SUPPORT VECTOR TRACKING OF OBJECT OF INTEREST WITH AFFINE MOTION MODELS

Lionel CARMINATI, Jenny BENOIS-PINEAU

LaBRI UMR-CNRS 5800
351 Cours de la liberation
33405 Talence (France)

ABSTRACT

In this paper we describe a new way to create an object oriented video surveillance system that monitors activity in a site using Support Vector Machine as a global framework both for the detection and tracking process. This harmonious solution using the same mathematical framework is applied to a real surveillance scenarios. It consists of detection of human face in a low cluttered environment and tracking of it in order to recognize an individual. Hence we propose a set of tools to detect human faces at different scales and poses in natural video, integrating SVM classifier and a adapted multi-resolution decision making. Once detected, the second step consists in tracking faces through a video stream. Here the merit of our work is to develop a full 6-parameters affine motion model tracking with SVM. This method fits well to the variety of real situation in our tracking scenario.

1. INTRODUCTION

The goal of this work is to provide an object oriented video surveillance system that monitors activity at a site over extended periods of time by locating and tracking humans in video surveillance environment. The need for detecting and tracking humans seems to be universal, and existing solutions are numerous[1, 2] , but detecting human body parts and tracking their motion is a challenging yet essential task for modeling, recognition and interpretation of human behaviors. In particular, tracking of at least the head or face is required in video surveillance applications. In both applications two tasks have to be fulfilled: a human face has to be "grabbed" by the system and it has to be tracked. While tracking, complex behaviours such as close up or shaking and so on are possible. Therefore more complex models than a translation of the head are necessary to estimate.

Tracking of objects of interest is a well know and studied problem. Starting from various methods with Kalman Filter based trackers [3], model-based one [4] or particular trackers [5], a new trend in research represent kernel-based method [6] trackers such as SVM Tracker proposed in [7]. Assuming a support vector machine trained to recognize vehicle, authors proposed a support vector tracking -SVT- which aims to locate the "maximum score sub-image" on a tessellation of the current image. From mathematical point of view it is a very seducing approach: the mathematical model can be used for object extraction -vehicle- and tracking of it. Furthermore from application point of view, such a framework allows for tracking of specific object getting real therefore, of parasite objects. The approach is nevertheless limited. Their SVM-based motion estimator aims at estimating optical flow and to some extent can be interpreted as a translational model for a given object. In the present paper we go further, remaining on the framework of SVM classifiers, we develop our previous study on face detector [8] and propose an SVM-based tracking of detected face with a full affine motion model. The global system is divided into two processes: a Support Vector Machine face detector and SVM tracker using the same set of parameters -kernel, training set, etc. Face detection step is preceded by a motion based detection, via Gaussian mixture, which performs detection of new object in the scene and provides reduction of number of samples to test by locating human faces only in motion area of the image.

We are working with surveillance scenario when a human is entering a room and problem is to localize his face in order to track and recognize it. In this case he/she can be approaching the camera or camera may zoom on the object. In both cases, we suppose that the camera motion is compensated by estimating global camera model as we do it in [9].

The paper is organized as follows: Section 2 presents the detection process of our approach, introducing Support Vector Machine theory applied to the problem of face detection. In Section 3, a tracking of object of interest is developed using SVM theory basis. Results obtained on typical video surveillance are described in Section 4 and conclusion and perspectives are given in Section 5.

2. DETECTION OF HUMAN FACES

The first stage of surveillance problem is to extract initial targets such as human faces from a video stream. Instead of searching for the whole frame, we suggest to detect objects of interest only in motion area assuming the problem can be described into two classes problem: "background" and "foreground objects". In [8], we proposed a detection method for moving areas based on modelling of luminance signal as a mixture of Gaussian [1] with a temporal regularization. We use the same method here to detect areas with proper motion and scan for a face inside.

The problem of face detection can be identified as a detection and localization task. We resort here to Support Vector Machines (S.V.M.) because they offer state-of-the-art capabilities in the context of supervised detection. Their effectiveness resides in part in the manner they address the fundamental issue of generalization[10].

Assuming we want to estimate a function $f : R^n \longrightarrow \{\pm 1\}$ using input-output training data $(x_1, y_1), ..., (x_l, y_l) \in R^n \times \{\pm 1\}$ such as f will correctly classify unseen examples (x, y), i.e., $f(x) = y$ for examples (x, y) that were generated from the same underlying probability distribution $P(x, y)$ as the training data. In such formulation it means that the data have to be separated into two classes. If we put no restriction on the class of functions that we choose our estimate f from, then, even if a function performs well on the training data, e.g. by satisfying $f(x_i) = y_i$ for all $i = 1, ..., l$ it is not guaranteed to generalize well to unseen examples. In classical learning theory, function f is chosen to minimize the training error or *empirical risk*,

$$R_{emp}(f) = \frac{1}{2l} \sum_{i=1}^{l} |y_i - f(x_i)| \tag{1}$$

Nevertheless the optimality of choice with regard to minimum of empirical risk does not imply a small test error, called *risk* $R(f)$

$$R(f) = \int \frac{1}{2} |y - f(x)| dP(x, y) \tag{2}$$

Assuming that we have a training data set of labeled examples linearly separable, we wish to determine, among the infinite number of linear classifiers that separate the data, which one will have the smallest generalization error 2. Support Vector Machine theory[10] shows that only few points of interest are used to determine hyperplane of classification. These data points are called *support vectors* which are the only ones relevant for the solution of the problem. In fact, all other data points can be deleted from the data set and the same solution would be obtained. Applied to the problem of face detection in video frames, the problem can be formulated as follows. Let us consider fixed size windows selected from image signal containing or not a human face. The training step will consist in construction of a classification surface from labeled examples. The generalization step consists in classifying windows from input images into two classes: "face" and "no face". Taking into account a large variability of image content of real scenes inside non-face windows, we followed the authors of [11] and chose a polynomial second order classifier.

Training is realized by selecting and labelling $N * N$ windows both on faces and on the natural background in full resolution video frames. The generalization step consists in scanning the input frame by $N * N$ retina and classifying it into "face" or "non-face" class by trained classifier. In order to enhance the robustness of the classification with regard to the face area size, we propose a multi-resolution scheme. Face detection scanning is performed at several scales -from top to the bottom of the multi -resolution pyramid- and all "face" labeled local windows are cross-checked. If several "face" labeled retinas are superimposed in the scaled images, the priority is given to the upper scale of the pyramid, ensuring the best covering of the face area.

3. SUPPORT VECTOR TRACKING

The goal of the whole system is to characterise the behaviour of objects of interest (human faces) in video scenes. Thus once the objects of interest is detected, the problem of tracking arises. A classifier trained to recognize a particular class of object can be used for tracking such an object by applying the classifier on each neighbourhood in a tessellation over some configuration space such as image translation, rotation and zoom. We followed authors of [7] where they seek to mitigate this. They also describe a complete method for tracking vehicle using a forward looking camera mounted on a moving vehicle to detect and track the rear-end of moving vehicle. Assuming only translational motion, the model they employed is sufficient for their purpose but cannot be employed in this way in real video surveillance environment when face is zooming or rotating for instance.

In this paper we develop a method of Support Vector Tracking which aims to follow object of interest undergoing affine transformations described by a complete -6 parameters- affine motion model. We so inject a Support Vector Machine into a Optical Flow estimator and we perturb the classification function f with respect to affine motion in image plane.

SVM theory demonstrates [10] that binary classification can be performed according to the sign of the classification function below:

$$\sum_{j=1}^{N} y_j \alpha_j k(I, x_j) + b \tag{3}$$

where x_j are the N support vectors obtained after training step, y_j their sign and α_j are their distance from the hyperplane. $k(I, x_j)$ is the kernel which computes a generalised inner product between I, the image region we wish to test, and the support vectors.

Now let us consider I_{init} the initial guess of position of moving object in the image. Assuming target object position is close to the initial position and following usual development of I_{final} in Taylor Series up to first order, we get $I_{final} = I_{init} + uI_x + vI_y$ where I_x and I_y are gradient on x and y respectively. In this work, we will assume so the optical flow (u, v) follows a 6-parameters affine motion model defined by $u = (a_0 + a_1 x + a_2 y)I_x$ and $v = (a_3 + a_4 x + a_5 y)I_y$.

By definition SVM score of I_{final} would be higher than I_{init} score, so we get

$$\sum_{j=1}^{N} y_j \alpha_j k(I_{final}, x_j) = max\{I | \sum_{j=1}^{N} y_j \alpha_j k(I, x_j)\}$$

where I correspond to all images to test around the initial position. If we plug definition of I_{final} in equation 3 we get

$$\sum_{j=1}^{N} y_j \alpha_j k(I_{init} + uI_x + vI_y, x_j)$$

which we have to maximize. Assuming a second order polynomial kernel given by $k(x, x_j) = (x \,.\, x_j)^2$ we introduce the energy function E we have to maximize.

$$\begin{aligned} E(a_0, .., a_5) &= \sum_{j=1}^{N} y_j \alpha_j k(I_{init} + uI_x + vI_y, x_j) \\ &= \sum_{j=1}^{N} y_j \alpha_j ((I_{init} + uI_x + vI_y) \,.\, x_j)^2 \end{aligned}$$

Deriving E with respect to each a_i and simplifying, we get the following equations

$$\begin{aligned} \frac{\partial E}{\partial a_0} &= 2\sum_{j=1}^{N} y_j \alpha_j (I_x \,.\, x_j)((I_{init} + uI_x + vI_y) \,.\, x_j) \\ \frac{\partial E}{\partial a_1} &= 2\sum_{j=1}^{N} y_j \alpha_j (xI_x \,.\, x_j)((I_{init} + uI_x + vI_y) \,.\, x_j) \\ \frac{\partial E}{\partial a_2} &= 2\sum_{j=1}^{N} y_j \alpha_j (yI_x \,.\, x_j)((I_{init} + uI_x + vI_y) \,.\, x_j) \\ \frac{\partial E}{\partial a_3} &= 2\sum_{j=1}^{N} y_j \alpha_j (I_y \,.\, x_j)((I_{init} + uI_x + vI_y) \,.\, x_j) \\ \frac{\partial E}{\partial a_4} &= 2\sum_{j=1}^{N} y_j \alpha_j (xI_y \,.\, x_j)((I_{init} + uI_x + vI_y) \,.\, x_j) \\ \frac{\partial E}{\partial a_5} &= 2\sum_{j=1}^{N} y_j \alpha_j (yI_y \,.\, x_j)((I_{init} + uI_x + vI_y) \,.\, x_j) \end{aligned}$$

such as $\forall i \in \{0, .., 5\}$ $\frac{\partial E}{\partial a_i} = 0$. In order to simplify the notation, considering matrices $A_{ij,0\leq i,j\leq 5}$, $B_{ij,0\leq i\leq 5,j=1}$ and $C_{ij,0\leq i\leq 5,j=1}$, the equation above, after development, can be described by the matrix form $CA = B$ with

$$\begin{aligned} C_{00} &= \sum_{j=1}^{N} y_j \alpha_j (I_x.x_j)^2 (I_x.x_j)^2 \\ C_{01} &= \sum_{j=1}^{N} y_j \alpha_j (I_x.x_j)^2 (xI_x.x_j)^2 \\ C_{02} &= \sum_{j=1}^{N} y_j \alpha_j (I_x.x_j)^2 (yI_x.x_j)^2 \\ C_{03} &= \sum_{j=1}^{N} y_j \alpha_j (I_x.x_j)^2 (I_y.x_j)^2 \\ C_{04} &= \sum_{j=1}^{N} y_j \alpha_j (I_x.x_j)^2 (xI_y.x_j)^2 \\ C_{05} &= \sum_{j=1}^{N} y_j \alpha_j (I_x.x_j)^2 (yI_y.x_j)^2 \\ C_{10} &= C_{01} \\ \vdots &\quad \vdots \quad\quad \vdots \\ C_{50} &= C_{05} \\ C_{51} &= C_{15} \\ C_{52} &= C_{25} \\ C_{53} &= C_{35} \\ C_{54} &= C_{45} \\ C_{55} &= \sum_{j=1}^{N} y_j \alpha_j (yI_y.x_j)^2 (yI_y.x_j)^2 \end{aligned} \quad (4)$$

,

$$A = (a_0, a_1, a_2, a_3, a_4, a_5)^T$$

and

$$B = \begin{pmatrix} -\sum_{j=1}^{N} y_j \alpha_j (I_x \,.\, x_j)(I \,.\, x_j) \\ -\sum_{j=1}^{N} y_j \alpha_j (xI_x \,.\, x_j)(I \,.\, x_j) \\ -\sum_{j=1}^{N} y_j \alpha_j (yI_x \,.\, x_j)(I \,.\, x_j) \\ -\sum_{j=1}^{N} y_j \alpha_j (I_y \,.\, x_j)(I \,.\, x_j) \\ -\sum_{j=1}^{N} y_j \alpha_j (xI_y \,.\, x_j)(I \,.\, x_j) \\ -\sum_{j=1}^{N} y_j \alpha_j (yI_y \,.\, x_j)(I \,.\, x_j) \end{pmatrix}$$

To estimate the parameter vector A we developed the following iterative process: During the first iteration at time t, we consider the initial sub-image I given by the position of previous face detection at time $t-1$. For each pixel of I, we estimate the coefficients a_i calculating $A = C^{-}1B$ which define a displacement vector (u, v) for each pixel I. Then we compute the energy function defined and then we repeat th process until energy function E becomes stable. In order to compute I at non integer positions, a bilinear interpolation is used.

4. RESULTS

The method proposed was tested both for detection and tracking performance. First, the classifier implementation[12] was trained with a database of 30*30 labelled face or non-face patterns. An upper bound penalty here was of 100 and a training set of 6977 cropped images (2429 faces and 4 548 non-faces) and 25 another faces extracted from video surveillance video were used. To test the run-time system we used the test database provided by the M.I.T. This set contains 24045 30*30 images (472 faces and 23 573 non-faces). The Cross-checking strategy was tested on 15 images extracted from two typical video surveillance sequences, acquired by a low-cost commercial camera in QCIF resolution, called "Jenny" and "Lionel". We compare the result of this strategy with results given by a "full search" approach which aims to locate face overall image at each level of the pyramid.

Strategy	**Full-Search**	**Cross-Checking**
Number of test	2746	1125
Recall Rate	85.1%	83.8%
Precision Rate	83.8%	95.7%
Global Time	1822ms	746ms

The table above illustrates full-search and cross checking strategies results. The global time given correspond to the total time used for classification. We notice that recall and precision are close the sames. Cross-checking improves precision rate with a small decrease of recall rate comparing to "full-search" method. The main advantage of cross-checking is time consuming improvement. This strategy divide by 3 the need time.

The behaviour of the tracking system in the case where the object of interest is detected is illustrated in Figure 1. Here when a strong motion occurs as depicted in the bottom images, the system fails to track face but in association with face detector, we can re-detect it.

The overall performance of the system in terms of computational time strongly depends on the limitation of search area for the detection algorithm, the quality of tracking and number of Support Vector obtained after training step. At present, in our experiments the face classification time for chosen window size is about 0.15 ms on a Pentium 4. With regard to Gaussian mixture detection, we choose K from 2 to 4 for computational reason. Obviously Gaussian mixture performance strongly depends on the size of the image we consider but motion and face detection is then performed within 10ms in CIF resolution. Actually tracking system is also time consuming but still near real-time due to high computational cost associated to number of support vector. We focus our attention to reduce their number significantly.

5. CONCLUSION AND PERSPECTIVES

In this work we proposed a full scheme for detection of ojects of interest such as human face and tracking it using the same mathematical framework: The Support Vector Machine. Compared to the state of the art, we formulated the problem and proposed a solution for estimation of complete affine first order model. The latter ensures the use of proposed trackers in various video surveillance, non machine interaction and indexing scenarios. A cross checking multi-resolution decision strategy allows better precision rate of decision step. Finally an intelligent cooperation between face and motion detections allows a near real-time performance of the system on low cost commercial PCs. Results obtained are encouraging and make us follow this work.

6. ACKNOWLEDGEMENTS

This research is funded by Centre National de Recherche Scientifique (CNRS) national grant together with VisualPix SA.

7. REFERENCES

[1] W. Grimson, C. Stauffer, R. Romano, and L. Lee, "Using adaptive tracking to classify and monitor activities in a site," *IEEE CVPR 1998* , pp. 22–29, 1998.

[2] T. Kanade, R. Collins, A. Lipton, P. Burt, and L. Wixson, "Advances in cooperatives multi-sensor video surveillance," in *Proceedings of DARPA Image Understanding Workshop*, **1**, pp. 3–24, 1998.

[3] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Transactions of the AMSE, Part D, Journal of Basic Engineering* **82**, pp. 35–45, 1960.

[4] J. M. Rehg and T. Kanade, "Model-based tracking of self-occluding articulated objects.," in *ICCV*, pp. 612–617, 1995.

[5] M. Isard and A. Blake, "Condensation conditional density propagation for visual tracking," *International Journal of Computer Vision* **29(1)**, pp. 5–28, 1998.

[6] D. Comaniciu, V. Ramesh, and P. Meer, "Kernel-based object tracking.," *IEEE Trans. Pattern Anal. Mach. Intell.* **25**(5), pp. 564–575, 2003.

[7] S. Avidan, "Support vector tracking.," *IEEE Trans. Pattern Anal. Mach. Intell.* **26**(8), pp. 1064–1072, 2004.

[8] L. Carminati, J.Benois Pineau, and M. Gelgon, "Human detection and tracking fro video surveillance applications in low density environment," *SPIE VCIP'2003 SPIE 0277 -786X* **5150**, pp. 51–60, 2003.

[9] M. Durik and J. Benois Pineau, "Robust motion characterisation for video indexing based on mpeg2 optical flow," *Content Based Multimedia Indexing* , September 2001.

[10] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer-Verlag, New York, ISBN 0-387-94559-8, 1995.

[11] E. Osuna, R. Freund, and F. Girosi, "Training support vector machines: an application to face detection.," in *IEEE CVPR*, pp. 130–136, 1997.

[12] M.I.T. SvmFu Version 3 Software developed by the Center for Biological and Computational Learning du M.I.T. http://five-percent-nation.mit.edu/SvmFu, 2001.

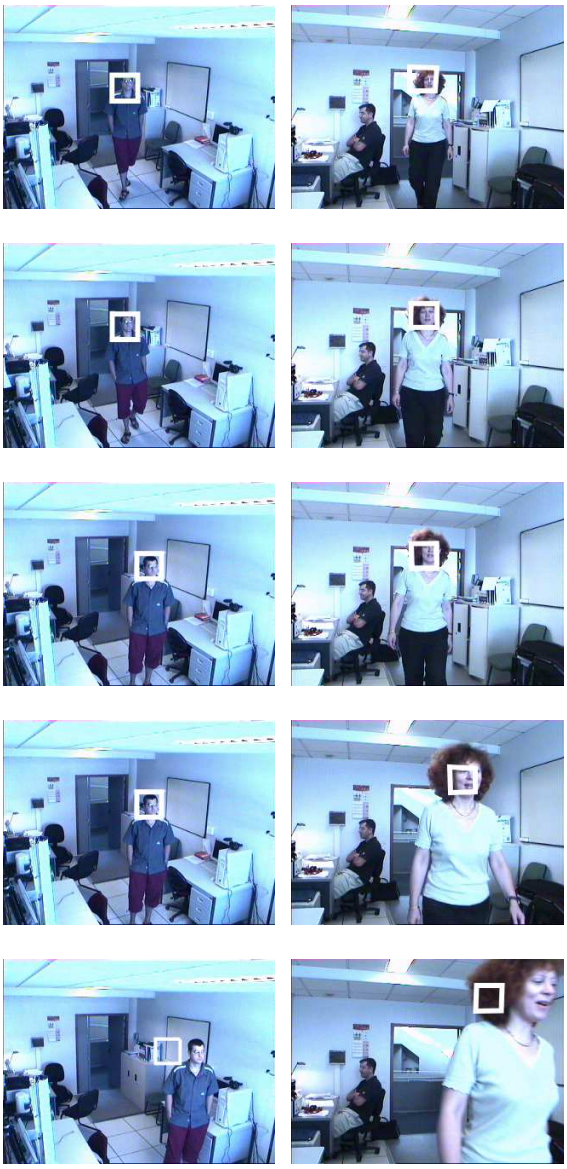

Fig. 1. Representation of detection and tracking techniques on two typical video surveillance. The first images on top give examples of Support Vector Machine face detection on real video-surveillance at rate 25fps on resolution CIF. The next images on left correspond to frames 22, 50, 80 and 150 acquired with a typical low cost camera. Right images correspond to frames 20, 40, 140 and 200. The bottom images illustrate when tracking process failed.