Monochrome Video Segmentation

Sunder Venkateswaran, Ankur Mohan and U. B. Desai
Department of Electrical Engineering
Indian Institute of Tecnology Bombay
Bombay, 400076
Email: ubdesai@ee.iitb.ernet.in

Abstract

To provide mutimedia applications with new functionalities, such as content-based interactivity and scalability, the new video coding standard MPEG-4 relies on content based representation. This means a prior decomposition of a video sequence into semantically meaningful, physical objects. We formulate this problem as one of separating foreground objects from the background based on motion information. We present in detail an unsupervised video segmentation technique which uses the watershed transformation for spatial segmentation. We first give an outline of the algorithm and then proceed to explain each of the steps in detail.

1 Introduction

Data and signal modeling for images and video sequences is experiencing important developments. Part of this revolution is due to a need to support a large number of multimedia services. Traditionally, digital images were represented as rectangular arrays of pixels and digital video was seen as a flow of frames. New multimedia applications imply a representation that is closer to the real world. For example in the MPEG-4 standard [1], a video sequence is considered to consist of independently moving objects and is encoded object by object. In this case, the representation based on an array of pixels is not appropriate if one wants to interact with objects in the image. As a consequence, the data modeling has to be modified and has to include regions of arbitrary shapes to represent objects. In the following, we make a distinction between an object which is a 2-D representation of an entity which has a semantic meaning and a region which is a connected component of space defined by a homogeneity criterion.

A number of techniques and algorithms for video segmentation have been proposed, each having its own features and applications. In [2] and [3] frames in a video sequence are spatio-temporally segmented one after another, independent of each other. The approach presented in [4]–[6] deals with the sequence as a 3D (2D plus time) signal and, therefore, performs a 3D segmentation. Here the sequence is split into 3D blocks of a given number of frames and these 3D blocks are segmented. In [7], motion projection is used for temporal tracking. Motion projection may result in uncovered and overlapping regions. These regions are considered as new regions.

Our technique consists of two phases: initial segmentation and temporal tracking. For initial segmentation, Gaussian blurring followed by a watershed transformation is used to segment the first frame of the video sequence into homogeneous regions. Then motion parameters of a simplified linear model are estimated for each region, and regions with a coherent motion are merged to form a moving object. Temporal tracking is used to segment the subsequent frames of the video sequence. It consists of motion projection, marker extraction and the modified watershed transformation. The algorithms for this technique and the results obtained are discussed in detail in the following sections.

2 Initial Segmentation

The technique is initialized by the spatio-temporal segmentation of the first frame of the sequence. It consists of three phases: spatial segmentation of the first frame of the sequence into regions based on luminance, motion parameter estimation for each region and merging of the regions to form moving objects.

2.1 Spatial Segmentation

The watershed algorithm is used to spatially segment the first frame of a video sequence. It consists of the following steps:

1. *Image Simplification:* The image is first prefiltered using a Gaussian low-pass filter. This smoothens the image and reduces the noise in it.

2. *Gradient:* The gradient vector at each pixel of this smoothened image is computed using the *Sobel* operator. The magnitude of this gradient is computed and used for the watershed transformation.

3. *Watershed Transformation:* Watershed transformation [8] is performed on the magnitude of the gradient. Every local minima of the gradient leads to a region in the resulting segmentation.

4. *Region merging based on gradient:* After the watershed transformation, some regions need to be merged because of possible over segmentation. In real images, there are thousands of intensity minima and associated watershed regions. In this case, the image is over segmented and the problem is identifying which watershed boundaries mark significant image structures. To reduce the number of regions, we go through a region merging step where adjacent regions are merged according to a criterion based on the gradient along the border of adjacent regions. If the length of the common contour between two regions is more than 5 pixels and the average gradient along this common contour is less than a threshold T_1, the two regions are merged. (The average gradient is equal to the sum of the gradient at each contour pixel divided by the total number of contour pixels).

Fig. 1 shows the first frame of the table tennis sequence and the spatial segmentation of that frame (after Step 4). Here we used a zero mean Gaussian filter of spatial variance 1 for simplification. The threshold T_1 was set to 15.

2.2 Motion Estimation

Motion estimation is required for motion-based region merging and also for motion projection. The spatial segmentation step partitions the image into homogeneous regions based on luminance. A moving object may be composed of several such regions following a coherent motion. Since the objective of video segmentation is to separate and track moving objects, the regions must be merged according to their motion information to form moving objects. The motion of a region is described by a linear simplified model:

$$\begin{bmatrix} dx \\ dy \end{bmatrix} = \begin{bmatrix} \cos\theta - 1 & \sin\theta \\ -\sin\theta & \cos\theta - 1 \end{bmatrix} \begin{bmatrix} x_i \\ y_i \end{bmatrix} + \begin{bmatrix} T_x \\ T_y \end{bmatrix}$$

where (dx, dy) is the motion vectors at position (x_i, y_i), θ is the rotation angle and T_x and T_y are the translations in the x and y directions.

The motion parameters are estimated using the *indirect parametric motion estimation method* [9]. In this method a dense optical field is first estimated using the Horn-Schunck method [10]. Based on the above estimated optical flow and using the spatial segments obtained from the algorithm of Section 2.2, the motion parameters of the respective regions can be computed. For this purpose a least mean square technique is used [11].

2.3 Region merging based on motion

Adjacent regions with coherent motion should be merged together to form moving objects. The similarity between two regions in terms of motion is measured by the increment of the mean-square motion compensation error. Region merging is realized in the following steps.

1. For every region R_i, motion parameters are estimated. The sum of square compensation error E_i is also calculated for each region. E_i is given by

$$E_i = \sum_j \left[(d_x - u_j)^2 + (d_y - v_j)^2 \right]$$

   Here j ranges over all the pixels (x_j, y_j) in the region R_i. u_j and v_j are the x and y components of the flow vector at pixel (x_j, y_j) estimated using the Horn-Schunck method. (d_x, d_y) is the motion vector at pixel (x_j, y_j) computed using the motion parameter model.

2. For every two adjacent regions R_i and R_j, a set of motion parameter is estimated for both of the regions combined together $(R_i \cup R_j)$. Let $E_{i,j}$ denote the sum of square compensation errors of the two regions when they are compensated using this set of motion parameters. The increment of mean square motion compensation error is calculated by $\Delta_{i,j} = (E_{i,j} - E_i - E_j)/(N_i + N_j)$, where N_i and N_j are the number of pixels contained in regions R_i and R_j, respectively.

3. If the value of $\Delta_{i,j}$ is smaller than a predefined threshold T_2 , then the corresponding regions are merged.

4. Update all of the E_i, $\Delta_{i,j}$, and N_i which are related to the merged regions. Go to Step 3 for merging other regions until every $\Delta_{i,j}$ is greater than T_2.

Fig. 2 shows the result of motion-based region merging for the spatial segmentation shown in Fig. 1. Here motion parameters are estimated with respect to the second frame of the sequence. In this example the threshold T_2 was set to 0.0005.

3 Tracking

Tracking in the temporal domain is performed after the spatio-temporal segmentation of the first frame, to segment the subsequent frames of the video sequence. It consists of four steps: Motion projection, extraction of markers, modified watershed transformation and region merging.

First the segmented objects in one frame are projected according to the estimated motion parameters into the next frame. This establishes a correspondence of moving objects between frames. As the projections do not form a complete and accurate segmentation of the next frame, pertinent parts of these projections are extracted as markers. The final segmentation is obtained from these markers and the modified watershed transformation.

3.1 Motion Projection

Let $f(x, y, t)$ and $f(x, y, t+1)$ denote two consecutive frames of a video sequence at times t and $t+1$ respectively. Suppose $f(x, y, t)$ has been segmented into n moving objects O_i, for $0 < i \leq n$, and motion parameters for each moving object have been estimated with respect to $f(x, y, t+1)$. Then, the correspondence of O_i in frame $f(x, y, t+1)$ can be roughly obtained by projecting O_i in frame $f(x, y, t+1)$ according to the motion information. Let (dx_i, dy_i) denote the motion field within moving object O_i, which is generated from the estimated motion parameters. The projection of O_i into $f(x, y, t+1)$ is described by:

$$P_i = \{(x + dx_i, y + dy_i) \mid (x, y) \in O_i\} \qquad (1)$$

The union of all of the projections P_i, for $0 < i \leq n$, may not cover the whole frame $t+1$, and one projection may overlap another due to occlusions between frames. Uncovered areas are segmented using the modified watershed transformation. The correspondence of overlapping areas is ambiguous. To solve this problem, projection error of intensity are used. Equation (1) indicates that pixel at (x, y, t) corresponds to pixel at $(x + dx, y + dy, t + 1)$ according to estimated motion parameters. If pixel at (x, y, t) really moves to pixel at $(x + dx, y + dy, t + 1)$, the projection error of intensity $|f(x, y, t) - f(x + dx, y + dy, t + 1)|$ must be very small. Hence, if two pixels $p_1(t)$ and $p_2(t)$ in frame t are projected to pixel $p(t + 1)$ in frame $t + 1$, the one which has smaller projection error is selected as the correspondence of $p(t + 1)$, and the other one is considered as a false correspondence.

3.2 Extraction of Markers

A marker is a set of pixels labelling an object. From the marker, the watershed transformation can locate the boundary of the object. Hence, a marker of an object should be completely included in the object. If a projected marker is completely included in the region of the corresponding moving object, it is a good marker for this object.

There are three cases in which the projected object should not be directly used as a marker. The first is that most of the projected object is included in the region of the corresponding moving object, but a small portion is protruding out. The second one is that the projected object significantly overlaps a new object which did not exist in frame t. In the third case, the moving object of frame t, exits from frame $t + 1$, so that the projected object is falsely included in the region of other moving objects. Considering the first case, a marker should be the internal area of the projected object. The area near the boundaries of the projected object should not be used as marker. In the second case two situations are possible. If the pixel at (x, y, t) correctly projects on to pixel at $(x+dx, y+dy, t+1)$, then the projection error $|f(x, y, t) - f(x+dx, y+dy, t+1)|$ would be small. On the other hand if pixel at (x, y, t) projects to a pixel, in frame $t + 1$, which corresponds to a new object, then the projection error would be large. In case three, since the object of frame t, has exited, the projection error would always be large. Hence the disturbance of new objects and disappearing objects can be eliminated by thresholding projection errors. Markers are therefore extracted in the following manner:

1) The interior area I_i of projection P_i is obtained by eroding P_i with a square structuring element of 5×5 pixels.

2) Compare the projection error of every pixel in I_i with a threshold T_3. The pixels with a projection error smaller than T_3 are taken as markers of the moving object.

3.3 Modified Watershed Transformation

After marker extraction, the modified watershed transformation is used to segment the $t + 1$ frame. The extracted markers are imposed as minima on the gradient image and used as initial catchment basins, irrespective of the gradient values within the area of the markers. The gradient minima in other areas are not suppressed. The flooding operation is now performed in a similar manner to the original watershed transformation.

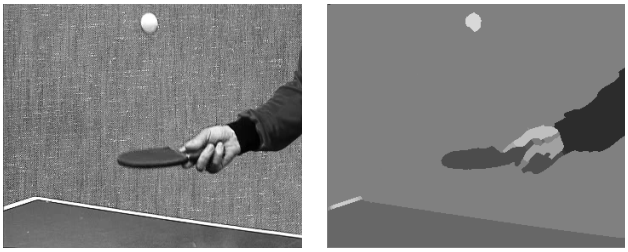
3.4 Region Merging

After the modified watershed transformation the resulting regions are merged according to motion information. This is because motion is not taken into account in case of new regions and new regions may be parts of a new object.

4 Results and Discussion

In this section the results obtained by using the algorithm with the table-tennis video sequence are discussed.

As explained earlier, directly applying the watershed algorithm to the input image results in a highly over-segmented image. To reduce the number of segments, we pass the first frame through a Gaussian low-pass filter of mean (μ) = 1 and standard deviation (σ) = 1. (σ^2 is the variance in the spatial domain, variance in the frequency domain will be scaled by a suitable factor). Fig. 1 shows the first frame of the table tennis sequence and the corresponding segmented frame.



(a) (b)

Figure 1: (a) The first frame of the tennis video sequence. (b) The spatial segmentation of this frame.

Once we have the segments for the first frame, we use the motion information between the first frame and the second frame to merge those regions which move in the same way. After this step, the resulting segments are semantically meaningful, i.e. the segments represent physical objects in the image. The results obtained on region merging between the first and the second frames are shown below. As described in the previous sections, we have used the Horn-Schunck method for motion vector estimation and a least square method for motion parameter estimation. Fig. 2 shows the result of motion based region merging for the segmentation of Fig. 1. Here, threshold T_2 was set to 0.0005.

Temporal tracking is performed after the first frame of the video sequence has been segmented into moving objects. The objective of temporal tracking is to segment the subsequent frames of the sequence and to establish a correspondence of moving objects between frames. Temporal tracking is performed in three steps:

1. Motion projection
2. Marker extraction

Figure 2: Result of motion based region merging for the first frame

3. Modified watershed transformation
4. Region merging.

We present below the marker image and the results obtained using the modified watershed transformation.

Figure 3: The marker image with the background as one of the markers.

In Fig. 4 since the background is one of the markers and the table is part of the background, the table does not appear in the segmented image. To remedy the problem, we eliminate the largest region as one of the markers. The largest region is most probably the background region and this will allow the segmented image to have objects in the background as segments.

It can be seen in the above images that the table tennis ball appears in its correct position. This is because only those pixels are taken as markers which satisfy the projection error criteria.

The segmentation of six frames of the tennis video sequence obtained by the above technique is depicted in Fig. 6. In the segmented video sequence, segments are demarcated by white boundaries. Notice that even though the ball moves significantly between frames it

Figure 4: The segmented image using the modified watershed algorithm.

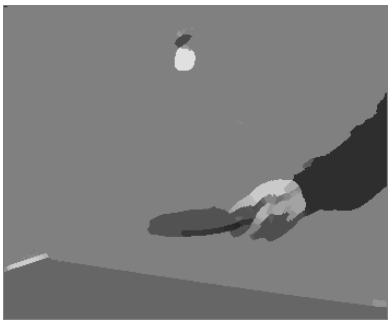


Figure 5: The table now appears as a separate segment when we no longer take the background as one of the markers.

is tracked correctly. Due to Gaussian blurring the background which has a noisy texture and several local minima appears as one segment. Also the racket is segmented as one object as it follows the same motion.

Fig. 7 shows the segmentation and tracking results for the "Claire" video sequence.

References

[1] MPEG-4 Video Group, "MPEG-4 video verification model version 7.0", *ISO/IEC JTC1/SC29/WG11, MPEG97/N1642*, Bristol, England, April 1997.

[2] F. Dufax, F. Moscheni, and A. Lippman, "Spatio-temporal segmentation based on motion and static segmentation," in *Proc. International Conf. on Image Processing,* Washington DC, vol. 1, pp. 306–309, Oct. 1995.

[3] J. G. Choi, S. W. Lee, and S. D. Kim, "Spatio-temporal segmentation using a joint similarity measure," *IEEE Trans. Circuits and Sys. for Video Tech.*, vol. 7, pp. 279–286, April 1997.

[4] M. Pardas and P. Salembier, "3D morphological segmentation and motion estimation for image sequences," *Signal Processing*, vol. 38, pp. 31–43, Sept. 1994.

[5] S. Rajala, M. Civanlar, and W. Lee, "Video data compression using three dimensional segmentation based on HVS properties," *Proc. International Conf. on Acoustics, Speech and Signal Processing*, New York, USA, pp. 1092–1095, 1988.

[6] P. Salembier and M. Pardas, "Hierarchical morphological segmentation for image sequence coding," *IEEE Trans. on Image Processing*, vol. 3, pp. 639–651, Sept. 1994.

[7] Y. Yokoyama, Y. Miyamoto, and M. Ohta, "Very low bit rate video coding using arbitrary shaped region-based motion compensation," *IEEE Trans. Circuits and Sys. for Video Tech.*, vol. 5, pp. 500–507, Dec. 1995.

[8] L. Vincent and P. Soille, " Watersheds in digital spaces: An efficient algorithm based on immersion simulations." *IEEE Trans. Pattern Analysis and Machine Intell.*, vol. 13, pp. 583–598, 1991.

[9] L. Torres and M. Kunt, *Video Coding: The Second Generation Approach*, pp. 232–234, Kluwer Academic Publishers, 1996.

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

[11] K. S. Arun, T. S. Huang and S. D. Blostein, "Least-Square Fitting of Two 3-D Point Sets", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 9, No. 5, pp. 698–700 September 1987.

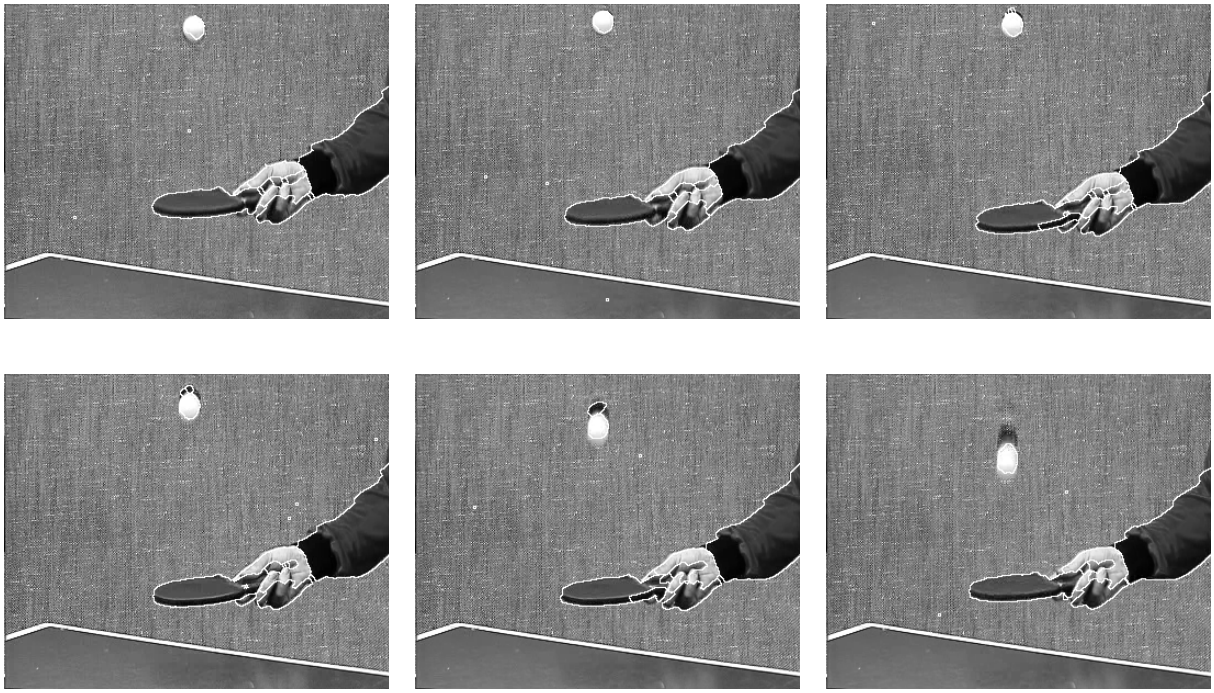


Figure 6: The segmented frames of the tennis video sequence

Figure 7: Result of segmentation for the “Claire” video sequence