A ROBUST MOTION ESTIMATION AND SEGMENTATION APPROACH TO REPRESENT MOVING IMAGES WITH LAYERS

Luis Torres, David García and Anna Mates

Dept. of Signal Theory and Communications
Universitat Politècnica de Catalunya
Gran Capità s/n, D5
08034 Barcelona, Spain
E-mail: {luis, albroto, almates}@gps.tsc.upc.es

ABSTRACT

The objective of this paper[1] is to provide a robust representation of moving images based on layers. To that goal, we have designed efficient motion estimation and segmentation techniques by affine model fitting suitable for the construction of layers. Layered representations, originally introduced in [4] are important in several applications. In particular, they are very appropriate for object tracking, object manipulation and content-based scalability which are among the main functionalities of the future standard MPEG-4. In addition a variety of examples are provided that give a deep insight into the performance bounds of the representation of moving images using layers.

1. INTRODUCTION

The basic idea behind the new standard MPEG-4 [1] is to represent the world *understood* as a composition of audio-visual objects, following a script that describes their spatial and temporal relationship. This type of representation should provide the possibility that the user interacts with the various audio-visual objects in the scene, in a way similar to the actions taken in everyday life. Although this content-based approach to the scene representation may be considered *evident* for a human being, it represents in fact a revolution in terms of video representation architecture used in the available standards, since it allows a *jump* in the type of functionalities that may be provided to the user. A scene represented as a composition of (more or less independent) audio- visual objects offers to the user the possibility to *play* with the scene content, by changing some of the objects characteristics (e.g. position, motion, texture or shape), by accessing only selected parts of scene or even by *cut and pasting* objects from one scene to another. Content and interaction are thus central concepts in MPEG-4 [2, 3].

An approach to represent the visual world in terms of objects (called Video Object Planes, VOP's, in MPEG-4) is the representation in layers proposed by Wang and Adelson [4]. Each layer contains three different maps: 1) the intensity map, 2) the alpha map, which defines the opacity or transparency of the layer at each point and 3) the velocity map which describes how the map should be warped over time. To clearly show the concepts behind the layered representation, Figure 1 advances our own results on the layered representation of the *Flower Garden* sequence. The representation has been able to extract four different layers corresponding to the tree, the houses, the garden and the background. The last image represents the reconstruction obtained from the different layers. In each layer all the visible information of the corresponding object is accumulated in one single extended frame. Notice that this representation allows manipulation of the content of the scene and besides each frame is defined only through the motion parameters what gives a very compact representation very useful in video coding applications.

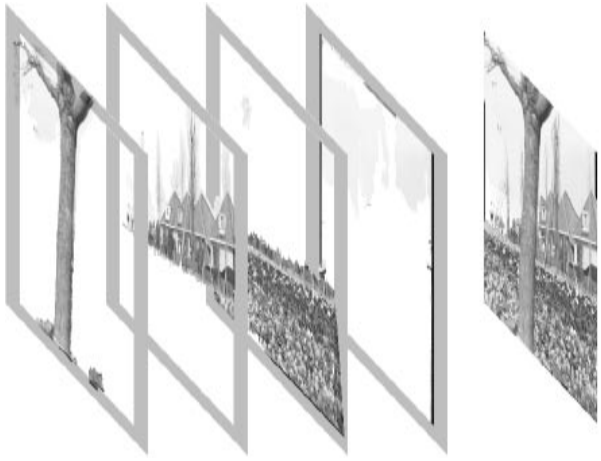


Figure 1. Representation of moving images using layers

Our construction of the layers follows conceptually the approach presented in [4], but using more efficient and robust techniques. The process implies two basic operations: 1) a local motion estimation and 2) a

[1] This work was supported in part by the CICYT grant of the Spanish government: TIC95-1022-C05

motion segmentation by affine model fitting. We have developed a simple and yet robust method to find the motion field. The technique is presented in Section 2. Once a set of affine models has been found, similar models are grouped based in a mean-square distance between the motion vectors of the pixels belonging to each of the models being compared. This gives an efficient motion segmentation strategy which is explained in Section 3. Section 4 provides some results of reconstructed sequences using this layered approach along with an example of object manipulation. Finally, Section 5 draws some conclusions.

2. ROBUST MOTION ESTIMATION

In order to estimate the motion field of the scene, we estimate first an initial approximation of the motion of some pixels located in a rectangular grid using block matching techniques. Given the position $\vec{r}$ of a pixel and a vector motion $\vec{v} = (v_x, v_y) \mid |v_x|, |v_y| \leq$ *maximum displacement*, the following cost function is defined:

$$G(\vec{r}, \vec{v}) = \sum_{\vec{r\prime} \in W(\vec{r})} \left\| I_t(\vec{r\prime}) - I_{t-1}(\vec{r\prime} - \vec{v}) \right\|^2 \qquad (1)$$

where $W(\vec{r})$ is a neighboorhod of $\vec{r}$ and I_t, I_{t-1} are the actual and previous images. The motion vectors of the remaining pixels are interpolated from these first ones. Due to local minima of the cost function, some motion vectors may be erroneous. To improve these motion vectors, a motion gradient is found for each pixel:

$$MG(\vec{r}) = \frac{1}{4} \sum_{\vec{d} = \begin{cases} (1,\ 0) \\ (0,\ 1) \\ (\text{-}1,\ 0) \\ (0,\ \text{-}1) \end{cases}} \left\| \vec{v}(\vec{r}) - \vec{v}(\vec{r} - \vec{d}) \right\|^2 \qquad (2)$$

Then for those pixels whose motion vectors are above the mean of all the gradients, the motion vectors are recalculated using a bigger $W(\vec{r})$. The interpolated motion vectors are refined with integer precision. The final motion vector for each pixel of the whole image is obtained through a refinement process based on one third-pixel accuracy.

Please notice that if occluded areas of I_{t-1} are uncovered in I_t, the cost function $G(\vec{r}, \vec{v})$ cannot give the correct motion estimation vector. To minimize this problem, the above explained technique is applied to estimate the motion field between I_t y I_{t+1}. This information is used to compensate for the errors of the first motion field (I_{t-1} and I_t). Figure 2 shows the motion field obtained for the first two images of the *Flower Garden sequence*. This figure shows a motion vector per pixel, an acceptable motion field and a quite good performance in the contours.

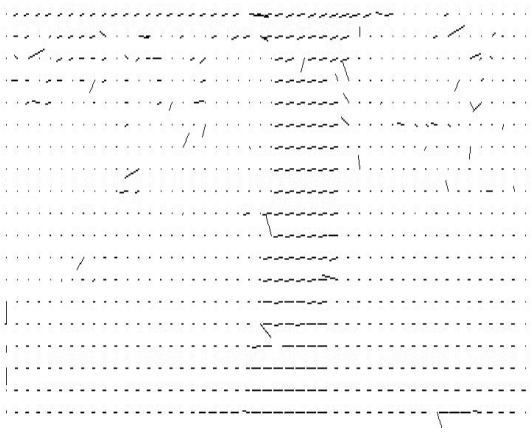


Figure 2. Motion field of the Flower Garden sequence between the first two images

3. ROBUST MOTION SEGMENTATION

The next step is to segment the motion vector field by fitting an affine motion model. The objective is that those regions of the image with the same affine model will be grouped. As we do not have a reliable initial image partition, the image is initially partitioned in rectangular regions. For each of these rectangular regions, the best affine model in the minimum square sense is found. Once a set of affine models have been found, similar models are grouped based in a mean-square distance between the motion vectors of the pixels belonging to each of the models being compared. Then, a motion segmentation process is applied through an iterative k-means clustering algoritm. Two models are grouped each time. The ordering in which all the model pairs are grouped is of paramount importance as it will fix the stability and convergence of the technique. The distance between two given affine models $\mathbf{a}_i$ and $\mathbf{a}_j$ belonging respectively to the regions $\mathcal{R}_i$ and $\mathcal{R}_j$, is defined as

$$dist(\mathbf{a}_i, \mathbf{a}_j) = \frac{1}{n} \sum_{\vec{r} \in \mathcal{R}_i, \nabla_j} \left\| V_{\mathbf{a}_i}(\vec{r}) - V_{\mathbf{a}_j}(\vec{r}) \right\|^2 \qquad (3)$$

where n indicates the number of pixels of $\mathcal{R}_i \cup \mathcal{R}_j$, $V_{\mathbf{a}_i}(\vec{r})$ and $V_{\mathbf{a}_j}(\vec{r})$ the motion vectors represented by $\mathbf{a}_i$ and $\mathbf{a}_j$ in the position $\vec{r}$. Once the distance between two affine models is defined, the grouping of models is done according to the following steps:

1. Find the pair (i, j) such that the distance between the corresponding models is minimum:

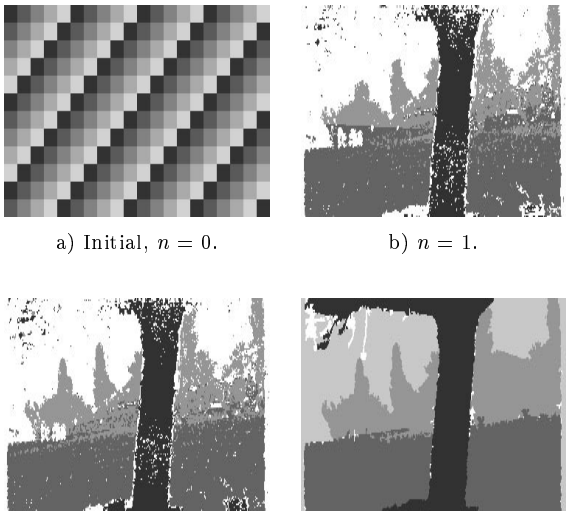
$$dist(\mathbf{a}_i, \mathbf{a}_j) \leq dist(\mathbf{a}_k, \mathbf{a}_l) \ \forall \ (k, l) \neq (i, j)$$

2. Group $\mathbf{a}_i$ and $\mathbf{a}_j$ in a single model, which gives the best estimation of the motion field of the regions $\mathcal{R}_i$ and $\mathcal{R}_j$.
3. Iterate the process while $dist(dist(\mathbf{a}_i, \mathbf{a}_j)$ is less than a prespecified threshold. A threshold of 1 pixel is a very reasonable option for most of the sequences.

Using this grouping technique, the segmentation algorithm proves to be very efficient as it gives very reliable final models at a fast convergence rate. We believe that this technique represents also an important improvement with respect to the original technique presented in [4]. After the completion of the k-means, unassigned pixels of the segmentation process are assigned using luminance criteria. As an example, Figure 3 has only needed four iterations to segment the motion field, starting from a list of affine models obtained from an original decomposition of the image in rectangular regions of 36×24 pixels. The motion information obtained in the first image is used as initial condition of the k-means algoritm in the following image, and so forth. Thanks to this projection process, the following images need only two iterations. Figure 4 and Figure 5 show the original sequence *Flower Garden* and its corresponding segmentation. We are currently working with better initial segmentations other than rectangular regions. In addition we have performed a series of experiments using an initial manual segmentation of the first image, that greatly improves the stability of the motion estimation process over time. This may show that for some applications, a semi-automatic procedure can be developed for the representation of moving images with layers.

4. RESULTS

The robust segmentation technique explained above provides a series of non-overlapped regions that will be used to form the final layers. Our layer synthesis approach follows very closely to the original one presented in [4]. To prove the feasibility of our approach, Figure 6 and Figure 7-b present the reconstruction of three consecutives frames of the *Flower Garden* sequence and the 15th frame of the *Mobile Calendar* sequence using the approach described above. Although it is not possible to fully appreciate in these still images the improvements of our method over the time domain, the reconstruction of the image it is believed to be of a very good quality. As a very basic example of object manipulation, Figure 7-c shows the same image with the ball removed. Please also notice the quality of the reconstructed image, specially in the location of the ball. All these sequences have been obtained using a rectangular initial partition. Improvements may be expected using more efficient initial segmentations.



a) Initial, $n = 0$. b) $n = 1$. c) $n = 3$. d) Post-processed after $n = 4$.

Figure 3. Results of the motion segmentation between frames #1 and #2 of *Flower Garden* after different iterations. White zones represent unassigned pixels

5. CONCLUSIONS

A robust representation of images using layers has been presented in this paper. To that end, robust motion estimation and segmentation techniques have been presented. Examples of image reconstruction and object manipulation have been provided which proves the feasibility of the approach.

REFERENCES

[1] ISO/IEC JTC1/SC29/WG11. MPEG-4 Proposal Package Description (PPD). July 1995.

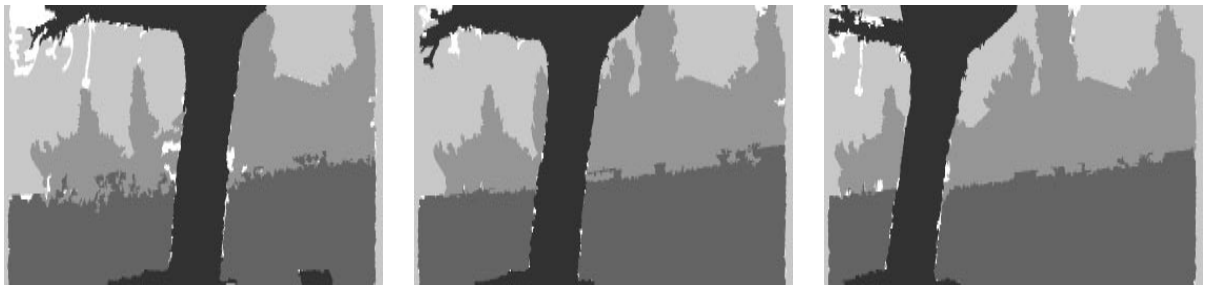
[2] F. Pereira. MPEG-4: a new challenge for the representation of audio-visual information. In *Picture Coding Symposium*, Melbourne, Australia, March 1996.

[3] L. Torres and M. Kunt. *Video coding: the second generation approach.* Kluwer Academic Publishers, Englewood Cliffs, 1996.

[4] J. Y. A. Wang and E. H. Adelson. Representing moving images with layers. *IEEE Transactions on Image Processing*, 3(5):625 – 638, September 1994.

a) #5 b) #15 c) #25

Figure 4. Original *Flower Garden Sequence*



a) #5 b) #15 c) #25

Figure 5. Segmentation of *Flower Garden*

a) #5 b) #15 c) #25

Figure 6. Reconstruction of *Flower Garden* **using the proposed approach**

a) Original b) Reconstruction c) Object manipulation

Figure 7. Reconstruction of *Mobile Calendar* **and object manipulation using layers**