verticais (v). As componentes possuem o mesmo tamanho da imagem de entrada e descrevem o campo de movimento no domínio da imagem.

Para o processo de detecção de eventos anormais em multidões, a diferença absoluta de cada elemento do fluxo ótico estimado de $k+1$ e k é calculada. Esta abordagem é realizada para compensar as altas variações de luminância nas cenas e afetam diretamente a técnica de fluxo ótico utilizada.

D. Transformada Discreta de Fourier

Optou-se na utilização da Transformada Discreta de Fourier (DFT) para a extração dos componentes da alta frequência, descrita na Equação 5.

$$F(k, l) = \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} f(i, j) e^{-i2\pi\left(\frac{ki}{N} + \frac{lj}{N}\right)} \tag{5}$$

De acordo com a Equação 5, os espectros de alta e baixa frequência são calculados após a transformação e adquiridos de cada componente sinusoidal e invariante da informação espacial da imagem. As altas frequências estão, em sua maioria, concentradas nas bordas, enquanto as baixas frequências estão no centro. Para uma melhor classificação, todos os valores de frequências nas direções horizontal (Vel_x) e vertical (Vel_y) são deslocados para o centro, como demonstrado na Figura 6, após o cálculo DFT.

Quando todas as magnitudes são deslocadas para o centro, uma inversão espacial de posições das altas frequências e baixas frequências é realizada. Dessa forma, os componentes de alta frequência são movidos para as bordas do espectro de magnitude e as baixas frequências movidas para o centro do espectro de magnitude. O deslocamento da DFT é uma etapa essencial para a classificação baseada nas distribuições espaciais dos componentes de alta frequência.

E. Extração dos Componentes de Alta Frequência do Campo de Movimento Estimado

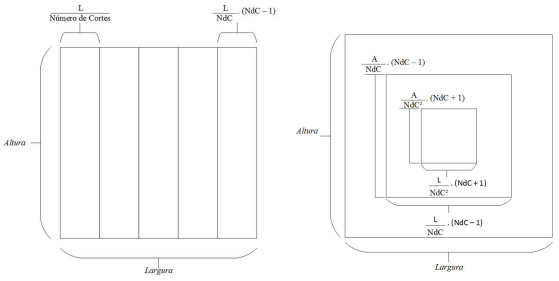
Para fazer a detecção e classificação proposta de eventos anormais, assume-se a hipótese de que quando há uma situação anormal, como uma briga generalizada, em um ambiente com muitas pessoas, grande parte dos componentes de altas frequências têm uma variação de amplitude, significando que estas variações podem ser detectadas, na dispersão espacial quando comparadas com a dispersão original desses componentes (no caso de situações sem a presença de ações de violência).

Para realizar a construção do descritor, duas abordagens que permitem avaliar estas características espaciais (retângulos

e aros retangulares) baseadas apenas na análise da dispersão espacial foram elaboradas, calculando as mudanças em regiões em que as altas frequências estão localizadas e descartando as regiões das baixas frequências que estão localizadas no espectro. Os cálculos destas regiões se dão pelas médias da soma de todos os elementos do espectro dentro da região específica. É uma abordagem simples, mas suficiente para indicar a comprovação da hipótese, que nesse caso forneceria informação suficiente para a construção de um descritor capaz de realizar a classificação proposta.

1) Extração por Retângulos: A extração por retângulos foi idealizada de acordo com a Figura 2a. A soma dos elementos é feita em cada divisão (retângulo) e a metodologia aplicada nas magnitudes deslocadas dos valores Vel_x e Vel_y. Na Figura 2a, L representa a largura e NdC o número de cortes/divisões retangulares.

2) Extração por Aros Retangulares: Pela disposição das frequências após o deslocamento da Transformada de Fourier, a extração de altas frequências em formatos de aros retangulares é a consideração mais natural. De acordo com a Figura 2b, cada divisão é considerada de forma distinta. Para cada aro retangular selecionado, os demais são descartados, preenchendo-os com valor zero. Após esse processo, a soma dos elementos do espectro de frequência do aro selecionado é realizada. Na Figura 2b, a variável A representa a altura, L a largura e NdC o número de cortes.



(a) Extração por retângulos. (b) Extração por aros retangulares.

Figura 2: Métodos de extração.

F. Classificação

Para a classificação, optou-se por uma técnica não supervisionada. Esta escolha foi adotada como um mecanismo para avaliar a capacidade de discriminação que um descritor possui, sem a utilização de nenhum treinamento. O classificador adotado foi o algoritmo *k-means* [MacQueen, 1967]. Nessa proposta, o interesse é feito em avaliar o potencial que os componentes de alta frequências provenientes da extração do fluxo ótico de um vídeo permite separar duas classes de ações contidas no vídeo (violência e não-violência). Utilizando a metodologia proposta e observando a dispersão dos pontos no gráfico valores na direção horizontal e vertical dos valores médios dos vídeos testados, fez-se o uso de agrupamento por centróide, definindo assim um agrupamento que permitiu a separação destas duas classes ($k = 2$). Na Seção IV serão apresentados maiores informações sobre o comportamento da dispersão obtida durante os testes realizados.

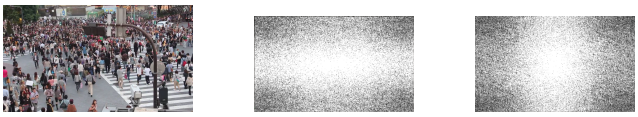
IV. Resultados Preliminares

A metodologia proposta foi testada em vinte arquivos de vídeo distintos, com duração de trinta segundos cada, separados em duas classes: Uma classe chamada atos violentos (dez vídeos); Uma classe chamada atos não-violentos (dez vídeos). A classe não-violentos possui vídeos com grandes quantidades de pessoas, mas sem atividades consideradas fora da normalidade. As Figuras 3 e 4 apresentam um *screen-shot* de alguns dos *frames* dos vídeos contendo atos violentos e não-violentos, respectivamente, usados para construir a base de dados de vídeos. Todos os vídeos foram retirados de vídeos públicos disponíveis na internet [1].

Figura 3: *Screen-shot* da Classe Violentos.

Figura 4: *Screen-shot* da Não-violentos.

A partir das Figuras 5 e 6, há uma forte evidência de que observando a dispersão de componentes, as duas classes propostas de classificação são distinguidas. Os resultados preliminares relatados nessa Seção cobrem apenas as condições mínimas necessárias para a validação de um descritor estável capaz de realizar o processo de reconhecimento e classificação não-supervisionada, proposto na Seção III.



(a) Vídeo original calmo (*Screen-shot*). (b) Componentes de amplitude horizontais da DFT sem violência. (c) Componentes de amplitude verticais da DFT sem violência.

Figura 5: Componentes de frequências extraídas da classe de não-violência.

(a) Vídeo original de violência (*Screen-shot*). (b) Componentes de amplitude horizontais da DFT com violência. (c) Componentes de amplitude verticais da DFT com violência.

Figura 6: Componentes de frequência extraídos da classe com violência.

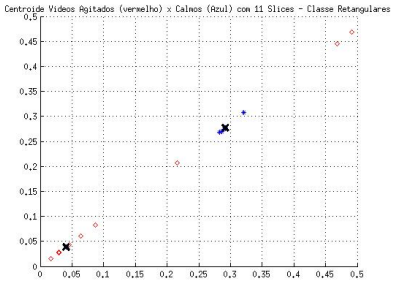
Assim, quando essas são comparadas, ocorre uma quantidade considerável de variações em amplitude e dispersão espacial dos coeficientes de alta frequência (as regiões de borda da imagem de amplitude da DFT). Também, como indicado

[1] Os vídeos utilizados para compor esta base de dados foram retirados do sítio https://www.youtube.com.

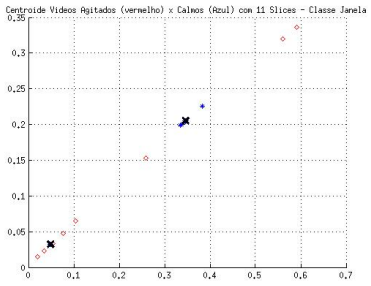
nas Figuras 5 e 6, os componentes de alta frequência do campo de movimento podem ser usados para realizar a detecção por meio de duas abordagens desenvolvidas: Uma que avalia a relação entre número de cortes e divisões utilizados e outra avaliando a forma dos cortes e divisões (retangular e aros retangulares).

A. Avaliação do Número de Cortes

Para o teste do número de cortes, calcula-se a média da energia da alta frequência (nas direções horizontal e vertical) das componentes da DFT, variando entre 3, 5, 7 e 11 cortes retangulares e aros para cada vídeo no processo de validação. Os resultados para o número de corte igual a 11 são apresentados na Figura 7. Na Figura 7, os pontos azuis representam a classe de vídeos não-violentos e os pontos vermelhos representam a classe dos vídeos violentos. Os eixos x e y descrevem as direções horizontal e vertical, respectivamente, das amplitudes dos componentes de alta frequência da DFT. Os pontos X (em preto) representam as centroides dos agrupamentos.



(a) Centroides da energia em retângulos



(b) Centroides da energia em aros retangulares

Figura 7: Número de cortes iguais a 11 - Energias horizontais e verticais da DFT.

B. Avaliação do Tipo de Cortes

As características de distribuição espacial dos componentes de alta frequência do campo de movimento após o processo de deslocamento da DFT, são os maiores indicadores para a detecção de eventos em multidões. Os resultados na Figura 7 mostram que há, claramente, nuvens de pontos da mesma classe agrupados, em ambos os métodos de extração. Apesar de pensar, *a priori*, que o método de extração por aros retangulares retornariam melhores resultados, observa-se que de acordo com a forma de corte mais simples obtida (retangular), é possível construir um descritor que permite discriminar os vídeos entre classes de atos calmos ou agitados em multidões. A partir do descritor, há evidências suficientes para a criação de um classificador. Como parâmetros para avaliar se a classificação foi bem sucedida, duas classes devem ser encontradas, "Agitado"ou "Calmo". O Verdadeiro Positivo é a classe "Agitado". Sabendo que as Equações 6, 7 e 8 são os cálculos de precisão, revocação e medida F, os seguintes resultados para ambos os métodos de extração foram atingidos: Precisão $= 63\%$, revocação $= 70\%$ e medida F $= 66.3\%$.

$$\text{Precisão} = \frac{\text{Verdadeiro Positivo}}{\text{Verdadeiro Positivo} + \text{Falso Positivo}} \tag{6}$$

$$\text{Revocação} = \frac{\text{Verdadeiro Positivo}}{\text{Verdadeiro Positivo} + \text{Falso Negativo}} \tag{7}$$

$$\text{Medida F} = 2 * \frac{\text{Precisão} * \text{Revocação}}{\text{Precisão} + \text{Revocação}} \tag{8}$$

V. Conclusão

Nesse artigo, novas estratégias para analisar a influência das altas frequências da Transformada Discreta de Fourier de um campo de movimento estimado foram apresentadas, além de notar como as altas frequências podem dar informações sobre a detecção de eventos em multidões. Usando os gráficos e informações mostradas na Seção IV, é possível afirmar que mesmo com uma classificação simplificada, os valores iniciais são aceitáveis, validando a metodologia.

Os próximos passos da pesquisa vão em direção à exploração das formas de cortes e a quantidade de cortes, para que um melhor resultado para o descritor proposto seja alcançado. Então, todos os esforços serão concentrados em projetar um classificador mais robusto e eficiente para um sistema completo de detecção de eventos em multidões. Outros trabalhos podem incluir a implementação das estratégias propostas em uma linguagem de programação de alto nível para habilitar a operação em cenários de tempo real (incluindo análise de tempo) e utilizando vídeos reais, além de fazer comparações com as últimas técnicas disponíveis em detecção e reconhecimento de eventos em multidões.

Referências

[Esen et al., 2013] Esen, E., Arabaci, M., and Soysal, M. (2013). Fight detection in surveillance videos. In *Content-Based Multimedia Indexing (CBMI), 2013 11th International Workshop on*, pages 131–135.

[Horn and Schunck, 1981] Horn, B. K. P. and Schunck, B. G. (1981). Determining optical flow. In *Artificial Intelligence*, number 17, pages 185–204.

[Husni and Suryana, 2010] Husni, M. and Suryana, N. (2010). Crowd event detection in computer vision. In *Signal Processing Systems (ICSPS), 2010 2nd International Conference on*, volume 1, pages V1–444–V1–447.

[Ke et al., 2007] Ke, Y., Sukthankar, R., and Hebert, M. (2007). Event detection in crowded videos. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8.

[Kruegle, 2011] Kruegle, H. (2011). *CCTV Surveillance: Video Practices and Technology*. CCTV Surveillance Series. Elsevier Science.

[Li et al., 2012] Li, G., Chen, J., Sun, B., and Liang, H. (2012). Crowd event detection based on motion vector intersection points. In *Computer Science and Information Processing (CSIP), 2012 International Conference on*, pages 411–415.

[Liao et al., 2011] Liao, H., Xiang, J., Sun, W., Feng, Q., and Dai, J. (2011). An abnormal event recognition in crowd scene. In *Image and Graphics (ICIG), 2011 Sixth International Conference on*, pages 731–736.

[Liu et al., 1998] Liu, H., Hong, T., Herman, M., Camus, T., and Chellappa, R. (1998). Accuracy vs efficiency trade-offs in optical flow algorithms. In *Computer Vision and Image Understanding*, number 72:3, pages 271–286.

[MacQueen, 1967] MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations.

[Xu et al., 2014] Xu, L., Gong, C., Yang, J., Wu, Q., and Yao, L. (2014). Violent video detection based on mosift feature and sparse coding. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 3538–3542.