Crowd Event Detection in Surveillance Video
An Approach based on Optical Flow High-frequency Feature Analysis

Ana Paula G. S. de Almeida, Vitor de Azevedo Faria and Flavio de Barros Vidal
Department of Computer Science, University of Brasilia, Brasilia, Distrito Federal, 70.910-900, Brazil

Keywords: Crowd Event Detection, Optical Flow, High-frequency Feature.

Abstract: Many real-world actions occur often in crowded and dynamic environments. Video surveillance application uses crowd analysis for automatic detection of anomalies and alarms. In this position paper we propose a crowd event detection technique based on optical flow high-frequency feature analysis to build a robust and stable descriptor. The proposed system is designed to be used in surveillance videos to automatic violence acts detection. Preliminary results show that the proposed methodology is able to perform the detection process with success and allows the development of an efficient recognition stage in further works.

1 INTRODUCTION

Many real-world actions occur often in crowded and dynamic environments. Video surveillance application uses crowd analysis for automatic detection of anomalies and alarms (Ke et al., 2007). In order, behavior of the crowd attracts many researchers interest because of its complexity and abstraction (Husni and Suryana, 2010). There are several obstacles, such as, occlusion, illumination changes and other obstacles that could influence detecting process, also there are some difficulties in analyzing crowd event.

Nowadays, the use of surveillance systems are higher in outside and inside environments, in public or private buildings, to maintain the safety of the users. But, considering this increase of equipment (cameras in general), it is difficult to check up all images captured from the many cameras and it requires many people to monitor all these images. Therefore, techniques that develop an accurate and efficient application to perform crowd event detection are very important for this high-growth market.

Supported by previous arguments, in this position paper we propose a initial development of a robust system able to crowd event detection based on features analysis from optical flow high-frequency components. The Section 2 describes the main related works about crowd event detection. In Section 3 and 4 the proposed methodology and initial results are presented, respectively. Conclusions and further work are discussed in Section 5.

2 CROWD EVENT DETECTION

In according to (Liao et al., 2011; Ke et al., 2007), crowd event detection is an important task in public security. It has become a major issue in public places such as subway stations, banks, squares, etc. (Li et al., 2012) describes that in crowd surveillance videos, event analysis is a critical research point. In this area, it mainly contains the following issues: Firstly, it is difficult to detect and track every identity independently, because huge numbers of motion objects are seriously occluded; Secondly, it is difficult to describe the spatial relations of the objects.

In (Garate et al., 2009), when a single object is difficult to be precisely detected and tracked in the crowd scene, the common solutions directly depend on the low-level video features analysis, through motion region detection and motion feature extraction.

For crowd event detection, there are some recent works that describe automatic techniques to detect exact timing when an event (as violence actions) occurs. For example, in (Xu et al., 2014) it is proposed a Bag-of-Words (*BoW*) classification model and a fusion between *BoW* and Motion SIFT (*MoSIFT*) algorithm to improve the previous method results showed in (Wang et al., 2012). In (Esen et al., 2013), a new model of motion feature, called Motion Co-Occurrence Feature (*MCF*) and an energy model based, using information from the velocity of the moving target estimated by optical flow and entropy approach for the disorder features.

G. S. de Almeida A., de Azevedo Faria V. and de Barros Vidal F..
Crowd Event Detection in Surveillance Video - An Approach based on Optical Flow High-frequency Feature Analysis.
DOI: 10.5220/0005355306290634
In *Proceedings of the 10th International Conference on Computer Vision Theory and Applications* (VISAPP-2015), pages 629-634
ISBN: 978-989-758-091-8

3 PROPOSED METHODOLOGY

For fight detection process in surveillance videos, we performed to each frame form of the video theses steps, as described in Figure 1: First, the Horn and Schunck optical flow (Horn and Schunck, 1981) is applied; From the dense estimated optical flow, the vertical and horizontal components are evaluated; The two-dimensional Discrete Fourier Transform is used in the estimated motion field (horizontal and vertical). The Discrete Fourier Transform divides motion field into components of low and high frequencies in specific regions.; Knowing that information, we propose two different spatial slicing methods (rectangular and squared) to extract the high and low frequencies components from the estimated motion field.

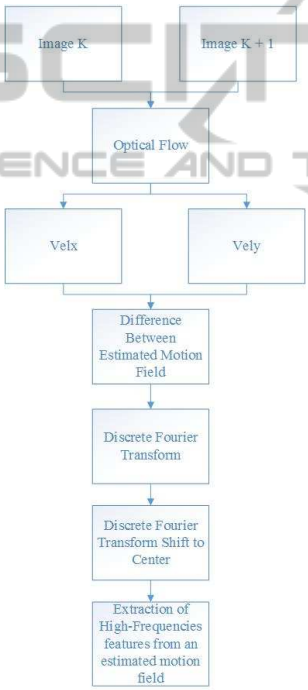


Figure 1: Proposed Methodology fluxogram.

In the following Sections the main ideas are described for each stage of the proposed methodology to be able to achieve results from the proposed approach.

3.1 Image Sequence Input

In this approach, it was used only images from a general surveillance video. A general surveillance video is captured from a fixed camera found at many cities nowadays (Kruegle, 2011). No other *a priori* information about the scenes is used and all image information processing is made in current video frame, defined in temporal index by $k + 1$ and compared with the anterior frame in time k.

3.2 Optical Flow

The optical flow approximates the image motion field by representing the apparent motion of the image brightness pattern on the image plane. In determining the optical flow, two aspects must be taken into account. One is related to the accuracy level of data concerning motion direction and intensity. The other aspect encompasses certain properties related to the computational load required for optical flow determination under minimal conditions of accuracy. The compromise between these aspects depends on the situation and the expected results. The trade-offs between efficiency and accuracy in optical flow algorithms are discussed by (Liu et al., 1998).

The methods of determining the optical flow can be divided (Barron et al., 1994) in: a) differential techniques; b) region-based matching; c) energy-based methods; and d) phase-based techniques. In this approach we considered the differential techniques. Among them, one has a particular interest; it uses spatiotemporal derivatives of the image brightness intensity defined by (Horn and Schunck, 1981).

3.2.1 Horn & Schunck Optical Flow Method

According to (Horn and Schunck, 1981), the optical flow cannot be calculated at a point in the image independently of neighboring points without introducing additional constraints. This happens because the velocity field at each image point has two components while the change in brightness at that point due to motion yields only one constraint. Before describing the method, certain conditions must be satisfied.

For convenience, it is assumed that the apparent velocity of brightness patterns can be directly identified with the movement of surfaces in the scene. This implies that, according the object surface that moves, it does not exist (or there is a little) brightness variation. This happens, for example, with objects of radial symmetry, low global contrast and high specular reflectance level. It is further assumed that the incident illumination is uniform across the surface.

Denoting $I(x,y,t)$ as the image brightness at time t of the image point (x,y). During motion, it is assumed that the brightness of a particular point is constant, that means

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

Expanding and rewriting the equation 1

$$I_x u + I_y v + I_t = 0 \quad (2)$$

where: I_x, I_y and I_t represent partial derivatives of brightness in x, y and t respectively; u and v are the $x-$ and $y-$velocity components.

Considering that, the brightness pattern can move smoothly and independently of the rest of the scene, there is a possibility to recover velocity information.

The partial derivatives of image brightness are estimated from the discrete set of image brightness measurements. To avoid problems caused by zero values for the derivatives in the spatio-temporal directions, the point of interest is located at the center of a cube formed by eight measurements as shown in figure 2 (Horn and Schunck, 1981).

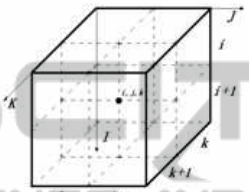


Figure 2: Estimating image partial derivates.

Each of the partial derivatives is estimated as the average of the four first differences taken over adjacent measurements

$$\begin{aligned} I_x &\approx \tfrac{1}{4}\{I_{i,j+1,k} - I_{i,j,k} + I_{i+1,j+1,k} - I_{i+1,j,k} + \\ &\quad I_{i,j+1,k+1} - I_{i,j,k+1} + I_{i+1,j+1,k+1} - I_{i+1,j,k+1}\} \\ I_y &\approx \tfrac{1}{4}\{I_{i+1,j,k} - I_{i,j,k} + I_{i+1,j+1,k} - I_{i,j+1,k} + \\ &\quad I_{i+1,j,k+1} - I_{i,j,k+1} + I_{i+1,j+1,k+1} - I_{i,j+1,k+1}\} \\ I_t &\approx \tfrac{1}{4}\{I_{i,j,k+1} - I_{i,j,k} + I_{i+1,j,k+1} - I_{i+1,j,k} + \\ &\quad I_{i,j+1,k+1} - I_{i,j+1,k} + I_{i+1,j+1,k+1} - I_{i+1,j+1,k}\} \end{aligned} \quad (3)$$

The additional constraint for the velocity calculation results from the assumption of smoothness of the velocity field.

A weighting factor α^2 is introduced to associate the error magnitude with quantization errors and noise.

The estimated values for velocities components to u_{k+1} and v_{k+1} are obtained from

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

$$v^{k+1} = \overline{v}^k - \frac{I_y\left[I_x\overline{u}^k + I_y v^k + I_t\right]}{\left(\alpha^2 + I_x^2 + I_y^2\right)} \quad (5)$$

In equations (4) and (5) $\overline{u}^k$ and $\overline{v}^k$ are the average velocities estimated from the Laplacian of the brightness pattern in iteration k, in which the neighboring pixels values are weighted with the mask shown in (Horn and Schunck, 1981).

3.3 Difference Between Estimated Motion Field

After evaluating the Horn and Schunck optical flow, two velocities components are estimated in horizontal(Vel_x) and vertical(Vel_y) direction. These components have the same size as the input image and describe the motion field in the image domain.

To crowd fight detection process, we choose to use the absolute difference between estimated flow in $k+1$ and k because in many situations, high luminance changes in the surveillance camera systems affect directly the proposed optical flow technique.

3.4 Discrete Fourier Transform

For high-frequency feature extraction we introduce information from two-dimensional Discrete Fourier Transform. The Discrete Fourier transform (DFT) (Oppenheim et al., 1999) represents any discrete function with sums of sines and cosines. Using the inverse discrete Fourier transform, it is possible to return to the original function without losing information. In image domain, it changes the domains from spatial to frequency domain. The two-dimensional Discrete Fourier Transform is described in Equation 6.

$$F(k,l) = \sum_{i=0}^{N-1}\sum_{j=0}^{N-1} f(i,j)e^{-i2\pi(\frac{ki}{N}+\frac{lj}{N})} \quad (6)$$

In according to Equation 6 whole frequencies (low and high) spectrum are evaluated after transformation and achieved from each sinusoidal components and invariant from spatial image information. The high frequencies are mostly concentrated in the borders, while the low frequencies are in the center. For improved classification purposes stage, after DFT transformation all Vel_x and Vel_y are shifted to the center (see Figure 3).

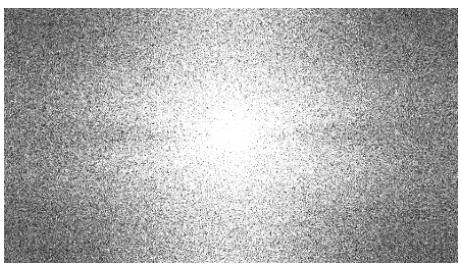


Figure 3: An example of vertical (Vel_y) magnitude components from DFT of estimated motion field.

When shifting all magnitudes to the center, it allows to invert the positions of high frequencies and low frequencies spatially. In this way, high frequencies components are in the borders of the magnitude spectrum and low frequencies are in the middle of the magnitude spectrum. The DFT shift procedure is a very important stage for the classification based on spatial feature from high-frequencies component.

3.5 Extraction of High-frequencies Features from an Estimated Motion Field

To perform crowd fighting detection in surveillance video systems, we assume an hypothesis that when in case of crowd fight occurrences many high-frequencies components from the motion field have amplitude changes (or have detectable amplitude changes) in spatial dispersion, if compared to the original spatial dispersion of theses components.

So in this stage we propose a spatial feature approach based on analysis only of the spatial dispersion, accounting in preliminary attempt changes in regions where the high-frequency is located, discarding regions that have elements of low frequencies probably are located in the spectrum.

In order to perform that, one method for spatial high-frequencies features analysis is prepared based on a division of the amplitude spectral region, is in Vel_x and Vel_y directions, in an odd number of slices, as described in Figure 4.

When great agitation (or any abnormal behavior), high-frequencies amplitude coefficients should increase and low-frequencies ones would became weaker.

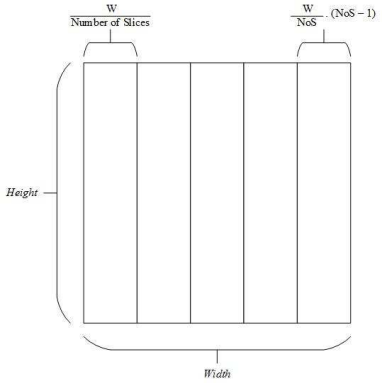


Figure 4: Vertical Slice Method.

For the vertical slice proposed method, Figure 4, we have H for image height, W that indicates image width, and NoS represents the quantity of desired slices, or the number of slices.

4 PRELIMINARY RESULTS

The proposed methodology was tested in twenty different video files, lasting thirty seconds each, in two classes: A class called violence acts (ten videos); A class called non-violence acts (ten videos). For example, a non-violence class is a situation in which there are many people crossing the street. Ten videos containing both class types, violence and non-violence acts, are used to validate the proposed methodology. The Figures 5-(a) and (b) are screenshots images from a violence and non-violence acts videos, respectively, used to build the video database. All files were captured from public videos available on internet[1].

(a) Violence acts video sequence.

(b) Non-violence acts video sequence.

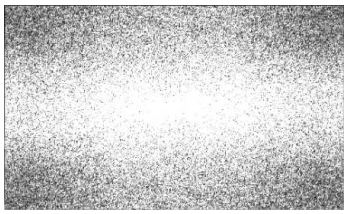
Figure 5: Screenshot images from the used video database.

In this work it is easily noticed that the proposed methodology, for a complete crowd event detection in surveillance video, requires to develop a methodology of detecting and recognition. Therefore, the preliminary results reported in this Section cover only the minimum necessary conditions in towards to build a stable feature descriptor able to be useful for the recognition process.

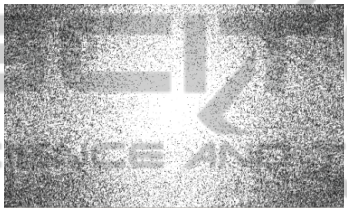
Applying the methodology proposed in the previous Section 3, we can see variations in the high-frequency components from videos with violence acts when compared to non-violence videos.

[1]The videos used in Figure 5-(a) are available in https://www.youtube.com/watch?v=dGGvCL9x6m8 and Figure 5-(b) in https://www.youtube.com/watch?v=8Q6NmFb0KuU.

(a) Original non-violence video (screenshot).



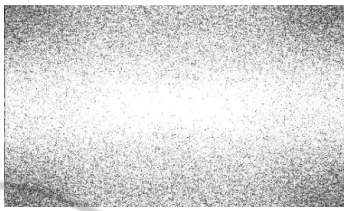
(b) Horizontal DFT amplitude components in non-violence video.



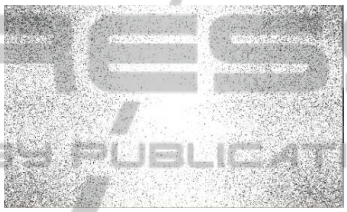
(c) Vertical DFT amplitude components in non-violence video.

Figure 6: Frequencies component extracted from a non-violence video.

(a) Original violence video (screenshot).



(b) Horizontal DFT amplitude components in violence video.



(c) Vertical DFT amplitude components in violence video.

Figure 7: Frequencies component extracted from a violence video.

In order, when comparing the two Figures 6 and 7 (violence and non-violence), occurs many amplitude variations and spatial dispersion of the high-frequency coefficients (border regions of the DFT amplitude image).

As indicated in Figures 6 and 7 that the high-frequency components from the motion field can be used to perform detection, and after the event recognition in crowds, two approaches are developed: One by evaluate the relation about number of slices are used and another by slices shape (rectangular and square).

4.1 Slice Number Tests

To evaluate the number of slices test, we calculate the mean of high-frequency energy (horizontal and vertical directions) of DFT components using 3, 5 and 11 rectangular slices, as described in Figure 8.

In Figure 8 the blue-mark is non-violence and red-mark is violence video classes. The x and y axes describe the horizontal and vertical directions, respectively.

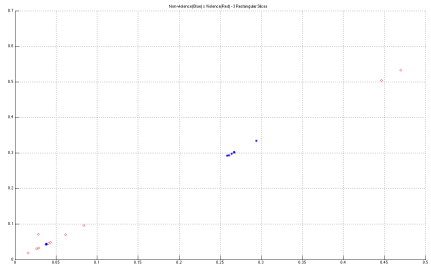
Due to the spatial distribution feature of high-frequency components of the motion field, after DFT shift process, it is the main indicative information for the event detection in crowds. The results in Figure 8 describes clearly a grouped cloud of points of the same class (non-violence), while the other cloud of points are spread in the graphic (violence class).

It is observed that in according to the slices shape are obtained (rectangular), we can build a descriptor that allows to discriminate videos between the violence and non-violence acts class in crowds. From this evaluated descriptor we have evidences to be able to classify using simple classifiers, as a Support Vector Machine or using an Artificial Neural Network.

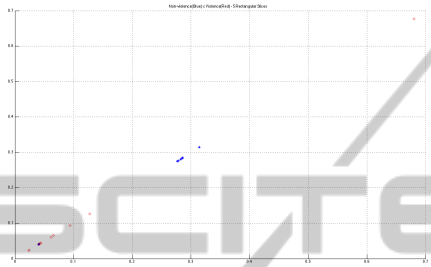
5 CONCLUSIONS AND FURTHER WORK

In this position paper we presented new strategies to analyze how the high-frequencies of estimated motion field and Discrete Fourier transform can give information about crowd event detection. Using the graphics and information showed in section 4, there are many evidences that it is possible to create a stable and reliable classifier to solve the proposed problem.

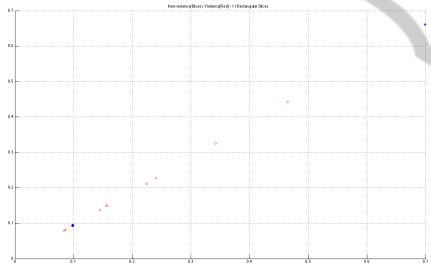
The next research steps will be given in towards of the exploration of slices shape and slice quantities



(a) 3 rectangular slices energy.



(b) 5 rectangular slices energy.



(c) 11 rectangular slices energy.

Figure 8: Slice number test - Horizontal and vertical DFT energy.

to achieve better results for the proposed descriptor. Then all efforts will be concentrate on designing a robust and efficient classifier for a full crowd event detection system.

Others works may include the implementation of the proposed strategies on a high level programming language in order to enable its operation in real time scenarios (including timing analysis) using a real videos and also perform more comparisons with the latest techniques available in crowd event detection and recognition.

REFERENCES

Barron, J. L., Fleet, D. J., and Beauchemin, S. S. (1994). Performance of optical flow techniques. In *International Journal of Computer Vision*, number 12:1, pages 43–77.

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.

Garate, C., Bilinsky, P., and Bremond, F. (2009). Crowd event recognition using hog tracker. In *Performance Evaluation of Tracking and Surveillance (PETS-Winter), 2009 Twelfth IEEE International Workshop on*, pages 1–6.

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

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

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, 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, 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.

Oppenheim, A. V., Schafer, R. W., and Buck, J. R. (1999). *Discrete-time Signal Processing (2Nd Ed.)*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

Wang, D., Zhang, Z., Wang, W., Wang, L., and Tan, T. (2012). Baseline results for violence detection in still images. In *Advanced Video and Signal-Based Surveillance (AVSS), 2012 IEEE Ninth International Conference on*, pages 54–57.

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.