

Robust Detection and Tracking of Multiple Objects in Cluttered Scenes

Li-Qun Xu

Content and Coding Lab, BT Exact, Adastral Park, Ipswich, UK

01473 648608

li-qun.xu@bt.com

Abstract

This paper addresses primarily the issue of robustly segmenting and tracking multiple objects in the cluttered outdoor dynamic scene. As compared with the state-of-the-art, two major contributions are 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. The problems of robust tracking, occlusions, object entries / exits are considered.

1. Introduction

The importance of accurate and robust detection and tracking of multiple moving objects in dynamic and cluttered visual scenes cannot be underestimated. 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 events analysis and management systems with a view to understanding the complex actions, interactions, and abnormal behaviours of objects in the scene.

Despite extensive recent research activities (see references) a number of technical challenges remain with real-world surveillance applications environment, such as *natural cluttered scene*; *repetitive background*; *illumination changes*; *occlusions* (inter-object, thin scene structures, large scene structures); *objects entries and exits*; *shadows and highlights* etc.

This paper discusses a robust multi-object tracking system in which several novel ideas are introduced to deal with the above challenging issues to a

considerable success, leading to the enhancement of several aspects of the state-of-the-art object tracking techniques. These include the use of techniques for false foreground pixels suppression; a novel framework for effective cast shadows / highlights removal whilst preserving original object shape; the integrated matching strategy 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.

2. Moving objects segmentation

We use the method due to Stauffer and Grimson [6] for adaptive background learning. Considering the noise in the background and camera jitters, the detected foreground pixels are first subjected to a false-foreground pixels suppression procedure. Following this, a further detection scheme is applied to rid these foreground pixels of possible cast shadows or highlights. The working mechanism of this novel scheme is shown in Figure 1, comprising four steps aiming at good shadows/highlights removal whilst preserving original object shapes:

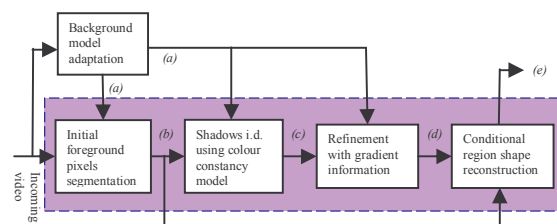


Figure 1: The schematic diagram of the novel shadows / highlights removal approach made up of four main processing steps. The input and output of each block are as follows: (a) the adaptive background image; (b) initial foreground segmentation result; (c) shadows removing using colour constancy model; (d) the result after shadows validation using gradient / texture information, generating the ‘skeleton’ image; and (e) final reconstructed foreground regions.

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

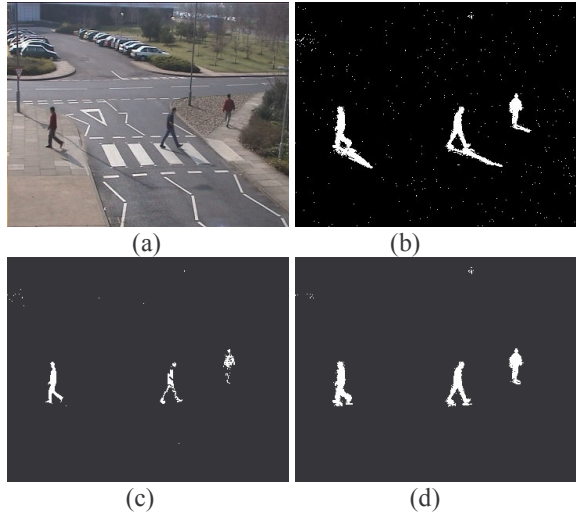


Figure 2: (a) A snapshot of a surveillance video sequence, the cast shadows from pedestrians are strong and large; (b) the result of initial foreground pixels segmentation, the moving shadows being included; (c) The 'skeleton' image obtained after the shadow removing processing; and (d) the final reconstructed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

3. Robust objects tacking

Following the proceeding processes, a connected component analysis is performed to label pixels into respective blobs. The blobs are temporally tracked throughout their movements within the scene by means of temporal templates.

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, shape, and colour of each object / candidate blob, namely:

- The velocity $\mathbf{v} = (v_x, v_y)$ 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;
- The orientation of the major-axis of the ellipse (θ);
- The dominant colour representation (\mathbf{c}_p), using the principal eigenvector of the aggregated pixels' colour covariance matrix of the blob [7].

Therefore at time t , we have, for each object l centred at (p_{lx}, p_{ly}) , a template of features:

$$M_l(t) = (\mathbf{v}_l, s_l, r_l, \theta_l, 1_l(\mathbf{c}_p))$$

Note that, instead of \mathbf{c}_p , we use $1_l(\mathbf{c}_p)$, or the value of 1.0, to denote the dominant colour of the template, and $d_k(\mathbf{c}'_p)$ defined in Eq. (1), to represent the colour similarity between the template l and candidate blob k :

$$d_k(\mathbf{c}'_p) = \frac{\mathbf{c}_p \bullet \mathbf{c}'_p}{\|\mathbf{c}_p\| \cdot \|\mathbf{c}'_p\|} \quad (1)$$

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

In addition, the mean $\bar{M}_l(t)$ 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. It is clear that the variance of each template feature should be analysed and taken into account in the following matching process to achieve a robust tracking result.

Matching procedure

We choose to use a parallel matching strategy in preference to the serial matching one such as that used in [7]. The next issue is to define a proper distance metric that best suits the problem under study. After some thoughtful analysis, a scaled Euclidean distance is adopted, assuming a diagonal co-variance matrix. For a heterogeneous data set, this is a reasonable distance definition.

The matching process can now be discussed:

Given, in frame t , for each object l being tracked so far, we have a template of features $M_l(t)$, its mean and variance vectors $(\bar{M}_l(t), V_l(t))$, the related set of Kalman filters $KF_l(t)$, the counter of tracked frames, ' $TK_counts = n$ ', and the counter of lost frames, ' $MS_counts = 0$ '. By Kalman prediction we also have the predicted (expected) values for l in frame $t+1$, or $\hat{M}_l(t+1)$:

- Step 1: For each new frame at time $t+1$, all the valid candidate blobs detected $\{k\}$ are matched against all the predicted object templates $\{l\}$. A ranking list is then built for each object l . The matching pairs with the lowest cost value that is also less than a threshold, THR , is identified as a match pair.
- Step 2: If object l is matched by the candidate blob k in frame $t+1$, then the ' TK_counts ' is increased by 1, and the normal updates for l are performed:

$M_l(t+1) = B_k(t+1)$ with $1_l(\mathbf{c}_p)$ replaced by $1_k(\mathbf{c}'_p)$, and $(\bar{M}_l(t+1), V_l(t+1))$ as discussed before, and correspondingly the Kalman filters $KF_l(t+1)$.

- **Step 3:** If object l has found no match in frame $t+1$, presumably lost or occluded, then $\bar{M}_l(t+1) = \bar{M}_l(t)$; 'MS_counts' is increased by 1. The object l is carried over to the next frame, with following exceptions:

a) If for object l 'MS_counts \geq MAX_LOST' (e.g., 10), then it is discarded, taken as either being still (merged into background) or entered into a building, car. If 'MS_counts $<$ MAX_LOST', the variance $V_l(t+1)$ is adjusted as below to assist the tracker to recover the lost object that may undergo unexpected or sudden movements.

$$\sigma_i^2(t+1) = (1 + \delta)\sigma_i^2(t)$$

b) As an error in the matching can occur simply due to the prediction errors, thus the prediction model should also be changed to facilitate the possible recovery of the lost tracking. Within the MAX_LOST period, Kalman filters are not used to update the template of features. Instead, for each feature an average of the last 50 correct predictions is used, which states as $M_l(t+1) = M_l(t) + \bar{M}_l(t)$.

- **Step 4:** For each candidate blob k in frame $t+1$ that is not matched, a new object template $M_k(t+1)$ is created from $B_k(t+1)$. Objects will be destroyed if they do not satisfy this condition.

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.

4. Experimental results

The system has been evaluated extensively, using IP-CCTV recordings as well as the benchmarking video sequences provided by PETS' 2001.

Figure 3 shows an example where the white van is occluded by a thin structure, or street light pole (top), and subsequently a group of people are largely blocked by the van for a few frames.

Problems occurred 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 decides that they have disappeared and it then 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 in other cases 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.

References:

- [1] A. Elgamal, R. Duraiswami, D. Harwood and L. Davis, *Proc. of the IEEE*, **90**(7), July 2002.
- [2] I. Haritaoglu, D. Harwood and L. Davis, *IEEE Trans. on PAMI*, **22**(8), August 2000.
- [3] T. Horprasert, D. Harwood and L. Davis, *ICCV'99 FRAME-RATE Workshop*.
- [4] O. Javed and M. Shah, *Proceedings of ECCV'2002*, LNCS 2353, pp. 343-357, 2002.
- [5] S.J. McKenna, S. Jabri, Z. Duric, A. Rosenfeld, and H. Wechsler, *CVIU*, **80**, 42-56, 2000.
- [6] C. Stauffer, W.E.L. Grimson, *IEEE trans. on Pattern Analysis and Machine Intelligence*, **22**(8), August 2000.
- [7] Q. Zhou and J.K. Aggarwal, *Proceedings of 2nd IEEE Int. Workshop (PETS'2001)*, Kauai, Hawaii, USA, 2001.

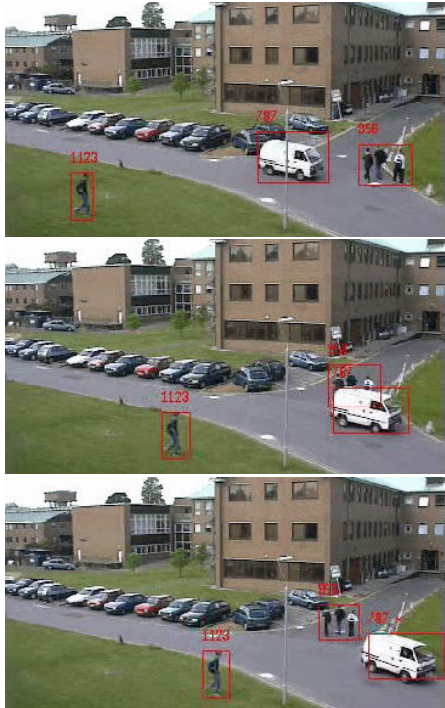


Figure 3: An example illustrating one of the difficult tracking situations that the system handles successfully, in which the moving white van, first occluded by the thin street light pole, then partially occludes the group of walking people.