Robust Motion Segmentation for Content-based Video Coding

F. Odone[1,2] , A. Fusiello[1] , E. Trucco[1]

[1] Department of Computing & Electrical Engineering Heriot-Watt University, Edinburgh, UK
[2] INFM-DISI, Dip.to di Informatica e Scienze dell'Informazione, Università di Genova, Genova, IT
{franci,fusiello,mtc}@cee.hw.ac.uk

Abstract

This paper presents a motion segmentation method useful for representing efficiently a video shot as a static mosaic of the background plus sequences of moving foreground objects. This generates an MPEG-4 compliant, content-based representation useful for video coding, editing and indexing. Segmentation of moving objects is carried out by comparing each frame with a mosaic of the static background, in which the ego-motion of the camera is compensated for with a robust technique. The automatic computation of the mosaic and the segmentation procedure are compared with the current literature and illustrated with real sequences experiments. An example of content-based manipulation is also shown.

Introduction

This paper presents a mosaic-based motion segmentation method for content-based, MPEG-4 compliant video coding, useful for indexing (Brunelli *et al*., 1999, Chang *et al*., 1997) and editing (Giaccone & Jones, 1998). The compact representation of a video shot we addopt is composed by a mosaic of the background and sequences of the foreground moving objects. This achieves high compression rates, since all the information about the background (which does not change) are processed and transmitted only once. These ideas fit into the new MPEG-4 standard (Koenen *et al.*, 1997), in which a scene is described as a composition of several Video Objects (VOs), encoded separately.

The starting point of our method is the construction of the mosaic of the background, obtained by estimating the relative motion between the camera and the static parts of the observed scene. The segmentation of moving objects is achieved by warping each image of the sequence into the mosaic reference frame and computing grey-level differences between the background and warped image: the main differences between the image and the corresponding part of the mosaic will lie in the areas of the image occupied by the moving objects, which do not appear in the mosaic. The proposed method has been tested on video shots acquired by a commercial hand-held camcorder, without using any special setup.

The structure of this paper is the following : the section ‘‘Mosaicing’’ introduces some key concepts of image sequence alignment and mosaic building in case of single relative motion between camera and observed scene. The section ‘‘Dealing with moving objects’’ extends the concepts previously introduced to a scene containing objects in motion , while section ‘‘Motion segmentation’’ explains how our approach to mosaic building can perform motion segmentation. The sections ‘‘Content-based representation’’ and ‘‘Content-based manipulation’’ describe how our motion segmentation method can be used for video coding and video editing. A few examples of mosaic contruction, video coding and decoding, and video manipulation are shown in the section ‘‘Results’’.

Mosaicing

Mosaicing is the automatic alignment of multiple images into larger aggregates onto a common reference plane (Szeliski, 1996). Image alignment relies on finding corresponding points over a sequence of the scene. This is usually achieved through an approximation of the 2D motion field, the optical flow (Barron *et al.*, 1992, Campani & Verri, 1992), that is, the apparent motion of the image brightness pattern. Direct minimization of discrepancy in pixel intensities have been widely used to align images (Irani *et al.*, 1996, Sawhney & Ayer, 1996, Szeliski, 1996).

We think that feature-based registration, although less common in mosaic applications (Zoghlami *et al*., 1997), is to be preferred for typical digital video sequences with high frame rate, for its lower computational complexity. Feature-based techniques, based on the tracking of two dimensional features such as corners, produce sparse 2D motion representations, using information only where it is most reliable. Also, this approach does not suffer from the aperture problem, (see, for example, Trucco & Verri (1998)) typical of the optical flow. Once a sparse 2D motion field is known, a global 2D motion model, i.e., an appropriate frame-to-frame transformation can be obtained. In some cases, two views of the same scene can be related by a transformation of the projective plane, called *homography*.

We assume, for now, that there is a single, relative motion between camera and scene. A homography (or *collineation*) is a non-singular linear transformation of the projective plane (Semple & Kneebone, 1952) into itself. Two images taken by a moving camera are related by a homography if the scene is planar or if the point of view does not change (the camera is rotating around its optical centre). In the general case (full 3D scene and arbitrary camera motion), the relationship between the two views can be casted terms of a homography plus a *parallax* term depending on the scene structure (Shashua & Navab, 1996). Four point correspondences in two images, if no three of them are collinear, determine a unique homography between the two images. When more point correspondences are available, an overconstrained linear system must be solved, for instance using a least squares estimate.

Let us suppose that we are given an image sequence with a negligible parallax and that point correspondences through the image sequence have been obtained by feature tracking (Shi & Tomasi, 1994, Fusiello *et al*., 1999). Then all the homographies between subsequent frames can be computed, and by composing them, it is possible to obtain transformations relating each image of the sequence an arbitrary reference frame. At this point we have the information we need to warp all the images onto a common reference plane and paste them into a mosaic.

Dealing with moving objects

In the the previous section an assumption of static scene was made. In that case, to calculate the homography from point correspondences, a least squares estimate is appropriate. In the case of multiple image motions, that is, when objects are moving in the scene, features attached to different objects have different motions, and a single homography cannot cater for all of them. Therefore a *robust* method must be employed in order to estimate the homography that explains the motion of the *majority* of the features, that is, the *dominant motion* (Irani *et al*., 1994). Unless the scene is cluttered with many moving objects, this is usually the relative motion of the camera with respect to the background (*ego-motion*).

We adopt Least Median of Squares (LMedS) (Rousseeuw & Leroy, 1987), a robust regression technique which has been used in many computer vision applications (Meer *et al*., 1991, Zhang, 1997). The optimal model represents the majority of data. Data points that do not fit into this model are *outliers*.

Assuming that camera motion is the dominant motion of the sequence, warping the images according to the homography describing the dominant motion yields a sequence where the background appears fixed (having compensated for camera motion), and the other objects are still moving.

In order to build a mosaic, a suitable filter must be used to assign grey levels to the mosaic pixels: if a median operation is chosen, moving objects get removed, and a mosaic of the background is obtained.

Motion Segmentation

The motion segmentation problem can be stated as follows: given a sequence of images, find the regions of the image, if any, corresponding to the different moving objects. This is equivalent to classify the pixels of each frame as either moving according to camera motion or independently.

Other approaches to segmentation through camera motion compensation have been used in the field of surveillance and targeting. In (Cohen & Medioni, 1999, Sawhney & Ayer, 1996) motion is computed at each pixel with a robust technique, and outliers masks correspond to the moving object. In (Giaccone & Jones, 1998) temporal analysis of gray levels, based on probabilistic models and a-priori information, is carried out in order to segment moving objects. Irani *et al.* (1996) use a local misalignment analysis based on the normal flow (Irani *et al*., 1994) between a dynamic mosaic and the original sequence. In our case, since the mosaic does not contain moving objects, a simple pixel-wise difference between mosaic and sequence frame produces a good segmentation of the areas in motion.

We actually achieve the segmentation of moving objects by computing the grey-level differences between the background and every frame of the motion-compensated sequence. Unfortunately the difference image is noisy for several reasons: object or illumination changes, residual misalignments, interpolation errors during warping, and acquisition noise. In order to extract only relevant moving objects from differences, we exploit temporal coherence by tracking the centroids of moving objects over the sequence.

Post-processing is also applied on the resulting maps, to improve segmentation. We use the morphological (Serra, 1982) operator *closure*, that is, *dilation* and *erosion* in cascade, to produce more compact regions, without adding noise and without altering the original shapes.

Content-based representation

In this section we describe how the segmentation method explained above can be used for MPEG-4 video encoding (Wang & Adelson, 1994, Koenen *et al.,* 1997).

In order to achieve content-based manipulation of image sequences, MPEG-4 relies on a segmented representation of the video data. A scene is considered to be composed of several Video Objects (VOs). Each VO is characterized by intrinsic properties such as shape, texture, and motion. In this context, "object" has a very general interpretation, and it is not necessarily a physical object. For example, the background region may be considered as one VO. A *sprite* consists of those regions of a VO that are present in the scene throughout the whole video segment. An obvious example is the "background sprite," that is, the mosaic of the background in a camera-panning shot.

Notice that MPEG-4 standard does not prescribe the method for creating VOs; it simply provides a standard convention for describing them, so that all compliant decoders are able to extract VOs from an encoded bit stream.

If we think of the mosaic of the background and the moving object sequence as VOs, the idea described in the previous section can be seen as an MPEG-4 compliant content-based encoding technique. A mosaic of the background of a video sequence is built and moving objects are segmented. The background sprite is transmitted to the receiver only once. Each moving object in the foreground is transmitted separately as an independent VO, its position described in the mosaic reference frame. All transformations between mosaic and original sequence are also needed; actually, it suffices to transmit all the homographies between consecutive frames, which allow us to relate any two sequence frames. When decoding, to re-build the original sequence, all we have do is to map the mosaic onto the frame of each image and paste the foreground onto it. Examples of coding/decoding are shown in the ‘‘Result’’ section.

Content-based manipulation

This section describes a particular content-based manipulation of a video sequence, in which the segmented representation is exploited to add a synthetic object (an advertising poster), to the background.

We first synthesize a fronto-parallel view of the background plane from the mosaic. This is known as *metric rectification* (Liebowitz & Zisserman, 1998) of a perspective image. A 3D plane and its perspective image are related by a homography, which is fully defined by the relative position of four points in the world plane. Once the homography is determined, the image can be backprojected onto the object plane. After inserting the synthetic object in the rectified mosaic, the mosaic is warped back onto its original plane. Then we use the decoding procedure described in the previous section to create a new sequence with the modified background. An example is shown in the ‘‘Results’’ section.

Results

This section shows some experimental results, obtained from video shots acquired with a commercial hand-held camcorder, no special setup nor calibration were used.

Figure 1 shows selected frames of the "Super5" sequence. This sequence is an outdoor scene with a car driving from the left to the right of the image field of view. The camera motion is mostly rotational, with a small translational component.

Figure 2 shows the mosaic of the background. In spite of the fact that the camera motion is not exactly rotational and the scene not planar, the registration obtained is very satisfactory. Note also that moving objects have been automatically removed without artifacts. Figure 3 illustrates a result of segmentation, showing a selected frame of the foreground sequence. In order to assess our coding technique, we encoded and decoded the "Super5" sequence and compared the result with the original one.

Figure 4 shows the same frames of Figure 1, after the coding and decoding transformations, whereas Figure 5 visualises the differences between the frame in the centre of Figure 4 and the original one, at the centre of Figure 1. As an image quality measure we computed the Peak Signal-to-Noise Ratio (PSNR) (Gonzales & Woods, 1992), based on the sum of squared differences between corresponding pixels of two images, one from the original sequence and the other from the coded-decoded one (Figure 6). The figure shows that the quality of the compression does not degrade too much throughout the sequence.

Finally, Figure 7 presents a result of video editing, where the "Heriot-Watt University" poster is inserted into the original sequence. On the left the metrically rectified mosaic is shown, where the artificial poster has been inserted, on the right there is a sample frame of the synthetic sequence.

More examples and sequences are available on the Internet at:
http://www.cee.hw.ac.uk/~franci/mosaic_demo/mosaic.html.

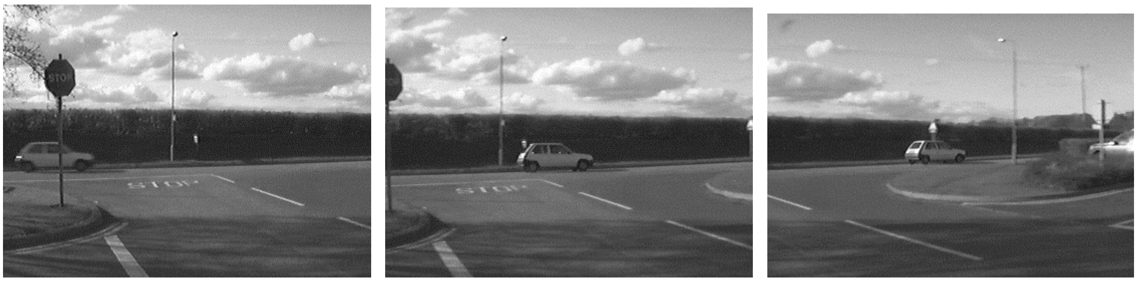

Figure 1: Frames 0, 16, and 40 from "Super5" sequence.

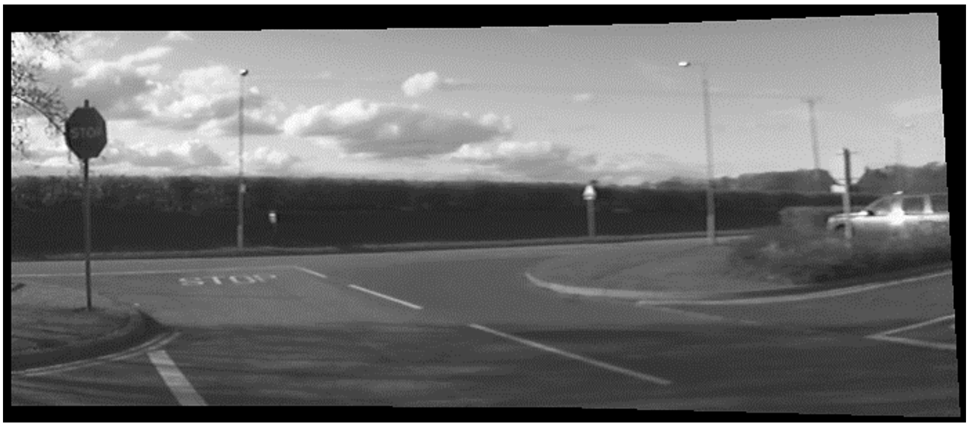

Figure 2: Mosaic of "Super5" (background sprite.)

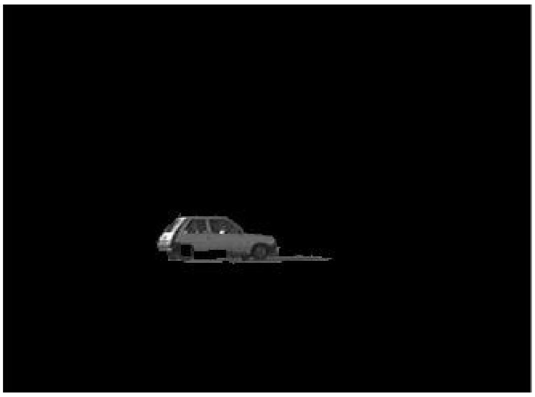

Figure 3: Example of moving object extracted from the sequences "Super5" .

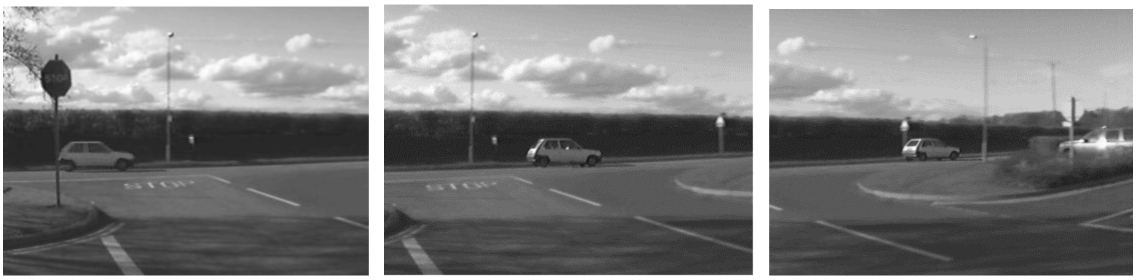

Figure 4: Frames 0, 16, and 40 from the coded-decoded sequence.

Figure 5: Differences between the encoded-decoded frame 16 and the original one.

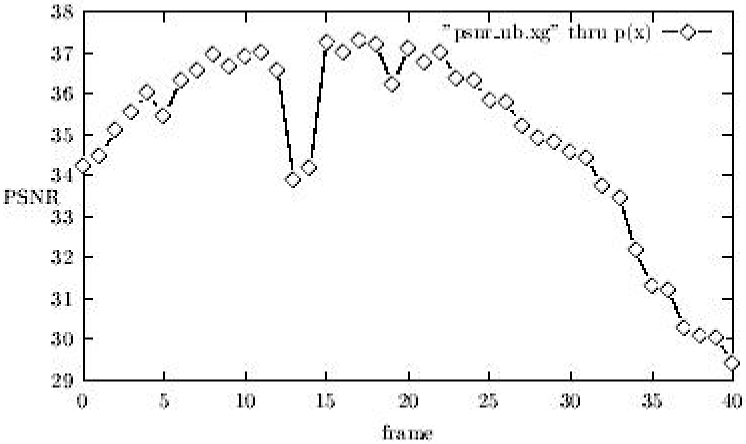

Figure 6 : Peak signal to noise ratio (dB) comparing the original and the encoded/decoded "Super5" sequence.

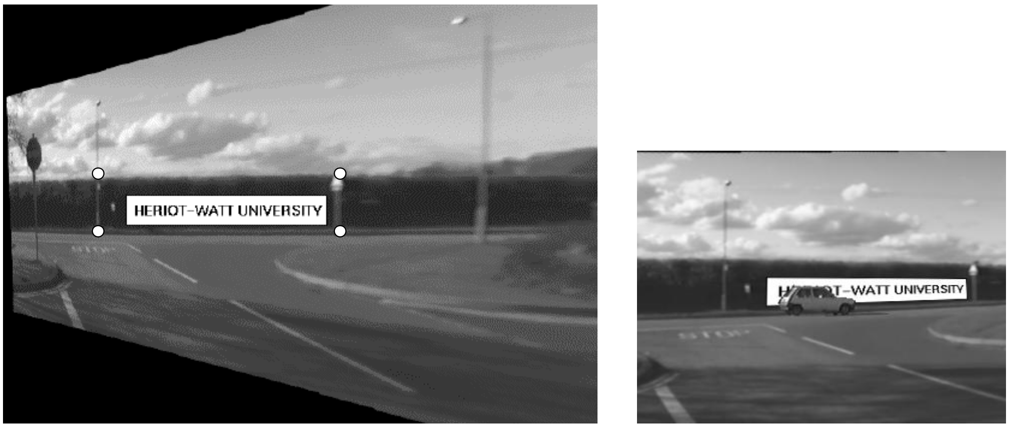

Figure 7: On the left the metrically rectyfied mosaic of the sequence "Super5": the four points that have been used to compute the homography are highlighted. On the right, a sample frame of the synthetic advertisement sequence.

Conclusions

This papers described a mosaic-based motion segmentation method that can be used to perform (MPEG-4 compliant) content-based coding of video sequences.

A feature-based motion estimation technique was preferred, since it is faster, and extracts information only where it is most reliable. A global transformation between each pair of images of the sequence was obtained by calculating homographies.

Segmentation was achieved producing a mosaic of the areas moving of the dominant motion, and comparing this static mosaic with all the frames of the sequence. This approach produced a content-based coding of the static background and the moving foreground objects.

At the present we assume that only one object is moving, but further work will address multiple object tracking and data association (Bar-Shalom & Fortmann, 1988). We reckon that this could be done without changes to the basic structure of the algorithm.

A number of experiments have been carried out to verify the quality of the sequences obtained after decoding. Very promising results have been obtained, where image quality is well preserved throughout the sequence.

Acknowledgements

We wish to thank Francesco Isgrò for the useful comments he made during the course of this work and Costas Plakas who wrote the feature tracker's code. Francesca Odone has been supported by a "Marie Curie" Training and Mobility Grant (ERB4001GT-97-3072). Andrea Fusiello has been supported by a EPSRC Grant (GR/M40844).

References

Bar-Shalom, Y. & Fortmann, T. E. (1988). *Tracking and data association.* Academic Press.

Barron, J. L.; Fleet, D. J.; Beauchemin, S. S.; & Burkitt, T. A. (1992). Performance of optical flow techniques. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 236--242.

Brunelli, R.; Mich O. ; & Modena C. M. (1999). A survey on the automatic indexing of video data. *Journal of Visual Communication and Image Representation*, 10:78--112.

Campani, M & Verri, A. (1992). Motion analysis from first order properties of optical flow. *Computer*

Vision, Graphics, and Image Processing, 56(1):90--107.

Chang, S. F. ; Chen, W.; Meng, H. J. ; Sundaram, H. ; & Zhong, D. (1997). Videoq: An automated content based video search using visual cues. In *Fifth ACM Multimedia Conference,* Seattle.

Cohen, I. & Medioni, G. (1999). Detecting and tracking moving objects in video surveillance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* pages II:319--325.

Fusiello, A. ; Trucco, E. ; Tommasini, T. ; & Roberto, V. (1999). Improving feature tracking with robust statistics. *Pattern Analysis and Applications,* 2(4):312--320.

Giaccone, P.R. & Jones, G. A. (1998). Segmentation of global motion using temporal probabilistic classification. In *British Machine Vision Conference,* pages 619--628.

Gonzales, R. C. & Woods, R. E. (1992). *Digital image processing*, Addison Wesley.

Irani, M.; Anandan, P. ; Bergen, J. ; Kumar, R. ; & Hsu, S. (1996). Efficient representations of video sequences and their applications. *Signal processing: Image Communication*, 8(4):327--351.

Irani, M. ; Rousso, B. ; & Peleg, S. (1994). Computing occluding and transparent motions. *International Journal of Computer Vision*, 12(1):5--16.

Koenen, R. ; Pereira, F. ; & Chiariglione, L. (1997). MPEG-4: Context and objectives. *Signal Processing: Image Communications,* 9(4):295--304.

Liebowitz, D. & Zisserman, A. (1998). Metric rectification for perspective images of planes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* pages 482--488.

Meer, P. ; Mintz, D. ; Kim, D. Y. ; & Rosenfeld A. (1991). Robust regression methods in computer vision: a review. *International Journal of Computer Vision,* 6:59--70.

Rousseeuw, P. J. & Leroy, A. M. (1987). *Robust regression & outlier detection*. John Wiley & sons.

Sawhney, H. & Ayer, S. (1996). Compact representations of videos through dominant and multiple motion estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(8):814--830.

Semple, J. G. & Kneebone, G. T. (1952). *Algebraic projective geometry*. Oxford University Press.

Serra, J. (1982). *Image Analysis and Mathematical Morphology*. Academic Press.

Shashua, A. & Navab, N. (1996). Relative affine structure: Canonical model for 3D from 2D geometry and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 18(9):873--883.

Shi, J. & Tomasi, C. (1994). Good features to track. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 593--600.

Szeliski, R. (1996) Video mosaics for virtual environments. *IEEE Computer Graphics and Applications,* 16(2):22--30.

Trucco, E. & Verri, A. (1998). *Introductory Techniques for 3-D Computer Vision*. Prentice-Hall.

Wang, J. Y. A. & Adelson, E. H. (1994). Representing moving images with layers. *IEEE Transactions on Image Processing*, 3(5):625--638.

Zhang, Z. (1997). Parameter estimation techniques: a tutorial with application to conic fitting. *Image and Vision Computing*, 15(1):59--76.

Zoghlami, I. ; Faugeras, O. ; & Deriche, R. (1997). Using geometric corners to build a 2D mosaic from a set of images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 420--425.