REAL-TIME AUTOMATIC DETECTION OF VIOLENT-ACTS BY LOW-LEVEL COLOUR VISUAL CUES

Alessandro Mecocci, Francesco Micheli

University of Siena- Dept. Information Engineering
Via Roma n° 56, 53100 Siena (Italy)
alemecoc@alice.it, michelifrancesco@libero.it

ABSTRACT

Automatic recognition of human activities is important for the development of next generation video-surveillance systems. In this paper we address the specific problem of automatically detecting violent interpersonal acts in monocular colour video streams. Unlikely previous approaches, only little knowledge is assumed about the acquisition setup and about the content of the acquired scenes. So the proposed approach is suitable in a wide range of practical cases. Reliability and general-purpose applicability is achieved by analysing low-level features (like the spatial-temporal behaviour of coloured stains), and by measuring some warping and motion parameters. In this way it is not necessary to extract accurate target silhouettes, that is a critical task because of occlusions and overcrowding that are typical during interpersonal contacts. A suitable index called Maximum Warping Energy (MWE) has been defined to describe the localized spatial-temporal complexity of colour conformations. Our experiments show that aggressive activities give significantly higher MWE values if compared with safe actions like: walking, running, embracing or handshaking. So it is possible to distinguish violent acts from normal behaviours even in presence of many people and crowded environments.
Homography is used to improve robustness by verifying the real targets nearness. False interactions because of perspective-induced occlusions are discarded.

Index Terms— Image Processing, Violent Acts Recognition

1. INTRODUCTION

Nowadays, it is important and urgent to develop automatic systems capable of monitoring those areas where quiet and pacific behaviours must be granted (airports, schools, rail stations, etc.), or of giving immediate alarm in case of danger in unsafe and isolated places. The current use of 24 hours digital video recording, is not satisfactory, at least for two reasons: first, a huge amount of digital video (comprising normal people doing normal actions) must be recorded to grant the ability of storing an extremely short chunk of useful data. This fact rises serious privacy issues. Second, such data can be used only for "post factum" reconstructions, they are not useful to prevent crimes nor to enable just-in-time counteractions.

This is why in this paper we present a real-time automatic system for detecting aggressive and suspicious acts. The main goals are: to identify those video intervals comprising violent or aggressive activities; to store a digital record of such activities, and to rise timely alarms, both locally and remotely. The key advantages are: a huge reduction of recorded data that focus only on suspicious actions (this reduces the privacy-invasion issue), and a proactive effect that enables well-timed reactions by police or other responsible entities.
Even if the problem of automatic violent acts detection is an important one, only few literature covers this topic. In [1], person-on-person violent actions are recognized by reasoning at a single target level. First, silhouettes of each person are extracted from the image, then they are segmented and the principal parts of the human body (head, neck, shoulders, limbs) are located. Finally, visual acts are classified by using: trajectories, accelerations and orientations of these principal parts. Unfortunately, this approach works well only in presence of few people (ideally only two interacting people), so their silhouettes can be accurately outlined. Besides, a low mutual occlusion rate is required, because the system cannot understand where the principal parts are located and to which silhouettes they belong to.
In [2] and [3], video and audio cues are jointly used to detect aggressive actions in a scene. These approaches give more robust results, because occlusions are more easily tolerated. Nevertheless silhouettes segmentation is still needed, so the occlusion issues remain.
Since we want a system that works in real operative conditions (in such cases the environment is quite crowded), we propose an approach based on global chromatic features extracted from moving object in the video stream. In other words, we do not try to detect violent actions by reasoning at a single target level, because it is almost impossible to get precise silhouettes of people when the occlusion level is high, and the camera external resolution is medium to low (as is typical in the surveillance field). Moreover, during the interaction of three or more persons, the complexity and the number of possible spatial arrangements grow exponentially, rapidly reaching the intractability limit. Instead, we use multiple spatio-temporal visual cues, to reach a reasonable certainty that an aggressive action is undergoing. In this sense, our philosophy is closer to that in [2] and [3], rather than in [1].

The principal idea is that, during violent actions, the interplay between people elicits higher accelerations and disordered movements of localized subparts in the scene. Moreover, because of the human bodies closeness, various subparts tend to be hidden and unhidden very often, causing changes in the chromatic appearance of such subparts. Therefore, our approach tries to solve the aggressive-act recognition problem, by analysing the spatial and temporal behaviour of colour stains into which each moving region in the scene can be segmented. Note that we speak about moving regions, meaning that it is not needed that such regions refer to a single target or silhouette. In this way the system easily copes with the occlusion problem (each region can be caused by the fusion of two or more targets, and can be imprecise because of low-level segmentation errors).

To classify the actions, we introduce a localized complexity index, that has been called *Maximum Warping Energy* (*MWE*). The term localized means that the measurements are not referred to absolute movements (in this case a running person would have a high movement level even if the action is not aggressive), but to relative motion, that is to movements referred to the centroid position of the moving region under analysis. In this sense the approach follows the old idea that the human visual system breaks up the perceived motion into two parts: the first is the common motion of the whole configuration; and the second one is the relative motion of each element within the configuration [4].

Therefore, our methods works as follows:

1) moving blobs in the scene are extracted by background estimation techniques
2) nearby blobs (Haussdorf distance) are grouped (they are delimited by minimum bounding boxes) to form Regions of Interest (*RoI*s) that are followed from frame to frame (by centroid tracking)
3) the colour values of each blob comprising each *RoI* are clustered to get a set of significant stains that is called *RoI* Colour Framework (*RoI*-CF)
4) the movements of each colour stain within each *RoI*-CF are used to get an estimate of the *MWE*
5) the nature of the undergoing interaction is decided based on the time behaviour of the *MWE*

The paper is organized as follows: paragraph two describes the flat field assumption used to discriminate visual occlusions from real physical proximity. Paragraph three and four describe the Colour Stain Framework used to extract visual cues to detect violent actions. Paragraph five describes the method used to analyse the motion of Colour Stains, while paragraph six discusses the results and future trends.

2. FLAT FIELD ASSUMPTION

In our experiments we noted that, in some critical situations, false alarms where given even if no aggressive action was occurring. Such problem mainly arise in those cases when moving targets interact because of visual occlusions. A typical case is when three or more targets run in opposite directions and fuse together because of perspective projection onto the image plane. The difference in spatial scale of the converging targets (that are far away between one another), and the severe occlusion that takes place, tend to give high values of the *MWE* index even if it is impossible for the targets to interact (they lay in locations far apart). To fight this negative effect, we assume a *flat ground* acquisition scenarios, i.e. we assume that all the targets move over a plane. This is not a restrictive hypothesis since most of working cases show a planar floor. Also, the *flat ground* assumption allows the use of homography to find out the relative position of targets (compensation of perspective distortion).

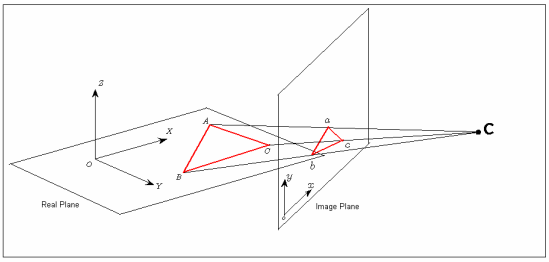

Figure 1: Example of homography. The real plane $Z = 0$ is mapped on the image plane (x, y) through the centre point C.

Using a pinhole camera model, the relation between the homogeneous coordinates of a 3D point P = [X, Y, Z, 1]T in the world coordinate system and the homogeneous coordinate of its projected point p = [x, y, 1]T in the image plane is given by the (3x4) matrix M:

$$\mathbf{p} = \mathbf{MP} \qquad (1)$$

M accounts both for the intrinsic and extrinsic parameters of the camera, and is obtained from the product of an upper triangular matrix (intrinsic part) and a rigid transformation matrix (extrinsic part). Its 12 coefficients $\{m_{ij}\}$ are obtained by using suitable calibration patches or by taking some measurements about the acquired scene [5].

It is easy to see that if the feet of moving people are assumed to stay in contact with the ground, the previous equation can be solved to recover the actual position of the target.

In practice, while two or more *RoI*s move, the contours of blobs comprising each *RoI* are analysed and used to estimate the position of the lowest points of the *RoI* (that are assumed to correspond to points on the world ground plane). Starting from these coordinates, the actual position of each *RoI* is estimated. When two or more *RoI*s are near enough on the world ground plane (say under a Euclidean-Hut whose dimensions are experimentally defined), they are considered as interacting. In the other case they are considered as occluded because of perspective (no physical interaction).

a b c

Figure 2: (a) Original frame; (b) Foreground Mask; (c) Segmented Foreground Object in **CIELab** colour coordinates.

3. BACKGROUND ESTIMATION AND *ROI* TRACKING

The blobs used to define each *RoI*, are got through a background estimation and maintenance approach. An adaptive statistical model based on three Gaussian mixtures (one for each colour channel) comprising five Gaussian each, is used in a way similar to that in [6].

Once all blobs have been extracted, nearby blobs are grouped and a minimum bounding box is used to delimit each group (*RoI*). The separation between individual blobs is measured according to the classic Hausdorff distance [7]. The distance threshold used to define a *RoI* is not critical, because the important information relates to low-level features given by colour stains movements (see below).

Once the *RoI*s have been identified, they are followed by a simple centroid tracking technique. Note that the *RoI* centroid is the centroid of the whole set of blobs comprising that *RoI*. Such simple tracking scheme can be used because we do not need a precise trajectory, since our goal is to compute an overall estimate of the global motion of each group (*RoI*) to evaluate the relative movements of each blob within the group. Besides, the centroid tracking method is very fast.

Let F_t be the frame at the time instant t, $\tilde{I}_t$ the matching binary segmented image provided by the background estimation module, and I_t the image obtained by masking F_t with $\tilde{I}_t$. Let N_t be the number of blobs (a blob is part of a segmented person, or a whole person, or a group of person, see Figure 2) in I_t. Such blobs are identified by B_i^t with $i = 1, \cdots, N_t$ and are characterized by their own centroids $\vec{\mathbf{C}}_i^t = (C_{i,x}^t, C_{i,y}^t)$.

The set of blobs is partitioned into subsets S_k so that $B_i \in S_k \;\; iff \;\; \exists B_z \,|\, d_A(B_i, B_z) < \theta \wedge B_z \in S_k$ where $d_A(\cdot,\cdot)$ is the Hausdorff distance while θ is a suitable threshold.

A *RoI* R_i is the rectangular region defined by the Minimum Bounding Box comprising all the blobs in S_j.

The centroid of each *RoI* is defined as the centroid of the blobs comprising it. The *RoI* tracker is a function $\phi(i) : \{1, \cdots, NR_t\} \rightarrow \{1, \cdots, NR_{t-1}\}$ which matches each *RoI* of a scene with a corresponding *RoI* in the next scene.

4. *ROI* COLOUR FRAMEWORK

To capture the variations because of eventual aggressive actions in the scene, we introduce the idea of Colour Framework (CF). A Colour Framework is a set of m binary images $J_{i,k}^t$ with $k = 1, \cdots, m$ associated to *RoI* R_i. Such images are obtained by segmenting the blobs of the *RoI* R_i according to a colour clustering algorithm. In our current implementation, to speed up the calculation, we use a **CIELab** colour space and the clustering is performed by simply subsampling the colour components in a uniform way. Be L the original number of quantization levels for each colour component ($L = 256$ in our case), and be Δ the subsampling factor, then the new number of quantization intervals for each component is L/Δ (we assume that L is a multiple of Δ).

The same quantization factor is applied to each component so we easily get: $m = (L/\Delta)^3$.

In our experiments good results have been obtained with $\Delta = 64$ *or* $\Delta = 128$ that gives 64 or 8 possible colour classes. The label for each colour class is assigned starting from 1 and increasing the value by 1 while scanning the **CIELab** resampled cube in a progressive order (**L** value first, from the lowest to the largest value, then the **a** value, then the **b** value).

To build the $J_{i,k}^t$ binary image we create a new image F_t^R by assigning to each pixel of F_t the label of its corresponding colour class. After that, each image $J_{i,k}^t$ is obtained by applying the following formula:

$$J_{i,k}^t(x,y) = \begin{cases} 1 & \text{if } F_t^R(x,y) = k \\ 0 & \text{if } F_t^R(x,y) \neq k \end{cases} \qquad (3)$$

where (x, y) denotes the pixel coordinates, while k is an integer running from 1 to *m*. Evidently each $J_{i,k}^t$ contains the pixels of the *RoI* whose colour belongs to the colour class k at time t. Each image $J_{i,k}^t$ comprises a certain number $n_{i,k}^t$ of blobs. Such blobs represent the stains of colour k within the *RoI* R_i, and are indicated by $R_{i,k,s}^t$ with $s = 1, \cdots, n_{i,k}^t$. Their centroids are denoted by $\vec{\mathbf{C}}_{i,k,s}^t$.

5. ANALYSIS OF COLOUR STAINS MOTION

The robustness of our approach is based on the possibility of judging the presence of violent acts by detecting the degree of variations and the temporal behaviour of some visual cues. In particular, the degree of motion of the colour stains with respect to the global motion, turned out to be a suitable descriptor. To analyse the stains motion, firstly the global motion of each *RoI* is estimated from frame F_{t-1} to frame F_t (that is match is established between a certain $\vec{\mathbf{C}}_i^t$ and a certain $\vec{\mathbf{C}}_{\phi(i)}^{t-1}$). Once

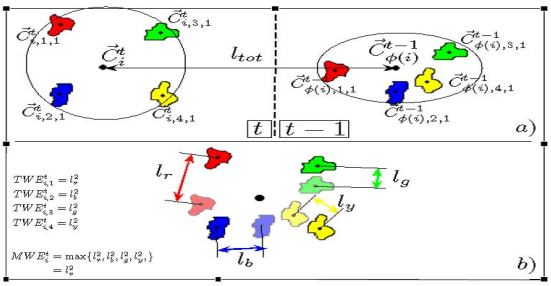

Figure 3: (a) *RoI* tracking frame by frame; (b) Colour stains warping after *RoI* motion compensation. Coloured double arrows mark the relative motion of each stain in the lapse [*t*1,*t*].

such a pairing has been found, the colour stains motion within each conformation is estimated by matching each blob in $J_{i,k}^{t}$ at time *t,* with a blob in $J_{\phi(i),k}^{t-1}$ at time *t*-1. The previous matching is repeated for the whole set of *m* images comprising each *RoI*-CF. The stains tracking is given by the function $\psi(p,q):\{1,\cdots,n_{i,k}^{t}\}\times\{1,\cdots,n_{\varphi(i),k}^{t-1}\}\rightarrow\{0,1\}$ defined as follows:

$$\psi(p,q)=\begin{cases}1 & \text{if } R_{i,k,p}^{t} \text{ match with } R_{i,k,q}^{t-1}.\\ 0 & \text{if } R_{i,k,p}^{t} \text{ don't match with } R_{i,k,q}^{t-1}.\end{cases} \quad (4)$$

To describe the spatial-temporal complexity of the colour stains conformation, we introduce a synthetic index that has been called *Total Warping relative Energy* $TWE_{i,k}^{t}$ of the stains of colour *k* belonging to *RoI* R_i at time *t*. We first define the warping energy of stains of colour *k* at time *t* as follows:

$$WE_{i,k}^{t}(p,q)=\|(\vec{\mathbf{C}}_{i,k,p}^{t}-\vec{\mathbf{C}}_{i}^{t})-(\vec{\mathbf{C}}_{\phi(i),k,q}^{t-1}-\vec{\mathbf{C}}_{\phi(i)}^{t-1})\|^{2} \quad (5)$$

Note that the previous energy refers to relative local motion of the stains within the *RoI*. At this point the *Total Warping relative Energy,* for stains of colour *k,* can be defined according to:

$$TWE_{i,k}^{t}=\sum_{p=1}^{n_{i,k}^{t}}\sum_{q=1}^{n_{\phi(i),k}^{t-1}}\psi(p,q)\cdot WE_{i,k}^{t}(p,q) \quad (6)$$

Since there are *m* different colour classes, then there will be *m* *Total Warping relative Energies.* Such energies are fused by a maximum operator to yield the *Maximum Warping Energy* MWE_{i}^{t} according to the following:

$$MWE_{i}^{t}=\max_{k\in[1,m]}\left\{TWE_{i,k}^{t}\right\} \quad (7)$$

Figure 3 shows the previous ideas applied to 4 colour stains. If MWE_{i}^{t} exceeds a predetermined threshold during a certain interval of time (currently a finite state machine is used that performs hysteresis thresholding while filtering out short and little variations), we decide that the activity within *RoI* $R_i \in I_t$ is violent.

6. RESULTS AND CONCLUSIONS

In our experiments we have used **CIELab** partitioned into 8 different colour classes. Many video sequences have been analysed related both to indoor and outdoor environments at different time during the day. Even if some events have been misclassified, the proposed approach is very promising and shows a very low false negative error level.

For example, Figure 4 shows the values of *MWE* in a difficult video where five people run towards each others from different directions. The physical interaction starts at frame 85. Note that running phase is not considered dangerous (low value of *MWE* before frame 85). The interaction becomes more and more violent culminating around frames 100 to 130. During this interval only the red line is visible, because all people in the scene are very near (actually fighting together) so a single *RoI* is detected. Besides, the MWE values are significantly high. After frame 130 people move alternately forward and backward (someone falls to the floor), so the number of blobs sharply changes in time (and other coloured lines appear, matching the different *RoI*s present in the scene). Violence disappear after frame 160 and *MWE* values drop accordingly.

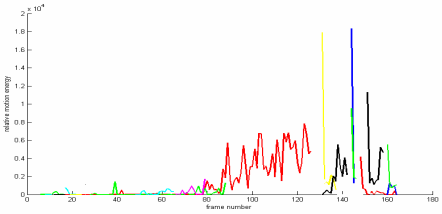

Figure 4: *Maximum Warping Energy* of some *RoI*s versus frame number. Each colour identifies a different *RoI.*

Our future works will be focused on improving the analysis of the time behavior of *MWE* through some kind of learning strategy. We will also improve the colour segmentation phase and we will integrate other sensors, like sound sensors.

7. REFERENCES

[1] Datta A., Shah M., Da Vitoria Lobo N., "Person-on-Person Violence Detection in Video Data", *icpr*, p. 10433 (ICPR'02) - Volume 1, 2002.

[2] Vasconcelos N., Lippman A, "Towards semantically meaningful feature spaces for the characterization of video content", Proc. ICIP, Volume 1, 1997, pp 25-28.

[3] Nam, J., Alghoniemy, M., "Audio-visual content-based violent scene characterization", ICIP 98, pp 353-357.

[4] Ramachandran V., S.M. Anstis, "The perception of Apparent Motion", Scientific American, pp. 80-87, June 1986.

[5] Forsyth D., Ponce J., "Computer Vision – A modern approach", Prentice Hall, 2003.

[6] Javed O., Shafique K., Shah M., "A Hierarchical Approach to Robust Background Substraction using Color and Gradient Information", Computer Vision Lab, School of Electrical Engineering and Computer Science, University of Central Florida.

[7] Preparata F. P., Shamos M. I., "Computational geometry, an introduction". Springer-Verlag, NY(1985).