SEGMENTATION AND TRACKING OF MULTIPLE OBJECTS IN VIDEO SEQUENCES

M. Aprile, A. Colombari, A. Fusiello, V. Murino

Dipartimento di Informatica, University of Verona
Strada Le Grazie 15, 37134 Verona, Italy
{colombar, fusiello, murino}@sci.univr.it

ABSTRACT

This paper describes a system that produces an object-based representation of a video shots composed by a background (still) mosaic and moving objects. Segmentation of moving objects is based on ego-motion compensation and on background modeling using tools from robust statistics. Region matching is carried out by an algorithm that operates on the Mahalanobis distance between region descriptors in two subsequent frames and uses Singular Value Decomposition to compute a set of correspondences satisfying both the principle of proximity and the principle of exclusion. The sequence is represented as a layered graph, and specific techniques are introduced to cope with crossing and occlusions.

1. INTRODUCTION

Digital video is nowadays widespread on the World Wide Web and in multimedia databases. Unfortunately, the usefulness of such large amount of information is limited by the effectiveness of the retrieval method. Whereas text documents are self-describing, digital video does not give any explicit description of its content (see [1] for a review on video indexing). Moreover, transmission of video requires high compression rates to make it viable.

By exploiting the object-based representation offered by MPEG-4 [2], video shots can be encoded as a stationary background mosaic – obtained after compensating for camera motion – plus moving objects (MOs) represented individually. This allows to achieve an high compression rate in the transmission of the sequence, since all the information about the background (which does not change) are sent only once. Besides, this representation of the video is also useful for editing, and it is a step toward its content-based description (as in the new MPEG-7 standard).

The challenge is to create a system that is able to do this segmentation automatically and accurately, and to cope with complex situations, such as crossing between MOs and occlusion with elements of the static background.

Several techniques have been proposed for motion segmentation (see [3] for a review), as temporal analysis of gray-level based on probabilistic models [4], robust motion estimation [5], or misalignment analysis based on the normal flow [6]. In [7], body parts are segmented and tracked, using a body model to help resolving ambiguities and tracking failures. Our tracking approach was inspired by [8], where a graph is used to represent objects and both shape and color features are used to match them.

In our work, MOs are obtained from the original video shot by differencing with the background. For each frame, the mosaic of the background is back-warped onto the frame and each pixel is labeled as belonging to a MO or not by comparing it with a statistical background model. Then, the resulting binary image is cleaned and connected regions (blobs) are identified as candidate MOs. The next step is to exploit temporal coherence: blobs are tracked (non-causally) through the sequence. Finally, noisy tracks are discarded and tracks belonging to the same object are merged. Our work builds on a previous research [9], and improves radically the blob tracking algorithm, allowing for occlusions between MOs, occlusions between a MO and a background object, MOs entering and leaving the scene at any point.

Specific contributions of this papers include the model of the background, based on robust statistics, and the blob matching technique based on a generalization of the method for feature-matching proposed in [10, 11].

1.1. Relationship to MPEG standards

The central concept in MPEG-4 is that of the Video Object (VO). Each VO is characterized by intrinsic properties such as shape, texture, and motion. MPEG-4 considers a scene to be composed of several VOs, which are separately encoded. In MPEG-7 the core element of content description is the Segment Description Scheme (DS), that represents a section of an audio-visual content, resulting from a spatial, temporal, or spatio-temporal partitioning. Segments does not need to be necessarily connected. A video shot is described by a VideoSegment DS. A StillRegion DS describes a spatial segment, and a Mosaic DS is a specialized type of StillRegion, used to describe a panoramic mosaic constructed by aligning and warping the frames of a VideoSegment. The MovingRegion DS represents a spatio-temporal segment, usually identified with an object. More details on MPEG-7 can be found in [12] and the other articles in the same issue.

2. MOTION COMPENSATION

Two pictures of the same scene are related by a (non-singular) linear transformation of the projective plane (or *homography*) in two cases: i) the scene is planar or ii) the point of view does not change (pure rotation). In these cases, which can be summarized by saying that there must be no *parallax*, images can be composed together to form a *mosaic*.

Inter-frame homographies computation is based on correspondences produced by the Kanade-Lucas-Tomasi (KLT) tracker [13], initialized with phase-correlation to reduce search range. As in [9], Least Median of Squares is used to be robust against tracking errors and features attached to moving objects. Finally, given the set of *inlier* point matches, the homography is computed according to a technique proposed by Kanatani[14], which obtains an optimal estimate and reduces the instability of images mapping even with a small overlap between frames. These homographies are then

combined to obtain frame-to-mosaic homographies and frames are warped accordingly and blended to produce a mosaic of the background (assuming that the majority of the tracked features belong to the background).

3. BACKGROUND MODELING

Starting from a single mosaic pixel P, a temporal line piercing all the aligned frames will intersect pixels that correspond to the background and pixels belonging to MOs. The color histogram of these pixels is modeled as a Gaussian distribution corrupted by outliers, corresponding to the MOs. Therefore, the median of the distribution – being a robust estimate of the mean – is taken as the background color and assigned to P: $\bar{c} = \underset{i}{\mathrm{med}}\{c_i\}$. As a result, only the pixels corresponding to the background contribute to the color of P: moving objects are removed. Actually, everything which keeps the same position in the mosaic for more than 50% of the time is included in the background.

Moreover, an estimate of the background color variability at that point is attached to each mosaic pixel P. A robust estimator of the spread of the distribution is given by the median absolute difference (MAD): $\mathrm{MAD} = \mathrm{med}_i\{|c_i - \bar{c}|\}$. It can be seen [15] that, for symmetric distributions, the MAD coincides with the *interquartile range*: $\mathrm{MAD} = (\xi_{3/4} - \xi_{1/4})/2$, where ξ_q is the qth quantile of the distribution (for example, the median is $\xi_{1/2}$). Hence, a pixel with color c, is deemed to belong to the background with 99.9% confidence if

$$|c - \bar{c}| < 5.2MAD \tag{1}$$

This rule comes from the robust statistics [15], where it is known as the the X-84 outlier rejection rule.

4. TRACKING MOVING OBJECTS

MOs are obtained from the original video shot by differencing with the background. Each frame is warped onto mosaic of the background and each pixel is labeled as belonging to a MO or not according to the rule given by Eq. (1). Then, the resulting binary image is cleaned with morphological filtering and connected regions (blobs) are identified as candidate MOs.

A layered graph is built, where each layer correspond to a frame and each vertex is a blob. An edge links two blobs from consecutive layers if they represent the same MO (or part of it) at different time. A *track* is a chain of nodes belonging to consecutive frames, each node belonging to a different frame. The union of several tracks forms a *path*. The goal is to find paths in the graph, each corresponding to a single MO.

4.1. Blob matching

In a first phase, tracks are constructed by matching blobs from one layer to the next. A dissimilarity (distance) measure between blobs is defined taking into account the appearance (shape and color) of the blob and its position. In particular, each blob is described by a feature vector $\mathbf{b}$ composed by: area, solidity, eccentricity, dimension of the bounding box, orientation[1], average color, contrast (standard deviation of the color) and position of the centroid. The dissimilarity of blobs I_i and J_j is computed as the Mahalanobis distance between the respective feature vectors:

$$d_{ij} = (\mathbf{b}_i - \mathbf{b}_j)^{\mathsf{T}}(\mathbf{\Lambda}_I + \mathbf{\Lambda}_J)^{-1}(\mathbf{b}_i - \mathbf{b}_j) \tag{2}$$

[1] See `regionprops` in the MATLAB Image Processing Toolbox

where $\mathbf{\Lambda}_I$ and $\mathbf{\Lambda}_I$ are the covariance matrices of the feature vectors in images I and J respectively.

Matching is carried out with a technique introduced by [10] and elaborated upon by [11], who proposed an algorithm based on the singular value decomposition (SVD) for associating features of two images.

Let $\{I_i\}_{1\ldots n}$ and $\{J_j\}_{1\ldots m}$ the two sets of blobs which are to be put in one-to-one correspondence. The first stage is to build a *proximity matrix* $\mathbf{G}$ of the two sets of features: $G_{ij} = e^{-d_{ij}/2}$. The next stage is to perform the SVD of $\mathbf{G}$

$$\mathbf{G} = \mathbf{U}\mathbf{S}\mathbf{V}^{\mathsf{T}}$$

where $\mathbf{U}$ and $\mathbf{V}$ are orthogonal and $\mathbf{S}$ is a non-negative $m \times n$ diagonal matrix. Finally, $\mathbf{S}$ is converted into a new $m \times n$ matrix $\mathbf{D}$ by replacing every diagonal element S_{ii} with 1, thus obtaining another matrix $\mathbf{P} = \mathbf{U}\mathbf{D}\mathbf{V}^{\mathsf{T}}$ of the same shape as the original proximity matrix and whose rows are mutually orthogonal. The element P_{ij} indicates the extent of pairing between the blobs I_i and J_j. This matrix incorporates the principle of proximity (that favours a match with the closest feature) by construction of $\mathbf{G}$ and the principle of exclusion (that prohibits many-to-one correspondences) by virtue of its orthogonality. If P_{ij} is both the largest element in its row and the largest element in its column, then I_i and J_j are regarded as corresponding with each other, provided that their Mahalanobis distance is below a certain threshold.

The use of Mahalanobis distance is customary in data association [16], but it is often used in a nearest-neighbour scheme; this approach extends it by introducing also the exclusion principle.

Our proximity matrix $\mathbf{G}$ generalizes the solution proposed by [11], because using Mahalanobis distance in a feature space allows to takes into account both appearance and spatial position (and possibly other features) in a consistent way.

Please note that this matching only produces chains of nodes (tracks). Many of them are due to noise, and only a few correspond to moving objects (or their parts). Tracks are classified according to temporal length and *size*, the latter being defined as the average blob area over the track. Bad tracks are those shorter than 5% of the longest sequence and smaller that 5% of the biggest sequence. Bad tracks are marked but not discarded, yet.

4.2. Tracks merging

A path represent the trajectory of a MO, therefore each path is uniquely associated to a MO. This is not true for tracks, as the tracks obtained from the previous search may be related to a part of an object (in the case of occlusion with a thin static element or because of over-segmentation). Tracks that may potentially correspond to this kind of situation have to be connected and merged into a path.

At both ends of each track a local search is carried out to find the blobs that could prolong the track. All the blobs are candidate, also the bad ones and those already belonging to a track (in this case we say that one track has a *collision* with the other one). The search area depends on the blob area and it is centered in the predicted position of the centroid, basing on the last 3 frames. The connection is established with template matching: the template is the blob with smaller area and target image is the other blob. If maximum correlation value is above a threshold, a connection from the template blob to the target blob is created. The search is repeated recursively, until either it fails or it finds a blob belonging to a track. In this way, besides recovering blobs that were not in a track, tracks representing fragments of the same MO can be connected.

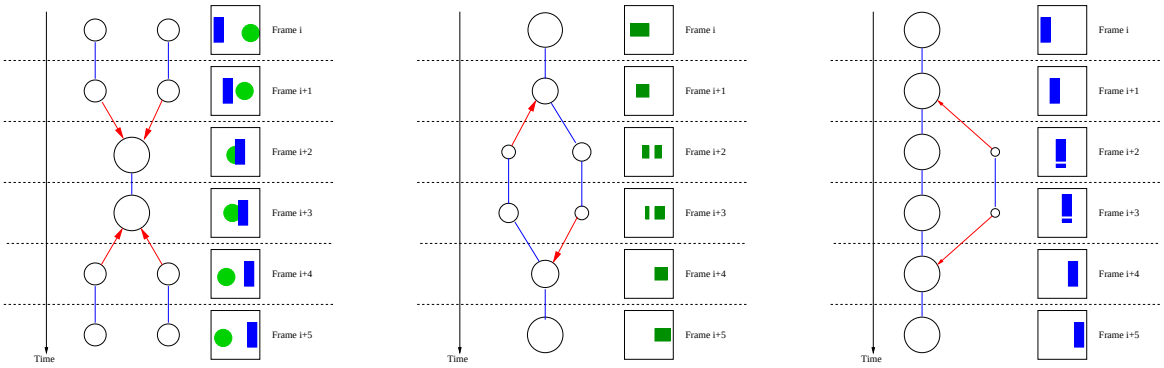

Fig. 1. Types of tracks collision. From left to right: crossing objects, occlusion by a static element, and fragmentation. The red edges (arrows) are those added in the prolongation step.

After this prolongation step, the different kinds of collision are analysed, namely: crossing between objects, occlusion by thin static structures, fragmentation of the objects (Fig.1). Only the colliding tracks corresponding to these situations are allowed to merge. At this point the response of classification is taken into account, and all the noisy tracks that did not merge with any good track gets removed.

Ideally, at this point, every path correspond to a MO, but the reverse is not true. If, for example, an object gets completely occluded, the prolongation step is not effective in this case and one could end up with two distinct paths associated to the same physical object.

As far as the coding is concerned this is uninfluential, but for the content-based representation one would like to preserve the object identity. Our solution is to analyze each pair of paths, and to compare their more representative blobs with template matching. If the maximum correlation is above a threshold the two paths are associated to the same MO.

5. RESULTS

In this section we report some results about object segmentation in two real sequences, taken with a digital hand-held camera, which give raise to collision of type 1 and 2, according to the description on Fig. 1. Radial distortion had been preliminary compensated by calibration [17]. Original sequences and more results, including some editing examples, are available on the web.[2]

In the first experiment we considered a video shot where two persons enter the scene from the opposite side and cross (Fig. 2). The camera does a panning motion, following first the man from left to right and then the woman from right to left. As an example of the segmentation yielded by our technique, Figure 2 shows some MOs extracted form the sequence. Figure 3 shows the mosaic of the background.

In the second experiment, we considered a sequence (Fig. 4) of a moving car. The camera does a panning motion, from left to right. When the car passes behind a pole it is divided in two parts, nevertheless our technique can recover it and recognize it as a single MO. Figure 5 shows the mosaic of the background.

The MPEG-7 compliant coding consists in describing the video sequence as composed by a Mosaic (a type of StillRegion) and MovingRegions. The mosaic is encoded as a still image and MovingRegions are encoded separately. Therefore one needs to send the mosaic, the moving objects and the frame-to-mosaic homographies. The decoder pastes the MOs onto the mosaic and warps it back to produce the original sequence.

Fig. 2. Selected frames from the "Lorena" sequence (top) and the result of segmentation (bottom).

Fig. 3. Mosaics of the "Lorena" background, obtained after global registration.

[2] http://profs.sci.univr.it/~fusiello/demo/motseg

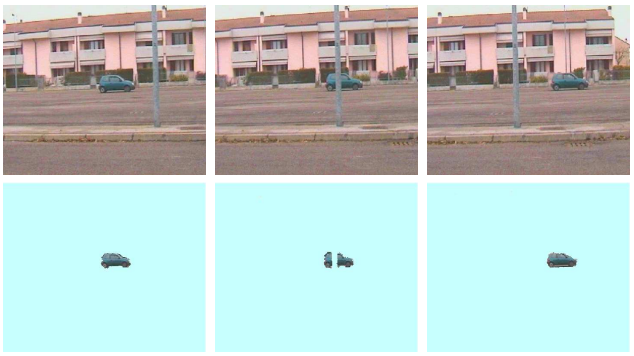

Fig. 4. Selected frames from the "Arosa" sequence (top) and the result of segmentation (bottom).

Fig. 5. Mosaic of the background for the "Arosa"sequence.

6. CONCLUSIONS

We presented a complete system which produces an object-based representation of a video shot, and, in particular, we addressed the problem of multiple objects segmentation and tracking. This paper builds on a previous work [9], and improves both segmentation and tracking. Segmentation is posed as an outlier rejection problem and solved by applying the X84 outlier rejection rule. Our region matching approach is a generalization of Scott and Longuet-Higgins algorithm for feature matching [10, 11], and it extends the classical nearest-neighbour data association scheme by implementing both the principle of proximity (in Mahalanobis distance) and the principle of exclusion. The proposed tracking technique is rather general, and can take into account occlusions between MOs, occlusions between a MO and a background object, MOs entering and leaving the scene at any point.

Our work can be extended in many ways. For example one might use the additional alpha channel in image representation for a more realistic blending of the object with the background [18].

7. REFERENCES

[1] R. Brunelli, O. Mich, and C. M. Modena, "A survey on the automatic indexing of video data," *Journal of Visual Communication and Image Representation*, vol. 10, pp. 78–112, 1999.

[2] R. Koenen, F. Pereira, and L. Chiariglione, "MPEG-4: Context and objectives," *Signal Processing: Image Communications*, vol. 9, no. 4, pp. 295–304, 1997.

[3] D. S. Zhang and G. Lu, "Segmentation of moving objects in image sequence: A review," *Circuits, Systems and Signal Processing*, vol. 20, no. 2, pp. 143–183, 2001.

[4] P.R. Giaccone and G.A. Jones, "Segmentation of global motion using temporal probabilistic classification," in *British Machine Vision Conference*, 1998, pp. 619–628.

[5] H. Sawhney and S. Ayer, "Compact representations of videos through dominant and multiple motion estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 8, pp. 814–830, August 1996.

[6] M. Irani, P. Anandan, J. Bergen, R. Kumar, and S Hsu, "Efficient representations of video sequences and their applications," *Signal processing: Image Communication*, vol. 8, no. 4, pp. 327–351, May 1996.

[7] S. Park and J.K. Aggarwal, "Segmentation and tracking of interacting human body parts under occlusion and shadowing," in *IEEE Workshop on Motion and Video Computing*, 2002, pp. 105–111.

[8] I. Cohen and G. Medioni, "Detecting and tracking moving objects in video surveillance," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1999, pp. II:319–325.

[9] F. Odone, A. Fusiello, and E. Trucco, "Layered representation of a video shot with mosaicing," *Pattern Analysis and Applications*, vol. 5, no. 3, pp. 296–305, August 2002.

[10] G. Scott and H. Longuet-Higgins, "An algorithm for associating the features of two images," in *Proceedings of the Royal Society of London B*, 1991, vol. 244, pp. 21–26.

[11] M. Pilu, "A direct method for stereo correspondence based on singular value decomposition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Puerto Rico, 1997, pp. 261–266.

[12] S. Jeannin and A. Divakaran, "MPEG-7 visual motion descriptors," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 11, no. 6, pp. 720–724, June 2001.

[13] C. Tomasi and T. Kanade, "Detection and tracking of point features," Technical Report CMU-CS-91-132, Carnegie Mellon University, Pittsburg, PA, April 1991.

[14] Kenichi Kanatani and Naoya Ohta, "Accuracy bounds and optimal computation of homography for image mosaicing applications," in *International Conference on Computer Vision*, Sept. 1999, vol. 1, pp. 73–79.

[15] F.R. Hampel, P.J. Rousseeuw, E.M. Ronchetti, and W.A. Stahel, *Robust Statistics: the Approach Based on Influence Functions*, Wiley Series in probability and mathematical statistics. John Wiley & Sons, 1986.

[16] I. Cox, "A review of statistical data association techniques for motion correspondence.," *International Journal of Computer Vision*, vol. 10, no. 1, pp. 53–66, 1993.

[17] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 11, pp. 1330–1334, 2000.

[18] M. Ruzon and C. Tomasi, "Alpha estimation in natural images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2000, pp. 18–25.