SPATIO-TEMPORAL SEGMENTATION BASED ON MOTION AND STATIC SEGMENTATION

Frédéric Dufaux[†], *Fabrice Moscheni*[‡] *and Andrew Lippman*[†]

[†]The Media Laboratory, Massachusetts Institute of Technology
Cambridge, MA 02139, USA

[‡]Signal Processing Laboratory, Swiss Federal Institute of Technology
CH-1015 Lausanne, Switzerland

ABSTRACT

The problem to segment an image sequence in terms of regions characterized by a coherent motion is among the most challenging in image sequence analysis. This paper proposes a new technique which sequentially refines the segmentation and the motion estimation by combining static segmentation and motion information. Simulation results show the efficiency of the proposed technique.

1. INTRODUCTION

This paper addresses the problem of segmenting an image sequence in terms of moving objects as well as to estimate the motion of these objects through time. This ill-posed problem is among the most challenging in image sequence analysis. There is indeed a very strong dependence between the processes to segment the moving objects and to estimate their motion. On the one hand, as the motion estimation depends on the region of support, a good segmentation is needed in order to precisely estimate the motion. On the other hand, as the moving regions are defined by a coherent motion, an accurate estimate of the motion is required to obtain a good segmentation. To overcome the above problem, various algorithms have been proposed.

In [1], a 3D segmentation based on luminance information is performed by morphological operators. The drawback of this technique is that the scene is segmented according to a criterion of uniform luminance instead of coherent motion.

Another approach consists in estimating first the motion in the scene by means of a dense optical flow. Based on this motion information, moving objects are then defined by the pixels whose motion is coherent. For instance, in [2] motion parameters are estimated from the optical flow by regression, and the spatio-temporal segmentation is then obtained by a clustering in the motion parameters space. The initial optical flow estimation, which plays a crucial role, relies on a non-parametric motion model [3]. This model requires an explicit constraint such as a smoothness constraint [4]. It results in inaccurate motion estimates on motion boundaries. Furthermore, the optical flow estimation relies on local computation and is therefore limited in accuracy.

In order to overcome the drawbacks of the above approaches, spatio-temporal segmentation techniques which refine the segmentation and the motion estimation have been proposed. In [5], the segmentation is expressed as a relaxation problem based on a Markov Random Field (MRF) modeling and a Bayesian criterion. With this approach, the spatio-temporal segmentation and the motion are simultaneously estimated. In contrast, the techniques in [6, 7, 8] sequentially refine the segmentation and the motion estimation. In [6], a single dominant motion is first estimated. Then, the current image is compared with the warped image, and new regions are defined as the areas corresponding to large prediction errors. The same procedure is applied to each of these new regions recursively. In [7] the motion estimates corresponding to each region are iteratively refined in a similar way. After each iteration pixels are assigned to one of the regions based on the prediction error. In [8], the motion estimation and segmentation is performed on multiple frames and takes into account the information provided by a static segmentation.

In the above techniques [5, 6, 7, 8], the support of the motion estimation corresponds to a whole region of the scene. Assuming that the pixels within this region undergo a coherent motion, a parametric motion model [3] can be introduced. The advantage of the parametric model, when compared to the non-parametric one, is that since all the pixels within the region of support

This work was partly supported by the Television of Tomorrow Program at the MIT Media Laboratory

can contribute to the motion estimation, robustness and high accuracy can be expected. Nevertheless, errors may occur when the support of the motion estimation does not correspond to an area characterized by a coherent motion. To overcome this problem, robust estimators which are less sensitive to outliers can be used [9].

In this paper, a new algorithm for the representation of a scene in terms of moving object is proposed. It distinguishes itself form other techniques by the following characteristics. In order to handle efficiently camera motion (e.g. zoom or pan), the latter is first removed through global motion estimation performed by a frame matching technique [10]. The image sequence is then pre-filtered to make it easier to segment. The subsequent spatio-temporal segmentation process takes into account the information provided both by static segmentation and motion estimation. Starting from the static segmentation, local motion estimation is performed. The local object-based motion estimation relies on a parametric affine motion model and is performed by a matching algorithm using a robust estimator. Regions corresponding to a failure of the motion estimation are further split by a clustering on the luminance. Conversely, regions characterized by a similar motion are merged by a clustering in the motion parameters space.

By exploiting the static segmentation, this algorithm assures very precise motion boundaries. In particular it avoids problems related to occlusion and uncovered background. Besides, the two-stage global/local matching motion estimation algorithm using robust estimator is characterized by its robustness and its resilience to noise leading to high performances. Therefore, the proposed algorithm is efficient to find the spatio-temporal meaningful entities existing in the scene.

2. SPATIO-TEMPORAL SEGMENTATION

The purpose of the proposed algorithm is to segment the regions characterized by a coherent motion, or in other words to determine the moving objects in the scene. The algorithm sequentially refines the segmentation and the motion estimation by efficiently combining static segmentation and motion information. More precisely, it starts from a static segmentation and splits or merges regions based on their motion information. This algorithm is described in more detail in the remaining of this section.

The motion arising in a scene can be decomposed into a global motion due to the camera (e.g. pan or zoom) and a local motion due to the displacement of the objects in the scene. In order to efficiently handle camera motion, a global motion estimation is first carried out by a frame matching technique [10] to remove the camera motion.

To make the image easier to segment, a spatial pre-filtering is then applied. The purpose of this pre-filter is to produce constant luminance regions delimited by sharp contours. The morphological operator *open-close by reconstruction* has been shown to be well adapted for this task [1]. It indeed produces flat zones while preserving the contour information. Therefore this operator is used in the remaining.

Following the pre-processing stage, the static segmentation is performed by a k-means clustering algorithm [11] on the luminance values. For each of the resulting static regions, affine motion parameters are computed using a matching technique (see Sec. 3). With this structure, the motion estimation is applied on a region characterized by a coherent motion, allowing therefore a precise and robust estimate. Nevertheless, to cope with a badly defined support of the motion estimation due to a failure of the segmentation, a robust estimator [9] less sensitive to outliers is used.

Because they are supposed to be the least significant, small or low contrast image features are lost during the segmentation process. However, some of these features might carry a visually important relevance. To partially overcome this problem, regions which are not well compensated are further split. The decision is taken based on a threshold on the prediction error. The split is then carried out by a static segmentation performed by k-means clustering on each of the selected regions. Furthermore, the new regions are restricted to the areas of the selected regions which correspond to a high prediction error. This refinement of the static segmentation allows to recover some of the small or low contrast features having a significantly distinct motion which can therefore be assumed to be visually important.

Afterwards, regions with similar motions are merged by applying a k-medoid clustering algorithm [11] in the motion parameters space. At this stage, the k-medoid is preferred to the k-means clustering for the following reason. The k-means is very sensitive to outliers, whereas in the k-medoid the centroid of each cluster is chosen among the elements of the input data resulting in a more robust clustering and less sensitivity with respect to outliers. This clustering in the motion parameters space results in regions characterized by coherent motion which can therefore be identified as the moving objects in the scene.

The main advantage of the above spatio-temporal segmentation algorithm is that as the boundaries are computed on the luminance signal, they can be very precisely located. In particular problems related to occlusion and uncovered background are avoided. The segmentation being applied between two frames only, a coherent segmentation through time is not granted. In order to overcome this problem, a tracking algorithm based on a Kalman filter [12] can be used.

3. LOCAL OBJECT-BASED MOTION ESTIMATION

In the proposed algorithm, the motion is estimated in two stages. First the camera motion is removed by global motion estimation [10]. The local motion due to the displacement of the objects in the scene is then handled by local object-based motion estimation. This section addresses this second stage. More precisely, the problem to estimate the motion parameters of a region characterized by a coherent motion is considered.

The problem of motion estimation is ill-posed. In order to regularize it, all motion estimation techniques rely on a motion model. According to [3], the motion models can be classified as non-parametric or parametric (in fact a further distinction is made between quasi-parametric or fully parametric). Non-parametric models rely on a dense local motion field (e.g. optical flow motion estimation techniques [4]) and requires and explicit constraint (e.g. smoothness constraint [4]). Due to the explicit constraint, motion cannot be accurately estimated on motion boundaries, and thus the latter cannot be precisely located. Moreover, the optical flow estimation involves only local computation and its accuracy is therefore limited. In contrast, the parametric models represent the motion of a large region by a single set of parameters, refer to as motion parameters. In this case, since all the pixels within the region of support contribute to the motion estimation, robustness and higher accuracy is obtained. As far as object-based motion estimation is concerned, as an object is characterized by a coherent motion, it is natural to apply a parametric model.

In order to estimate the parameters of a parametric motion model two different approaches have been investigated. The techniques in the first class are composed of two steps: the computation of a dense optical flow using a non-parametric technique (i.e. with an explicit constraint), followed by the modeling of the motion vectors by a set of motion parameters [2]. As they do not compute the motion parameters from the luminance signal itself, these methods can be referred to as indirect. Their drawback is that they depend strongly on the efficiency of the non-parametric motion estimation technique.

The second class of parametric motion estimation techniques directly estimates the parameters of the motion model as in [6, 3, 8]. The computation of the motion field is implicitly constrained by the motion model itself, and consequently no additional explicit constraint is required. As the estimation is carried out on the luminance signal itself, these techniques can be seen as direct. Differential techniques are the most widely used to solve this problem [6, 3, 8]. They are based on a Taylor expansion of the luminance signal. The method in [5] relies on a Markov Random Field (MRF) modeling and a Bayesian criterion.

An alternative approach for direct parametric motion estimation is the matching technique [10]. In contrast with the differential and Bayesian techniques, the matching technique does not rely on a model of the luminance. Therefore, it is characterized by its robustness and its resilience to noise. In [10], this technique has proved to outperform differential and regression techniques to estimate the camera motion. For this reason, this technique is adopted in the algorithm proposed in this paper. The matching motion estimation relies on an affine motion model which allows to represent the motion of a planar surface under orthographic projection. To decrease the computational complexity and to allow a non-exhaustive search while avoiding local minima, a Gaussian pyramid structure of the input images is built. The final motion parameters at one level propagate as initial estimates on the next level. A deterministic relaxation scheme is applied during the propagation stage. It compares the motion parameters obtained for neighboring regions and selects the one providing the lowest prediction error. This deterministic relaxation scheme allows to avoid local minima.

In the parametric motion estimation, pixels within a region are assumed to undergo a coherent motion. Errors may occur when the support of the estimation is not well defined and the latter hypothesis does not hold. To overcome this problem, a robust estimator is used [9]. This estimator is less sensitive to outliers and provide a reliable motion estimate although two or more motions are present in the region.

4. SIMULATION RESULTS

Simulation results are presented in this section. Figures 1 and 2 show a frame of the two test sequences "Car" and "Tennis" and the corresponding final spatio-temporal segmentation obtained by the proposed algorithm. The latter generated 9 and 7 moving regions for "Car" and "Tennis" respectively. In particular, the moving objects are effectively segmented, showing the efficiency of the method. Furthermore, the motion boundaries are very precisely located.

5. CONCLUSION

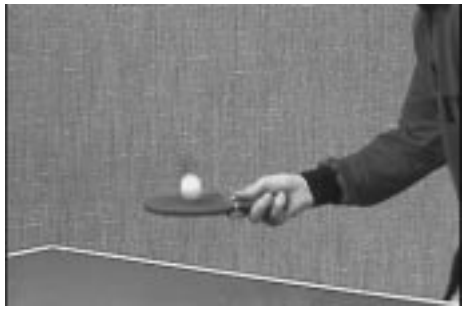
A new algorithm to segment an image sequence in terms of regions characterized by a coherent motion has been proposed. It efficiently combines static segmentation and motion information. In particular, the method leads to very precisely located boundaries as the latters are computed on the luminance signal. Furthermore, motion is robustly computed by a global estimation which remove the camera motion, followed by a local estimation using a matching technique and a robust estimator.

a)

b)

Figure 1: a) a frame of "Car", and b) the final spatio-temporal segmentation.

a)



b)

Figure 2: a) a frame of "Tennis", and b) the final spatio-temporal segmentation.

Future research works will investigate the use of a tracking algorithm based on Kalman filter [12] in order to determine the evolution of the detected objects and therefore to assure a coherent segmentation through time.

6. REFERENCES

[1] M. Pardas and P. Salembier. 3D morphological segmentation and motion estimation for image sequences. *Signal Processing*, vol. 38, no. 2, pp. 31-43, September 1994.

[2] J.Y.A. Wang and E.H. Adelson. Spatio-temporal segmentation of video data. In *SPIE Proc. Image and Video Processing II*, volume 2182, San Jose, CA, February 1994.

[3] P. Anandan, J.R. Bergen, K.J. Hanna, and R. Hingorani. Hierarchical model-based motion estimation. In M.I. Sezan and R.L. Lagendijk, editors, *Motion Analysis and Image Sequence Processing*, pages 1–22. Kluwer Academic Publishers, 1993.

[4] B.K.P. Horn and B.G. Schunck. Determining optical flow. *Artif. Intell.*, vol. 17, pp. 185-203, 1981.

[5] P. Bouthemy and E. François. Motion segmentation and qualitative dynamic scene analysis from an image sequence. *Int. Journal of Computer Vision*, vol. 10, no. 2, pp. 157-182, 1993.

[6] M. Hoetter and R. Thoma. Image segmentation based on object oriented mapping parameter estimation. *Signal Processing*, vol. 15, no. 3, pp. 315-334, October 1988.

[7] S. Peleg and H. Rom. Motion based segmentation. In *IEEE Proc. Int. Conf. on Pattern Recognition*, pages 109–113, Atlantic City, NJ, June 1990.

[8] P. Schroeter and S. Ayer. Multi-frame based segmentation of moving objects by combining luminance and motion. In *Proc. EUSIPCO 94*, Edinburgh, U.K., September 1994.

[9] P.J. Rousseeuw and A.M. Leroy. *Robust Regression and Outlier Detection.* Wiley, New York, 1987.

[10] F. Moscheni, F. Dufaux, and M. Kunt. A new two-stage global/local motion estimation based on a background/foreground segmentation. In *IEEE Proc. ICASSP'95*, Detroit, MI, May 1995.

[11] L. Kaufman and P.J. Rousseeuw. *Finding Groups in Data: an Introduction to Cluster Analysis.* Wiley, New York, 1990.

[12] Y. Bar-Shalom and T.E. Fortmann. *Tracking and Data Association.* Academic Press, 1988.