Robust Tracking and Object Classification towards Automated Video Surveillance

Jose-Luis Landabaso[1], Li-Qun Xu[2], Montse Pardas[1]

[1] Technical University of Catalunya, Barcelona, Spain
[2] BT Exact, Adastral Park, Ipswich, UK

Abstract. This paper addresses some of the key issues in computer vision that contribute to the technical advances and system realisation for automated visual events analysis in video surveillance applications. The objectives are to robustly segment and track multiple objects in the cluttered dynamic scene, and, if required, further classify the objects into several categories, e.g. single person, group of people or car. There are two major contributions being presented. First, an effective scheme is proposed for accurate cast shadows / highlights removal with error corrections based on conditional morphological reconstruction. Second, a temporal template-based robust tracking scheme is introduced, taking account of multiple characteristic features (velocity, shape, colour) of a 2D object appearance simultaneously in accordance with their respective variances. Extensive experiments on video sequences of variety real-world scenarios are conducted, showing very promising tracking performance, and the results on *PETS2001* sequences are illustrated.

1 Introduction

Accurate and robust segmentation and tracking of multiple moving objects in dynamic and cluttered visual scenes is one of the major challenges in computer vision. It is particularly desirable in the video surveillance field where an automated system allows fast and efficient access to unforeseen events that need to be attended by security guards or law enforcement officers as well as enables tagging and indexing interesting scene activities / statistics in a video database for future retrieval on demand. In addition, such systems are the building blocks of higher-level intelligent vision-based or assisted information analysis and management systems with a view to understanding the complex actions, interactions, and abnormal behaviours of objects in the scene.

Vision-based surveillance systems can be classified in several different ways, considering the environment in which they are designed to operate i.e. indoor, outdoor or airborne; the type and number of sensors; the objects and level of details to be tracked. In this paper our focus is on processing videos captured by a single fixed outdoor CCTV camera overlooking areas where there are a variety of vehicle and/or people activities.

There are typically a number of challenges associated with the chosen scenario in the realistic surveillance applications environment: natural cluttered background, repetitive background, illumination changes, occlusions, objects entries and exits, or shadows and highlights.

Over the recent years there have been extensive research activities in proposing new ideas, solutions and systems for robust object tracking to address the above situations [1]. Most of them adopt the 'background subtraction' as a common approach to detecting foreground moving pixels, whereby the background scene structures are modelled pixel-wise by various statistically-based learning techniques on features such as intensities, colours, edges, textures etc. The models employed include parallel unimodal Gaussians [2], mixture of Gaussian [3], nonparametric Kernel density estimation [4], or simply temporal median filtering [5]. A connected component analysis (*CCA*) [6] is then followed to cluster and label the foreground pixels into meaningful object blobs, from which some inherent appearance and motion features can be extracted. Finally, there is a blob-based tracking process aiming to find persistent blob correspondences between consecutive frames. In addition, most application systems will also deal with the issues of object categorisation or identification (and possibly detailed parts analysis) either before [7] or after [5] the tracking is established.

Regarding the matching method and metric, the heterogeneous nature of the features extracted from the *2D* blobs has motivated some researchers to use only a few features, *e.g.* the size and velocity in [8] for motion correspondence, and the size and position with Kalman predictors in [3]. Others using more features decide to conduct the matching in a hierarchical manner, for instance, in the order of centroid, shape and colour as discussed in [5]. Note that if some domain knowledge is known, *e.g.*, the type of an object to be tracked being a single person, then more complex dynamic appearance models of the silhouettes can be used [7]. Also, in [4] special probabilistic object appearance models have been used to detect and track individual persons who start to form a group and occlude each other [9].

In this paper we describe a robust multi-object tracking and classification system in which several novel ideas are introduced. These include the use of false foreground pixels suppression; the cast shadows / highlights removal; and the matching process using the scaled Euclidean distance metric in which a number of features characterising a foreground object are used simultaneously, taking into account the scaling and variance of each of the features. The method is not only very accurate, but also allows an easier inclusion of other extracted features, if necessary, leaving room for future enhancement. The system also further incorporates a classification module to classify each persistently tracked object, based on the analysis of local repetitive motion changes within the blob representation over a period of time. Figure 1 depicts schematically the block diagram of our object tracking and classification system.

The paper is structured as follows. In the next section the techniques for pixel-domain analysis leading to the segmented foreground object blobs are described. Section 3 is devoted to discussion on issues concerning robust object tracking, including the use of temporal template; the matching procedure, and the object entries and exits. Section 4 describes the object classification approach adopted.

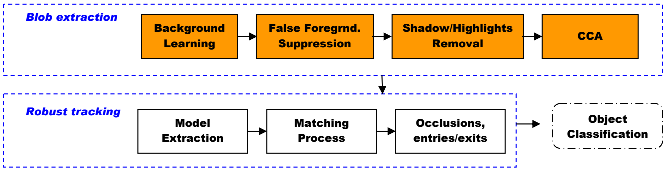

Fig. 1. The system block diagram showing the chain of functional modules.

Section 5 illustrates the experimental evaluations of the system. And finally, the paper concludes in Section 6.

2 Moving Objects Segmentation

The first issue to solve in the chain of the proposed surveillance system is the segmentation of those image pixels that do not belong to the background scene. As in [8] the adaptive background subtraction method proposed by Stauffer and Grimson [3] is adopted. A mixture of K Gaussian distributions is used to model RGB colour changes, at each pixel location, in the imaging scene over the time. With each incoming frame the Gaussian distributions are updated, and then used to determine which pixels are most likely to result from a background process. This model allows a proper representation of the background scene undergoing slow lighting and scene changes as well as momentary variations such as swaying trees / flags with winds.

The foreground pixels thus obtained, however, are not exempt from false detections due to noise in the background and camera jitters. A false-foreground pixels suppression procedure is introduced to alleviate this problem. Basically, when a pixel is initially classified as a foreground pixel, its 8-connected neighbouring pixels models are examined. If the majority of these models, when applied to this pixel, agree that its a background pixel, then its considered as a false detection and removed from foreground.

Once the foreground objects pixels have been identified, a further scheme is applied to find out if some of these foreground pixels correspond to areas likely to be cast shadows or specular reflections. The working mechanism of this novel scheme is the following: As the first step, a simplified version of the technique discussed in [10] is used to evaluate the variability in both brightness and colour distortion between the foreground pixels and the adaptive background, and possible shadows and highlights are detected. It was observed though that this procedure is less effective in cases that the objects of interest have similar colours to that of presumed shadows. To correct this, an assertion process comparing the gradient / textures similarities of the foreground pixels and corresponding background is incorporated. These processing steps effectively removing cast shadows also invariably delete some object pixels and distort object shapes. Therefore, a morphology-based conditional region growing algorithm is employed to recon-

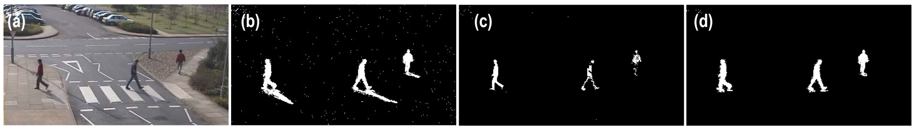

Fig. 2. (a) A snapshot of a surveillance video sequence, the cast shadows from pedestrians are strong and large; (b) the result of initial foreground pixels segmentation, the moving shadows being included; (c) The "skeleton" image obtained after the shadow removing processing; and (d) the final reconstructed objects with error corrections.

struct the objects shapes. This novel approach gives favourable results compared to the current state-of-the-art to suppress shadows / highlights. Figure 2 illustrates an example processing result.

3 Robust Objects Tracking

After the cast shadows / highlights removal procedure, a classical 8-connectivity connected component analysis is performed to link all the pixels presumably belonging to individual objects into respective blobs. The blobs are temporally tracked throughout their movements within the scene by means of temporal templates.

3.1 Temporal Templates

Each object of interest in the scene is modelled by a temporal template of persistent characteristic features. In the current studies, a set of five significant features are used describing the *velocity* $\boldsymbol{v} = (v_x, v_y)$ at its centroid (p_x, p_y); the *size*, or number of pixels, contained (s); the *ratio* (r) of the major-axis vs. minor-axis of the best-fit ellipse of the blob [11]; the *orientation* of the major-axis of the ellipse (θ); and the *dominant colour* representation $(\boldsymbol{c_p})$, using the principal eigenvector of the aggregated pixels' colour covariance matrix of the blob.

Therefore at time t, we have, for each object l centred at (p_{lx}, p_{ly}), a template of features $M_l(t) = (\boldsymbol{v_l}, s_l, r_l, \theta_l, 1_l(\boldsymbol{c_p}))$

There are two points that need special clarification as follows:

a) Prior to matching the template l with a candidate blob k in frame $t+1$, centred at $({p'}_{kx}, {p'}_{ky})$ with a feature vector $B_k(t{+}1) = (\boldsymbol{v'_k}, {s'}_k, {r'}_k, {\theta'}_k, d_k(\boldsymbol{c'_p}))$, Kalman filters are used to update the template by predicting, respectively, its new velocity, size, aspect ratio, orientation in $\hat{M}_l(t+1)$. The velocity of the candidate blob k is calculated as $\boldsymbol{v'_k} = ({p'}_{kx}, {p'}_{ky})^T - (p_{lx}, p_{ly})^T$

b) Instead of $\boldsymbol{c_p}$, we use $1_l(\boldsymbol{c_p})$, or the value of 1.0, to denote the dominant colour of the template, and $d_k(\boldsymbol{c'_p})$, to represent the colour similarity between the template l and the candidate blob k: $d_k(\boldsymbol{c'_p}) = \frac{\boldsymbol{c_p} \cdot \boldsymbol{c'_p}}{\|\boldsymbol{c_p}\| \|\boldsymbol{c'_p}\|}$

It is only after a match (in section 3.2) is found that the template's dominant colour is replaced with that of the matched candidate.

The mean $\overline{M_l}(t)$ and variance $V_l(t)$ vector of such a template are updated when a candidate blob k in frame $t+1$ is found to match with it. And they are computed using the latest corresponding L blobs that the object has matched, or a temporal window of L frames (*e.g.*, $L = 50$). With regard to individual Kalman filters $KF_l(t)$, they are updated only by feeding with the corresponding feature value of the matched blob.

3.2 Matching Procedure

We choose to use a parallel matching strategy in preference to the serial matching one such as that used in [5]. The main issue now is the use of a proper distance metric that best suits the problem under study. Obviously, some features are more persistent for an object while others may be more susceptible to noise. Also, different features normally assume values in different ranges with different variances. Euclidean distance does not account for these factors as it will allow dimensions with larger scales and variances to dominate the distance measure.

One way to tackle this problem is to use the Mahalanobis distance metric, which takes into account not only the scaling and variance of a feature, but also the variation of other features based on the covariance matrix. Thus, if there are correlated features, their contribution is weighted appropriately.

However, with high-dimensional data, the covariance matrix can become non-invertible. Furthermore, matrix inversion is a computationally expensive process, not suitable for real-time operation. So, in the current work a scaled Euclidean distance, shown in (1), between the template l and a candidate blob k is adopted, assuming a diagonal covariance matrix. For a heterogeneous data set, this is a reasonable distance definition.

$$D(l,k) = \sqrt{\sum_{i=1}^{N} (x_{li} - y_{ki})^2/\sigma_{li}^2} \tag{1}$$

where the index i runs through all the features of the template, and σ_{li}^2 is the corresponding component of the variance vector $V_l(t)$. Note especially that for the colour component, $x_{li} = 1.0$ is assumed for the object l, and $y_{ki} = d_k(\boldsymbol{c'_p})$ for the candidate blob k.

3.3 Occlusions Handling

In the current approach, no use is made of any special heuristics on the areas where objects enter/exit into/from the scene. Objects may just appear or disappear in the middle of the image, and, hence, positional rules are not necessary.

To handle occlusions, the use of heuristics is essential. Every time an object has failed to find a match with a candidate blob, a test on occlusion is carried out. If the object's bounding box is overlapped with some other object's bounding box, then both objects are marked as ‘occluded’. This process is repeated until all objects are either matched, marked as occluded, or removed after missing for MAX_LOST frames.

As discussed before, during the possible occlusion period, the object template of features are updated using the average of the last 50 correct predictions to obtain a long-term tendency prediction. Occluded objects are better tracked using the averaged template predictions. In doing so, small erratic movements in the last few frames are filtered out. Predictions of positions are constrained within the occlusion blob.

Once the objects are tracked, the classification challenges can be addressed.

4 Object Classification

The goal is to classify each persistently tracked object as being a single person, a group of people or a vehicle. The procedure employed is based on evaluating internal motion within the tracked object blob over T consecutive frames, which is similar to that discussed in [8].

First, a translation and scale compensations of the object over time is needed. Translation is done by using a bounding box centred on the tracked object. The bounding box is then resized to a standard size to compensate for scale variations.

Second, the internal motion is computed as the blob changes in consecutive frames using the XOR operator $D_t(i,j) = B_t(i,j) \oplus B_{t-1}(i,j)$ followed by accumulating these changes over the last T frames: $A(i,j) = \sum_{\tau=u}^{T} D_{t-\tau}(i,j)$.

Finally, all $A(i,j)$ corresponding to the pixels in the top and bottom section of the object are added together (2), considering that the only repetitive movement observed for walking persons are in the top (arms), and bottom (legs) sections.

$$\overline{A} = \frac{\sum_{i=0}^{X} \left(\sum_{j=0}^{Y/3} A(i,j) + \sum_{j=2Y/3}^{Y} A(i,j) \right)}{X \cdot Y} \qquad (2)$$

where X and Y are the width and height of the scale-compensated object blob.

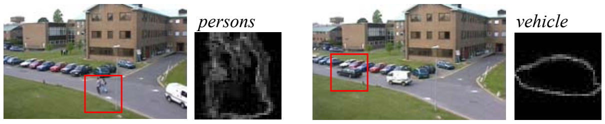

Fig. 3. $A(i,j)$ for a group of persons and a car. $A(i,j)$ is depicted in grey scale with white values denoting higher motion. The left image shows much higher internal repetitive movements, especially in the upper and bottom sections.

At this point, a threshold can be defined. An object is identified as non-rigid moving object such as a person or a group of people if its value is above the threshold; otherwise it is classified as a vehicle. The choice of the threshold depends on T. In our tests a threshold of 1 proved to classify most of the objects correctly when using a value of $T = 50$ (2 secs. at 25 fps).

5 Experimental Results

The system has been extensively evaluated in several scenarios and conditions, with, among others, the benchmarking video sequences provided by *PETS 2001*. Original testing images are compressed in *JPEG* format, and we have used subsampled versions of size 384×288. Apart from the *JPEG* compression artefacts, the sequences also contain a few other difficulties, including thin structures, reflections, illumination changes, swaying leaves in trees and window reflections in outdoor scenarios, shadows, etc. The system has dealt with all these problems successfully, and handles well with the occlusion situations, when the movement of the blobs is easily predictable, as in figure 4.

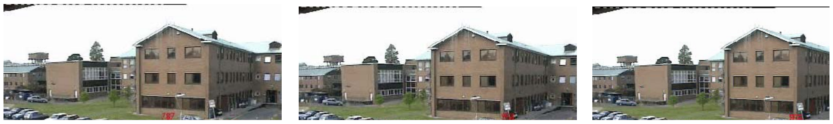

Fig. 4. An example illustrating one difficult tracking situation: a white van is occluded by a thin structure (a street light pole) and a group of people is largely blocked by the van for a few frames. These and other tracking results are accessible to view at *URL*: `http://gps-tsc.upc.es/imatge/_jl/Tracking.html`

Problems occur when a few individually moving objects join each other and form a group. These objects are correctly tracked within the limit of pre-defined MAX_LOST frames as if they were occluding each other. Beyond the limit the system creates a new template for the whole group. Other problems may appear when objects abruptly change their motion trajectories during occlusions: sometimes the system is able to recover the individual objects after the occlusion, but on other occasions new templates are created.

Regarding shadows and highlights they are handled correctly in most cases, though very long cast shadows may not be completely removed sometime.

Finally, objects are correctly classified for over 80% of the frames, using the majority voting classification result via a slide window of W frames, *e.g.* $W = 50$.

6 Conclusion

In this paper, we have presented a robust vision-based system for accurate detection, tracking as well as categorical classification of moving objects in outdoor environments surveyed by a single fixed camera. Each foreground object of interest has been segmented and shadow removed by an effective framework. The 2D appearances of detected object blobs are described by multiple characteristic cues. This template of features are used, by way of scaled Euclidean distance matching metric, for robust tracking of the candidate blobs appeared in the new frame. In completing the system we have also introduced technical solutions to dealing with false foreground pixels suppression, temporal templates adaptation, and have discussed briefly the issues for object classification based on motion history. Experiments have been conducted on real-world scenarios under different weather conditions, and good and consistent performance has been confirmed.

Future work includes resolving the difficult problems of individual moving objects joining-separating-joining by using more persistent appearance modelling; multi-camera cooperative tracking and occlusion handling.

Acknowledgments

This work was performed at the Content and Coding Lab, BT Exact, UK, where JLL was supported by a BT Student Internship, in connection with the EU Framework V SCHEMA project (IST-2000-32795). JLL also acknowledges the bursary from Spanish national project grant number TIC2001-0996.

References

1. Wren, C., Azarbayejani, A., Darrell, T., Pentland, A.: Pfinder: Real-time tracking of the human body. IEEE Transactions on Pattern Analysis and Machine Intelligence (1997)
2. Jabri, S., Duric, Z., Wechsler, H., Rosenfeld, A.: Detection and location of people in video images using adaptive fusion of color and edge information. Proceedings of International Conference on Pattern Recognition (2000)
3. Stauffer, C., Grimson, W.: Learning patterns of activity using real-time tracking. IEEE Transactions on Pattern Analysis and Machine Intelligence (2000)
4. Elgamal, A., Duraiswami, R., Harwood, D., Davis, L.: Background and foreground modeling using nonparametric kernel density estimation for visual surveillance. Proceedings of IEEE (2002)
5. Zhou, Q., Aggraval, J.: Tracking and classifying moving objects from video. Proceedings of Performance Evaluation of Tracking and Surveillance (2001)
6. Horn, K.: Robot Vision. MIT Press (1986)
7. Haritaoglu, Harwood, D., Davis, L.: W4: Real time surveillance of people and their activities. IEEE Transactions on Pattern Analysis and Machine Intelligence (2000)
8. Javed, O., Shah, M.: Tracking and object classification for automated surveillance. Proceedings of European Conference on Computer Vision (2002) 343–357
9. McKenna, S., Jabri, S., Duric, Z., Rosenfeld, A., Wechsler, H.: Tracking groups of people. Proceedings of Computer Vision and Image Understanding (2000)
10. Horpraset, T., Harwood, D., Davis, L.: A statistical approach for real-time robust background subtraction and shadow detection. Proceedings of International Conference on Computer Vision (1999)
11. Fitzgibbon, A., Fisher, R.: A buyer's guide of conic fitting. Proceedings of British Machine Vision Conference (1995) 513–522