BUILDING A NORMALITY SPACE OF EVENTS
A PCA Approach to Event Detection

Angelo Cenedese
Department of Engineering and Management, University of Padova, Stradella S.Nicola 3, 36100 Vicenza, Italy
angelo.cenedese@unipd.it

Ruggero Frezza
Department of Information Engineering, University of Padova, Via Gradenigo 6/B, 35131 Padova, Italy
frezza@dei.unipd.it

Enrico Campana, Giambattista Gennari, Giorgio Raccanelli
Videotec S.p.a., Via Friuli 6, 36015 Schio, Italy
enrico.campana@gmail.com, giorgio.raccanelli@videotec.com, giambattista@videotec.com

Keywords: Event detection, Principal Component Analysis (PCA), Video Analysis.

Abstract: The detection of events in video streams is a central task in the automatic vision paradigm, and spans heterogeneous fields of application from the surveillance of the environment, to the analysis of scientific data. Actually, although well captured by intuition, the definition itself of event is somewhat hazy and depending on the specific application of interest. In this work, the approach to the problem of event detection is different in nature. Instead of defining the event and searching for it within the data, a normality space of the scene is built from a chosen learning sequence The event detection algorithm works by projecting any newly acquired image onto the normality space so as to calculate a distance from it that represents the innovation of the new frame, and defines the metric for triggering an event alert.

1 INTRODUCTION

Within the paradigm of automatic computer vision, it is of paramount importance for many application that the system is able to automatically discern objects, features, events in a frame sequence. This task, which may be considered trivial when performed by human subjects, is in reality of great complexity when undertaken by a synthetic intelligence, and it represents a canonical problem in cognitive sciences. Moreover, although of general interest in the analysis of images and video streams, the detection of events, the focusing on their evolution in time, and the understanding of the scene, are key issues in the solution to the surveillance problem, which involves all these steps.

In this context, research on themes regarding surveillance and monitoring systems, has become more and more active during the last few years, supported by technologies that are becoming increasingly pervasive in our everyday life so that huge amount of data are made available (cctv systems, wireless networks, distributed sensor architectures)(CogViSys, 2007)(VSAM, 2007).

As might be expected, also scientific literature and conferences have been quite a fertile land for papers and presentations on the subject. We recall here a couple of special issues published by international journals (Collins et al., 2000) (Regazzoni et al., 2001), workshops held in conjunction with international conferences (ICML06, 2006) (ECCV06, 2006), and a recent survey paper (Hu et al., 2004), which are a good indication of what the current state of the art is.

In this work, we focus our attention on the anomaly and event detection phase, which aims at providing the Operator with metadata information of low semantic level derived from the raw data acquired from the scene. This information is the base to produce a high-level semantic description of the scene.

The first challenge in approaching the problem of event detection regards its definition and formalization. In actual fact, it is a very common task in many domains to monitor and analyze routinely collected data in search for what differs from and stands out of the *normality*, which constitutes the event or the anomaly. When considering event detection in a video stream, the event $\mathcal{E}$ is any image portion that differs from the rest of the image stream in some sense. Since mathematically the *difference in some*

sense is linked to some kind of norm, it is one main concern to try and find a procedure that highlights events as the result of norm measure. Then, the space of events is partitioned into three classes: *Normal events*, which trigger no alarm in the detection, *negligible anomalies* (e.g. natural movements, shadows) whose alarm trigger should be filtered, and *detectable anomalies* that are the true events.

2 STATE OF THE ART

In literature, several algorithms are reported that perform event detection and have been applied in different contexts.

Some of these are based on the so-called *background subtraction* techniques (Jain et al., 1977), which range from simple frame differencing, to more complex probabilistic modeling schemes. The main drawbacks of background subtraction are that the methodology is extremely sensitive to change in illumination and variation of weather conditions, hence the use in outdoor environment may be critical; also, it integrates no spatial correlation, therefore it is scantly robust with respect to spurious signals or artifacts present in the frame, which can generate false alarms. Moreover, these techniques cannot cope with natural movements of the background, changing in the background geometry, or camera oscillations, resulting again in the generation of a great amount of false alarms. A partial solution to the issues raised by the previous discussion is given by the Radial Reach Filter algorithm (Satoh et al., 2002), where the thresholding operations to discern between background and foreground in the image are performed at pixel level, but taking into account a sort of local texture. One main concern remains: The fact that the threshold values are commonly chosen by the operator, implies that the performance of the algorithm is depending on the operator experience and skills.

Starting from these basic techniques, other solutions have been developed, some employing a mixture of Gaussians to build a multimodal background model (Friedman and Russell, 1997), some resorting to the identification of shapes in the image with a dictionary of objects, which results in algorithms of non general application. Many approaches to parametric modeling have been proposed, among which hidden Markov models have become more popular in the field because they naturally enclose spatial and temporal information (Clarkson and Pentland, 2000). In these models, despite the high-quality detection provided, the main problem is the need of training the model over a really extensive sample set and the fact that often in video sequences there is no prior knowledge available to support the definition of a parametric model. A non parametric approach is considered in (Zelnik-Manor and Irani, 2001) that regards the event as a stochastic process sampled in time and space to build an empirical distribution associated to the event, in this way allowing to handle a wide range of events. A combined detection and tracking method for complex events is proposed in (Medioni et al., 2001). This work addresses the problem also in the case of moving cameras, resorting to a preprocessing procedure (Image Stabilization) to filter out the camera movement, and relying to the normal component of the optical flow field for the computation of the residual motion and the detection task.

A PCA approach has been explored by Monnet and colleagues, to model background in dynamic scenes (Monnet et al., 2003).

For a more exhaustive overview we refer the reader to the survey paper (Radke et al., 2005).

3 RATIONALE & ALGORITHM

The basic idea is to steer away from a pixel-based approach, in which the frame image is a board where to perform calculation on pixels, and move towards an information content approach, where the frame image is a container of information, regardless to the representation of the image as a matrix of pixels. As a first step in this direction we introduce the concept of *region frame*, which is a section of the image plane frame whose dimension is defined according to the information content of the image sequence in the sense that it is related directly to the size of the desired object or action to be detected. In this way the spatial correlation within each frame is enhanced. As a consequence, the size of the region assumes the role of characteristic dimension: The image is no longer analyzed pixel by pixel, but region by region, thus circumventing issues related to electronic (pixel) noise or spurious pixel signals.

The algorithm is organized through a two-step process, namely the learning phase and the (proper) detection phase. Starting from a sequence of frames (and the regions within), a *normality space* is built during the learning phase, which is supervised by an operator. Then, during the detection phase, any newly acquired frame is projected onto the normality space that has been built during the learning phase of the algorithm. The projection on the normality model yields the innovation brought in by the new frame sequence, that is the detection of unexpected events.

In the remainder of the paper, we will make use

of notions of SVD decomposition, projection over a subspace(Golub and VanLoan, 1996).

Be the image I_t at time t of $(n_y \times n_x)$ pixels, the image plane of I_t is partitioned into N regions $\{\mathcal{R}_i, i = 1,\ldots,N\}$, which can be imagined without loss of generality as regions of $(n \times m)$ pixels. We denote with $I_{t,i} := I_t|\mathcal{R}_i$ the restriction of I_t to the region $\mathcal{R}_i$. $I_{t,i}$ is then transformed into the image vector $y_{t,i}$ of size nm, by piling the columns of $I_{t,i}$ (the procedure can also be performed by rows).

Starting from a learning set of T frames $\left\{I_t^L, t = 1,\ldots,T\right\}$, and a partition of the image frame into N regions $\{\mathcal{R}_i\}$ we produce the sequence $\left\{y_{t,i}^L, t = 1,\ldots,T\right\}$ for each region of the image plane to compose the *normality matrix* $Y_i = \left[\, y_{1,i}^L \,|\, y_{2,i}^L \,|\, \ldots \,|\, y_{T,i}^L \,\right]$, whose size is $(nm \times T)$, where typically $nm > T$. The columns of Y_i report the information on region $\mathcal{R}_i$ in time, while the eigenvectors of $Y_i Y_i^\top$ are the principal components.

The SVD decomposition applied to Y_i yields $Y_i = U_i \cdot S_i \cdot V_i^\top$, with U_i column-orthonormal of dimension $(nm \times T)$, and V_i unitary square matrix of size $(T \times T)$, S_i diagonal matrix of the singular values $\sigma_{j,i}$. We remind that for the properties of the SVD algorithm the first columns of U_i corresponding to non-null singular values form a base for Y_i.The higher the value of the $\sigma_{j,i}$, the more relevant the principal component $u_{j,i}$ is to describe the scene in $\mathcal{R}_i$. Therefore, in order to retain only the main information from the learning set, a truncation at singular value $\sigma_{r,i} > 0$ is operated. It follows:

$$Y_i^r = U_i^r \cdot S_i^r \cdot (V_i^r)^\top, \tag{1}$$

where U_i^r, S_i^r, and V_i^r, are the restrictions of U_i, S_i, and V_i, and the vector space $\mathcal{U}_i := \text{range}\{Y_i^r\} = \text{span}\{u_{1,i}, u_{2,i}, \ldots, u_{r,i}\}$.

It is now to understand how to choose the truncation index r. Firstly, we observe that the principal components corresponding to null singular values are negligible, which gives a first value for r, be it $r = r_0 - 1$ (corresponding to $\sigma_{r_0} = 0$). Then, the selection of the *optimal* value is done following an intensity principle, that is by considering the first r components of Y_i retaining up to the $I_i^{\%}(r)$ of image intensity, $I_i^{\%}(r) = \frac{\sum_{j=1}^{r} \sigma_{j,i}}{\sum_{j=1}^{r_0-1} \sigma_{j,i}}$. This value follows automatically from the SVD procedure, and depends on the information content of the region of interest. Therefore, we introduce the concept of *normality* through the following

Definition 1 (Normality Model.) *Given a learning set of images $\{I_t\}$ and a set of image plane regions $\{\mathcal{R}_i\}$, interpreted as a sequence of matrices $\{Y_i^r\}$, the Normality Model is formed by:*

- *the base $\mathcal{B}_i = \{u_{1,i}, u_{2,i}, \ldots, u_{r,i}\}$ of subspace $\mathcal{U}_i$;*
- *the threshold $T_i^\sigma = \sigma_{r+1}$, stating an upper bound to the norm of the projection error of a vector in the space generated by the columns of Y_i on $\mathcal{U}_i$.*

In reality, the normality model is build from the learning sequence in a slightly more complex way, in order to ensure validation and consistency of the model. The learning sequence is subdivided by sampling into a number of subsequences: The first subsequence $Y_i^r = Y_i^{r(1)}$ is then used to build a first instance of the normality space $\mathcal{U}_i = \mathcal{U}_i^{(1)}$ and a related threshold value $T_i^\sigma = T_i^{\sigma(1)}$. The remaining subsequences $Y_i^{r(*)}$ are used to validate it: Each image (column) vector $y_{t,i}^{L*}$ of $Y_i^{r(*)}$ is projected onto the candidate $\mathcal{U}_i$

$$\mathbf{P}_{\mathcal{U}_i}(y_{t,i}^{L*}) = U_i^r \cdot (U_i^r)^\top \cdot y_{t,i}^{L*},$$

and the projection error $e_{\mathcal{U}_i}(y_{t,i}^{L*})$ is computed

$$e_{\mathcal{U}_i}(y_{t,i}^{L*}) = y_{t,i}^{L*} - \mathbf{P}_{\mathcal{U}_i}(y_{t,i}^{L*}).$$

A comparison among the norm of all the $e_{\mathcal{U}_i}(y_{t,i}^{L*})$ from the validation set and the threshold value T_i^σ, allows to understand whether the subsequence used for learning is a good learning set, or, conversely, if there is need to extend it so as to include the image vectors that are not well included in $\mathcal{U}_i$. The procedure derives a new $\hat{Y}_i^r$ from $Y_i^{r(1)}$ to build a further instance of the normality space $\mathcal{U}_i$. If the whole learning sequence is chosen appropriately, the procedure finally converges to a *validated space* $\hat{\mathcal{U}}_i$, with base $\hat{\mathcal{B}}_i$, neatly representing the essence of what happened during the learning period, and threshold value $\hat{T}_i^\sigma$.

Since $\hat{T}_i^\sigma$ is the threshold value discriminating the maximum allowed projection error during the learning phase, it is reasonable to use during the detection phase a less strict value, in order to accept in the normality domain also events that are similar to those of the learning set but have never been observed before. To this aim, the threshold value is altered by resorting to an additive or a multiplicative factor, according to the characteristics of the region of interest. In particular, if the scene is characterized by intense dynamics, the threshold value $\hat{T}_i^\sigma$ is already high, so it is advisable to use only an additive term α_i, to modify slightly $\hat{T}_i^\sigma$ in order to reduce the number of false positives without introducing further false negatives ($\check{T}_i^\sigma := \hat{T}_i^\sigma + \alpha_i$). Differently, when the scene is static, $\hat{T}_i^\sigma$ is low. The use of an additive correction law would result in an excessive generation of false negatives; conversely, the multiplicative term introduces a scaling factor in the learning threshold value ($\check{T}_i^\sigma := \alpha_i \hat{T}_i^\sigma$). In both cases, the α_i value is tuned according to experimental observations.

The detection phase of the algorithm follows basically the procedure illustrated for the validation of the normality space: Each newly acquired image frame I_t is pre-processed according to the same region decomposition of the learning phase, and reshaped into the image vectors $\{y_{t,i}\}$, where the i index spans the region set. Then, each vector $y_{t,i}$ is projected onto the correspondent $\hat{u}_i$ to determine whether and at which degree it is included in span$\{\mathcal{B}_i\}$: That is, the norm of the projection error is compared with the modified threshold value $\hat{T}_i^{\sigma}$.

An exhaustive campaign of simulations and experiments have been performed, regarding the detection of anomalous events in heterogeneous environment: In particular, we focused on the analysis of traffic flow video sequences, and the detection of people in outdoor environments in presence of wind acting on natural objects such as trees and bushes, and change of light. The algorithm has also been implemented and tested in real life situation, such as the task of monitoring the behavior of people at a fair, with consistent results.

In addition, it is remarkable how the Event Detector has proven robust to use in common situations as the detection of forgotten objects, and prohibited directions in crowd flow, because by studying the normality of the scene, the algorithm implicitly includes these events. Nonetheless, there is currently under development a dedicated tool that implements a set of rules to specifically manage these situations.

4 CONCLUSIONS

In this paper we present a novel approach to event detection by resorting to the SVD technique. The core contribution is the capability of building a vector space summarizing *what is normal in the scene* with only little supervision by the operator, who has just to choose an appropriate learning sequence. The algorithm works by projecting newly acquired images onto the so-constructed normality space, in search for innovation that is, in the case of a surveillance systems, the presence of events of some kind. Moreover, we employ an object-oriented approach by analysing regions of the image related to the characteristic size of the event of interest instead of single pixels or local textures as done in previous works.

From the preliminary results obtained so far using indoor and outdoor sequences in operating conditions, the results are quite promising, showing good robustness and high performance.

REFERENCES

Clarkson, B. and Pentland, A. (2000). Framing through peripheral perception. In *Proceedings of the IEEE international conference on image processing (ICIP 2000), Vancouver, Canada*, pages 38–41.

CogViSys (2007). http://cogvisys.iaks.uni-karlsruhe.de/. [online].

Collins, R. T., Lipton, A. J., and Kanade, T. (2000). Special issue on video surveillance. *IEEE Trans. On Pattern Analysis and Machine Intelligence*, 22(8).

ECCV06 (2006). 6th ieee international workshop on visual surveillance. In conjunction with the 9th European Conference on Computer Vision 2006, Graz, Austria.

Friedman, N. and Russell, S. (1997). Image segmentation in video sequences: A probabilistic approach. In *Annual Conference on Uncertainty in Artificial Intelligence*, volume 2, pages 175–181.

Golub, G. and VanLoan, C. (1996). *Matrix Computations*. Johns Hopkins University Press, Baltimore.

Hu, W., Tan., T., Wang, L., and Maybank, S. (2004). A survey on visual surveillance of object motion and behaviours. *IEEE Trans. On Systems, Man and Cybernetics Part C, Applications and Reviews*, 34(3):334–352.

ICML06 (2006). Machine learning algorithms for surveillance and event detection. In conjunction with the International Conference on Machine Learning 2006, Carnegie Mellon University, Pittsburgh, PA.

Jain, R., Militzer, D., and Nagel, H. (1977). Separating nonstationary from stationary scene components in a sequence of real world tv-images. In *International Joint Conference on Artificial Intelligence*, pages 612–618.

Medioni, G. G., Cohen, I., Bremond, F., Hongeng, S., and Nevatia, R. (2001). Event detection and analysis from video streams. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(8):873–889.

Monnet, A., Mittal, A., Paragios, N., and Ramesh, V. (2003). Background modeling and subtraction of dynamic scenes.

Radke, R. J., Andra, S., Al-Kofahi, O., and Roysam, B. (2005). Image change detection algorithms: A systematic survey. *IEEE Transactions on Image Processing*, 14(3):294–307.

Regazzoni, C., Ramesh, V., and Foresti, G. L. (2001). Special issue on video communication, processing and understanding for third generation surveillance systems. *Proceedings of the IEEE*, 89(10).

Satoh, Y., Tanahashi, H., Wang, C., Kaneko, S., Niwa, Y., and Yamamoto, K. (2002). Robust event detection by radial reach filter. In *16th ICPR, International Conference on Pattern Recognition*, volume 2, pages 623–626.

VSAM (2007). http://www.cs.cmu.edu/ vsam/. [online].

Zelnik-Manor, L. and Irani, M. (2001). Event-based analysis of video. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR 2001), Kauai, Hawaii, December 2001*, pages 123–130.