Robust 2D Moving Object Segmentation and Tracking in Video Sequences

VASILE GUI, FLORIN ALEXA, CATALIN CALEANU, DANIELA FUIOREA
Department of Electronics and Telecommunications
„Politehnica” University Timisoara, Romania
Bd. V. Parvan No. 2, 30223 Timisoara, Romania

Abstract: − Unsupervised motion segmentation and tracking in video sequences is a complex task, requiring robust estimation and flexible modeling. The paper presents an unsupervised method of moving object segmentation and tracking in video sequences captured by static cameras. Central to our work is the nonparametric density estimation and the mean shift algorithm for finding local maxima of the probability density. Foreground segmentation obtained from background estimation is combined with simultaneous region tracking and segmentation followed by connectivity-based moving object segmentation, in order to obtain an efficient processing algorithm. Preliminary tests asses the viability of the proposed approach.

Keywords — Video segmentation, moving object tracking, nonparametric kernel density estimation.

1 Introduction

Motion analysis has become more and more important in many applications such as video surveillance, human-computer intelligent interface, content-based image sequence representation or video conferencing. All require the ability to segment and track moving objects in complex environments. Motion information may be needed to further characterize undergoing activities in the scene, like in the surveillance or human computer interface applications. Also, motion information can be extracted prior to segmentation and tracking, in order to facilitate these processes. Video segmentation, object tracking and motion estimation are actually closely related problems.

The most direct approach to motion estimation is based on the computation the optical flow [1]. The advantage of the optical flow approach is that it generates a dense motion field. However, a major problem with optical flow computation is its high sensitivity to noise. To alleviate it, edge preserving filtering methods can be used. Alternatively, optical flow can be obtained from segmented images, as recently proposed by Zitnick et al. [2].

Motion estimation is particularly difficult in flat image regions, since motion does not affect the observed intensity or color vectors in such regions. This is why an important trend in motion estimation and tracking is based on the detection of a relatively small number of salient feature points in successive frames and to match them. A major advantage of point based motion analysis techniques is the potential invariance to shades, intensity or color changes. Edges, corners and local orientation are among the most used features for point based motion analysis, object tracking or registration [3]. Point matching algorithms should work well in the presence of noise and geometric distortions. Consequently, a lot of work has been directed toward robust solutions, which can tolerate high percentage of outliers in the set of matched points. See for example [4][5].

Point based motion estimation produces a sparse motion information, needing further processing. By contrast, segmentation based solutions generates the complete motion field. A good segmentation produces regions with sharp borders, avoiding the problem of correspondence ambiguity in flat image regions. After solving the region correspondence problem, motion information can be extracted from such correspondences. In the present paper we follow this approach. The quality of the segmentation-based motion estimation depends on the quality of segmentation [6] as well as on the method used to extract the motion parameters from pairs of regions. In this work, we use a robust mean shift based tool for segmenting and tracking both the background and the moving objects.

2 Brief review of the mean shift algorithm

Given a sample of N d-dimensional data points, $\mathbf{x}_i$, drawn from a distribution with multivariate probability density function $p(\mathbf{x})$, an estimate of this density at $\mathbf{x}$ can be written as [7]:

$$\hat{p}_{\mathbf{H}}(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N} K_{\mathbf{H}}(\mathbf{x}-\mathbf{x}_i) \qquad (1)$$

where

$$K_{\mathbf{H}}(\mathbf{x}) = |\,\mathbf{H}\,|^{-1/2}\ K\left(\mathbf{H}^{-1/2}\mathbf{x}\right) \qquad (2)$$

is the kernel function depending on a symmetric positive definite $d\times d$ matrix, called bandwidth matrix. Frequently $\mathbf{H}$ has a diagonal form or even the form $\mathbf{H}= h^2\mathbf{I}$, assuming the same scale h for all dimensions, i.e. a single scale parameter and an isotropic estimator, K_h.

An efficient way to find local maxima of the estimated PDF is through the mean shift algorithm. Given the PDF estimated with the radial symmetric kernel K with profile k,

$$\hat{p}_{h,K}(\mathbf{x}) = \frac{c_{k,d}}{nh^d}\sum_{i=1}^{n} k\left(\left\|\frac{\mathbf{x}-\mathbf{x}_i}{h}\right\|^2\right) \quad , \qquad (3)$$

the mean shift vector is proportional to the normalized gradient of the estimated PDF
and is found to be:

$$\mathbf{m}_{h,G}(\mathbf{x}) = \frac{\sum_{i=1}^{n}\mathbf{x}_i g\left(\left\|\frac{\mathbf{x}-\mathbf{x}_i}{h}\right\|^2\right)}{\sum_{i=1}^{n} g\left(\left\|\frac{\mathbf{x}-\mathbf{x}_i}{h}\right\|^2\right)} - \mathbf{x}\,, \qquad (4)$$

with

$$g(x) = -k'(x)\,. \qquad (5)$$

The mean shift vector points into the direction of the maximum increase of the PDF, estimated with kernel K. Using the mean shift vector at a location $\mathbf{y}$, a gradient ascent algorithm can be used to find the location of the maxima of the estimated PDF closest to the starting location. This can be simply done by iterating the equation

$$\mathbf{y}_{j+1} = \frac{\sum_{i=1}^{n}\mathbf{x}_i g\left(\left\|\frac{\mathbf{y}_j-\mathbf{x}_i}{h}\right\|^2\right)}{\sum_{i=1}^{n} g\left(\left\|\frac{\mathbf{y}_j-\mathbf{x}_i}{h}\right\|^2\right)}, \quad j = 1,2,... \qquad (6)$$

until convergence is obtained.

3 Motion segmentation and tracking

A block diagram of the proposed motion segmentation and tracking method is shown in figure 1. In the present work, the video stream is supposed to be generated by a fixed camera, facilitating efficient background estimation. A generalization of the method for moving or PZT cameras is possible, by solving the dominant motion estimation and compensation problem first.

3.1 Background estimation

Background modeling is commonly carried out at pixel level. Each pixel is represented by a feature vector, such as intensity or color, disparity, depth etc. In the present work, we use only color information. A suitable way to model the static background is through a random vector with an associated probability density function (PDF). The background, $\mathbf{b}$, at a pixel is then defined as

$$\mathbf{b} = \arg\max_{\mathbf{x}}\ \hat{p}_{h,K}(\mathbf{x}) \qquad (7)$$

that is the feature vector (color) maximizing the PDF estimated from a set of N frames. In our work, the nonparametric density estimation technique first proposed by Elgamal [8] is used, with some improvements of the method, as reported in [9][10][11]. A recursive, mode tracking algorithm is used in order to reduce the computational complexity from O(2N), corresponding to Fast Gauss Transform algorithms [12][13], to O(N^0), that is independent on the size of the frame buffer. Also reduced sensitivity to noise and fast adaptation to background changes were obtained, by using an adaptive learning rate, based on a cumulative error measure.

3.2 Background segmentation

Background segmentation can be viewed as a two class discrimination problem. Initially, background subtraction is obtained by thresholding the background probability density estimates. This is equivalent to the color similarity test,

$$\| \mathbf{x}_i - \mathbf{y}_b \| < th_s \ . \qquad (8)$$

When motion information becomes available, foreground object model can be used to obtain a better foreground/background segmentation by comparing the estimated probability densities of each pixel color vector in its background and foreground distributions.

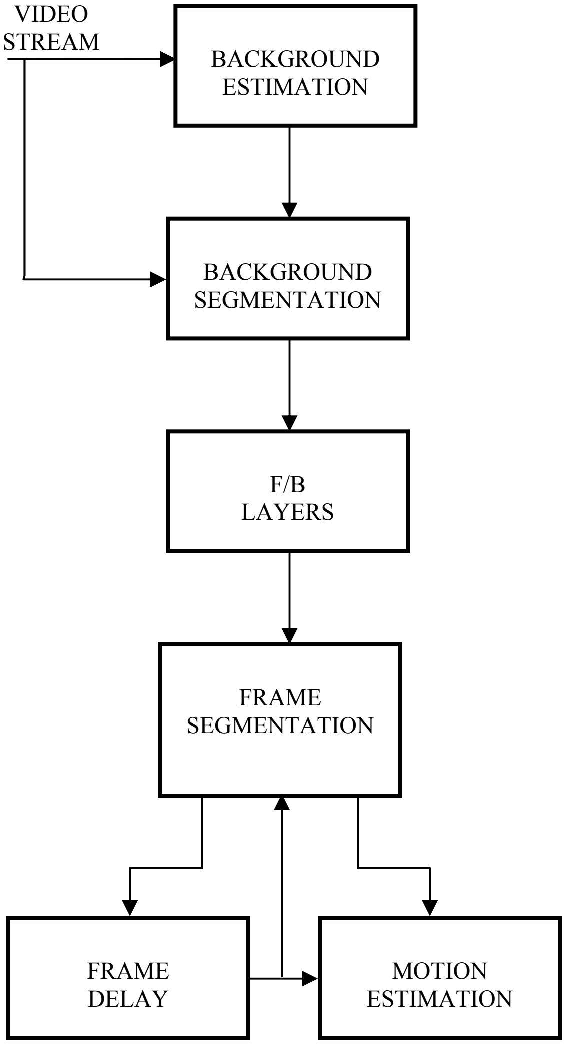

Fig. 1. Block diagram of the motion segmentation and tracking system

3.3 Frame segmentation

The first frame of a video shot is segmented by a mean shift clustering method described in [14]. Our data vector is formed by the space coordinates of the pixels, or the domain information and the color of pixels (range) information and the $L{*}u{*}v{*}$ color space:

$$\mathbf{x} = [x, y, L*, u*, v*]^T \ . \qquad (9)$$

Since space and range data have different scales, we use renormalize them prior to clustering and then use a single scale, h as in equation (6). Mean shift segmentation consists of the following steps:

Step1. Starting from each pixel, $\mathbf{x}$, make $\mathbf{y}_0$=$\mathbf{x}$ and iterate equation (6) to find its closest mode in the PDF.
Step2. Group all pixels that converged to the same mode in one region.
Step3. Merge all adjacent regions with mode differences smaller than a certain threshold.
Optionally, very small regions can be merged to the most similar adjacent region. In this work, we did not merge small regions, as they can provide very useful motion information.

As a result of the intra frame segmentation, each pixel is assigned a label, k, representing the index of the segment/region R_k containing it. Additionally, pixels from this frame are classified and labeled as foreground or background pixels (F/B), based on similarity to the estimated background. Note that pixels have two labels: a regional label and a foreground/background label.

Starting from the second frame, for each pixel data, $\mathbf{x}_i$, the best match in a search window centered on the pixel's location is found in the previous frame $\mathbf{y}_c$. Again, the mean shift equation (6) is iterated to match the pixel to a mode from the previous frame, $\mathbf{y}_c$. The color similarity test between $\mathbf{x}_i$ and $\mathbf{y}_c$ given by equation (8) is used to asses the match. If the test is passed, the pixel is assigned the same labels as $\mathbf{y}_c$. Otherwise, the pixel belongs to a new region and remains initially unlabeled. Generally, most pixels pass the similarity test, actually being simultaneously tracked and labeled. Moreover, the correspondence problem is also solved and the computational burden is kept at a very low level. For the rest of the pixels, the same 2D mean shift segmentation is used as for the initial frame and new region labels are assigned. Those pixels are assigned as foreground pixels, and newly appearing objects if they do not match the background estimated at their location. If they do, they belong to uncovered background and classified correspondingly. Previous frame foreground regions which have not found correspondence in the current frame are marked as occluded.

3.4 Object segmentation, tracking and motion estimation

Objects like people in surveillance sequences are deformable and perform complex movements. Arms, legs, head and shoulders do not move in the same way. To segment correctly such objects, very general region grouping criteria have to be chosen. Ultimately, the rules need to concentrate on the most stable feature of an object: it cannot be split into several pieces. By contrast, different objects may join temporarily. For example, a person may take a bag, carry it and then place it somewhere in the background.

To obtain object segmentation, regions are merged according to the following rules [15]:

Rule 1: The regions have to be labeled as belonging to the class of foreground regions.

Rule 2: The regions have to be adjacent in all frames where they co-exist.

Foreground region motions are modeled as piecewise constant within segmented regions. A region, R_k, is represented by its PDF mode vector, $\mathbf{y}_k$, estimated from all its samples. If the region R_k from the current frame is matched with region R_p from the previous frame, its motion vector is defined as the difference vector of the spatial components of the mode vectors.

4 Results and discussion

Results obtained with our video object segmentation and tracking method are illustrated bellow. Figure 2 illustrates stages of the segmentation of a video sequence. Frame 72 from the original image is shown in figure 2a). The foreground/background segmentation mask is shown in figure 2b). In figure 2c, results of the spatial segmentation of the foreground image with marked borders can be seen, while in figure 2d) the foreground object extracted for the same frame, after object segmentation, is represented.

We tested the proposed object segmentation and tracking method on several images. The connectivity condition and spatial position of the object proved to be reliable for video object segmentation and tracking. Experiments demonstrated better performance on images which can be well represented by regions with piecewise constant models. In such images, the mean shift segmentation is remarkably stable. The simultaneous tracking and segmentation approach was adopted with the main purpose to further enhance the stability of the frame segmentation. However, the problem is not completely solved, as it can be seen from the next example.

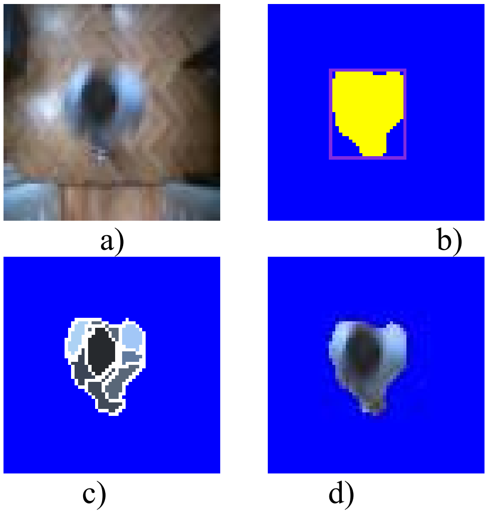

Fig. 2. a) Input image frame; b) foreground/background segmentation mask with bounding rectangle; c) region segmentation; d) object segmentation

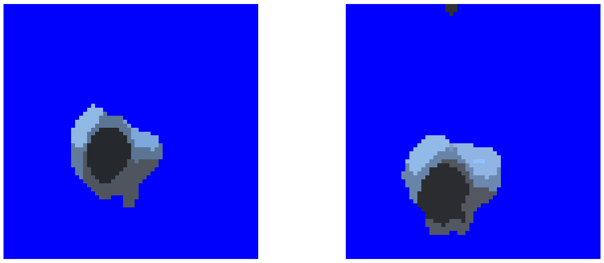

Figure 3. Results of frame segmentation for two consecutive frames

Figure 3 Shows results of frame segmentation in two consecutive frames of the same video sequence. Although all regions from the first frame have found

correspondence in the second one, the shapes of some regions change significantly from frame to frame. This is a result of several interacting factors, like self shadows, reflections and motion blur. Region shape deformations affect the accuracy of the region based motion estimation. We believe that estimating region motions from density modes of coordinates instead of simple mean (centroid) coordinates result in less sensitivity to shape deformation, as pixels less typical for a region contribute less in the estimation. Quantitative assessment of this fact remains a subject of our future work.

5 Conclusions

In this work, we presented a method to segment and track moving objects in video sequences. The main tool is the mean shift, a robust clustering method funded on nonparametric density estimation. By embeding the intra frame segmentation within a tracking framework, efficient implementation is obtained. Connectivity based object segmentation ensures high flexibility in modeling appearance changing objects.

References

1. S.S. Beauchemin and J.L. Barron, “The computation of optical flow”. *ACM Computing Surveys,* 27(3), 1995, pp 433-467.
2. C.L. Zitnick, N. Jojic and S.B. Kang, “Consistent segmentation for optical flow estimation”. In ICCV 2005, pp 1308-1315.
3. J. Shi and C. Tomasi, “Good features to track”. In Proc. CVPR 1994, pp 593-600.
4. M. A. Fischler and R. C. Bolles, “Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography”. *Comm. Assoc. Comp.Mach*, 24(6):381–395, 1981.
5. P. Meer, “Robust techniques for computer vision”, *Emerging Topics in Computer Vision*, G. Medioni and S. B. Kang (Eds.), Prentice Hall, 2004, pp. 107-190.
6. Y-J. Zhang, Ed. “*Advances in Image and Video Segmentation*”. IRM Press, Hershey, PA, 2006.
7. D.Comaniciu, P.Meer, “Mean shift: A robust approach toward feature space analysis”, *IEEE Trans. Pattern Anal. Machine Intell,* Vol. 24, No.5, 2002, pp. 603-619.
8. A. Elgamal, R. Duraiswami, D. Harwood, L. Davis, “Background and foreground modeling using nonparametric kernel density estimation for visual surveillance”, Proceedings of the IEEE, Vol. 90, No.7, 2002, pp 1151-1162.
9. C. Ianăşi, V. Gui, C.I. Toma, D. Pescaru, “A fast algorithm for background tracking in video surveillance using nonparametric kernel density estimation“, *Facta Universitatis (Niš)*, Vol. 18, No. 1, 2005, pp 127-144.
10. C. Ianăşi, V. Gui, F. Alexa, C.I. Toma, „Fast and Accurate Background Subtraction for Video Surveillance, Using an Adaptive Mode –Tracking Algorithm”. *Proceedings WSEAS Conference on Dynamical Systems and Control*, Venice 2005.
11. C. Ianăşi, V. Gui, F. Alexa, C.I. Toma, “Noncausal adaptive mode-tracking estimation for background subtraction in video surveillance”, WSEAS Transactions on Signal Processing Issue 1, Vol, 2, Jannuary 2006, pp 52-56, ISSN 1790-5022.
12. A.Elgamal, R.Duraiswami, L.S.Davis, “Efficient kernel density estimation using the Fast Gauss Transform with applications to color modeling and tracking”, *IEEE Trans. Pattern Anal. Machine Intell*. Vol. 25, No. 11, 2003, pp. 1499-1504.
13. J. Yang, R. Duraiswami, N. Gumerov, L. Davis, “Improved Fast Gauss Transform for efficient kernel density estimation”, *IEEE Intl. Conference on Computer Vision, ICCV*, 2003, pp. 464-471.
14. Y.H. Gu and V. Gui, “Joint space-time-range mean shift-based image and video segmentation”. Invited paper in Y-J. Zhang, Ed. “*Advances in Image and Video Segmentation*” IRM Press, Hershey, PA, 2006.
15. V. Mezaris, I. Kompatsiaris, M.G. Strintzis, “Video object segmentation using Bayes-based temporal tracking and trajectory-based region merging”. *IEEE Trans. on Circuits and Systems for Video Technology*, Vol. 14, No. 6, pp 782-795, June 2004.