Motion and Segmentation Prediction in Image Sequences Based on Moving Object Tracking

Adrian G. Borş Ioannis Pitas
Department of Informatics
University of Thessaloniki
Thessaloniki 540 06, Greece
E-mail: {adrian, pitas}@zeus.csd.auth.gr

Abstract

The image sequence is represented as a set of moving regions which make up moving objects. Motion, position and graylevel (or color) information is used for segmenting the moving objects. A criterion is proposed for modeling the 3-D motion and segmentation. After identifying the occluding regions, the moving objects are tracked over the next frames. Prediction is employed for estimating the future moving object position and its optical flow.

1 Introduction

Various approaches have been proposed for optical flow estimation and motion segmentation [1]. The maximization of the *a posteriori* probability has been considered in [2]. A classification approach was proposed for jointly segmenting the moving objects and their corresponding optical flow in [3]. Median Radial Basis Function (MRBF) algorithm which relies on robust statistics was employed for estimating the moving object characteristic vectors [3].

We provide a classification based criterion for the 3-D segmentation of the image sequence. After estimating the boundaries of the moving objects and their corresponding optical flow in 2-D we want to see how they change over the next frames [1]. Tracking algorithms are usually employed for estimating the moving object position in the following frames [4, 5]. Certain regions do not have a clear match in the previous frame [1, 4]. After finding the undetermined regions, their component pixels are classified by means of the already trained MRBF network. Consequently, the moving objects are matched from one frame to the next one. After estimating the 3-D description of the moving sequence, we can derive a prediction model. This model assumes that object velocity depends on the velocities of the same moving object in the previous frames. The predicted image, is compared with the real frame (if it is available) in order to certify the accuracy of the 3-D model.

2 Spatio-temporal motion and segmentation estimation

An image sequence $f(t)$, $t = 1, \ldots, K$ is composed of a set of moving regions $\{X_i(t), i = 1, \ldots, N\}$ with the properties :

$$f(t) = \cup_{i=1}^{N} X_i(t) \tag{1}$$

$$X_j(t) \cap X_k(t) = 0, \forall j \neq k \tag{2}$$

Each subset $X_k(t)$ is associated to a five-dimensional representative vector $\mu_k = [\mathcal{S}_k, \mathcal{M}_k]$, describing the optical flow $\mathcal{M}_k$ and segmentation information $\mathcal{S}_k$ associated with a certain moving region. The still image feature vector $\mathcal{S}_k$ contains the location and the characteristic graylevel. $\mathcal{S}_k$ is directly related to the segmentation label of the moving region k while $\mathcal{M}_k$ represents the velocity of the respective moving region. The classification in moving regions is done according to the Bayesian theory :

$$P(\hat{\mu}_k(t), t = 1, \ldots, K-1 | f(t), t = 1, \ldots, K) > P(\hat{\mu}_j(t), t = 1, \ldots, K-1 | f(t), t = 1, \ldots, K). \tag{3}$$

After applying the Bayes theorem and expressing the probabilities from one frame with respect to the previous frames, we obtain :

$$\begin{aligned} &P(\hat{\mu}_k(t), t = 1, \ldots, K-1 | f(t), t = 1, \ldots, K) = \\ &\ldots = \Pi_{t=p}^{K} [P(f(t) | \hat{\mu}_k(j), f(j), j = 1, \ldots, t-1] \\ &\Pi_{t=p}^{K} [P(\hat{\mu}_k(t-1) | \hat{\mu}_k(j), f(j), j = 1, \ldots, t-2, \\ &f(t-1)] \frac{P(\hat{\mu}_k(j), j = 1, \ldots, p-1 | f(j), j = 1, \ldots, p)}{\Pi_{t=p}^{K} P(f(t) | f(j), j = 1, \ldots, t-1)} \end{aligned} \tag{4}$$

The first probability in the derived expression is associated with the reconstruction of a frame based on the previous $t-1$ frames and their correspondent feature vectors. The second probability is associated with the dependence of the feature vector at the moment $t-1$ with respect to the values of the same feature vector at the previous moments. This probability represents the tracking of the feature vectors over several frames. The third probability models the moving object characteristics evaluated in the first p frames. The denominator represents the dependency of a certain frame on the previous ones and can be neglected.

The image is split in blocks and a feature vector denoted as $\mathbf{u}_{IJ}$ containing the location, the graylevel and the motion vector is associated with each block at site (I,J). The third probability from (4), when considering $p=2$, can be further decomposed :

$$P(\hat{\mathcal{M}}_j,\hat{\mathcal{S}}_j|f(t),f(t-1)) = P(f(t)|\hat{\mathcal{M}}_j,\hat{\mathcal{S}}_j,f(t-1)) \cdot\frac{P(\hat{\mathcal{M}}_j|\hat{\mathcal{S}}_j,f(t-1))P(\hat{\mathcal{S}}_j|f(t-1))}{P(f(t)|f(t-1))} \quad (5)$$

where $P(\hat{\mathcal{S}}_j|f(t-1))$ represents the *a priori* probability of the segmentation and $P(\hat{\mathcal{M}}_j|\hat{\mathcal{S}}_j,f(t-1))$ is the probability of the optical flow estimation depending on the segmentation map and image [2]. After expressing each probability as an energy function, we model these probabilities with Gaussian functions. The Gaussian function associated with a moving region and implemented by a hidden unit of the RBF network is :

$$\phi_j(\mathbf{u}_{IJ}) = \exp\Big[-(\mathbf{u}_{IJ}-\hat{\mu}_j)^T\hat{\mathbf{\Sigma}}_j^{-1}(\mathbf{u}_{IJ}-\hat{\mu}_j) -WDFD(\hat{\mathcal{M}}_j)\Big] \quad (6)$$

where $\hat{\mu}_j$ and $\hat{\mathbf{\Sigma}}_j$ are the estimates of the center vector and covariance matrix of the Gaussian function and $WDFD(\hat{\mathcal{M}}_j)$ represents the weighted displaced frame difference - a measure of confidence in the motion estimation algorithm [3]. The output layer function consists of a weighted sum of hidden-unit outputs, scaled to the interval $(0,1)$ by a sigmoidal function :

$$Y_k(\mathbf{u}_{IJ}) = \frac{1}{1+\exp\Big[-\sum_{j=1}^{L}\lambda_{kj}\phi_j(\mathbf{u}_{IJ})\Big]} \quad (7)$$

for $k=1,\ldots,N$, where λ_{kj} is the parameter associated to the connection between the hidden unit j and the output unit k, L is the number of basis functions and N is the number of moving objects. The training of the basis function parameters is done using an outlier rejection algorithm based on the MRBF algorithm [3]. The number of moving objects is not known *apriori* and is determined according to a compactness criterion. The output units associate moving regions with their corresponding moving objects. The training algorithm is unsupervised and provides an estimate for $P(\hat{\mu}_k(j),j=1,\ldots,p-1|f(j),j=1,\ldots,p)$ [3].

3 Tracking the moving objects

A component of the first product in the decomposition from (4) represents the reconstitution of the current frame based on the previous image segmentation. For $p=2$ we can express this probability as an energy function measuring the accuracy of displacing the moving regions according to their movement :

$$P(f(t)|\hat{\mu}_k(t-1),f(t-1)) \simeq \exp[-|\cup_i^N (X_i(t-1)\oplus\hat{\mathcal{M}}_i(t-1))-f(t)|] \quad (8)$$

where $X_i(t-1)\oplus\hat{\mathcal{M}}_i(t-1)$ represents the translation of the moving region $X_i(t-1)$, obtained from the segmentation of the frame $f(t-1)$, with the motion vector $\hat{\mathcal{M}}_i(t-1)$. The maximization of this probability represents the minimization of the difference between the given frame and the frame reconstructed based on the previous frame, its segmentation and correspondent optical flow. It can be observed that by displacing each individual object $i=1,...,N$, from $f(t-1)$ (8), certain regions have uncertain assignment. The condition from (1) is not any more respected when regions from one frame do not have a correspondent in the next frame (uncovered regions) and (2) is not validated when two different objects project in the same region of the next frame (occluding regions). Both situations occur in the regions from the margins of the objects. After detecting these particular regions we extract their correspondent feature vectors $\mathbf{u}_{IJ}$ associated to their location, graylevel and displacement. The trained MRBF network, can be used at the pixel resolution in a multiresolution approach [3]. We apply the already trained MRBF network only in the regions decided as uncertain according to (8).

The components of the second product of the expression (4) representing the dependency of a feature vector on the values of the same future vector in the previous frames, can be expressed as an energy function :

$$P(\hat{\mu}_k(t-1)|\hat{\mu}_k(j),f(j),j=1,\ldots,t-2,f(t-1)) = \exp[-(\hat{\mu}_k(t-1)-\sum_{i=1}^{H}W_k\psi(\hat{\mu}_k))] \quad (9)$$

where $\psi(\hat{\mu}_k)$ are a set of functions modeling the variation of the k-th object feature vector in time and

W_k are their associated weights. RBF functions, similar with that from (6) but having as inputs the feature vectors at various time intervals, can be used for modeling complex variations as those produced by chaotic time series and they can be employed in the place of $\psi(\hat{\mu}_k)$ as well. However, in most of the cases, moving objects have smooth motion, which can be modeled by a linear system. In this case, the model (9) can be simplified :

$$P(\hat{\mu}_k(t-1)|\hat{\mu}_k(j), f(j), j=1,\ldots,t-2, f(t-1)) = \exp[-(\hat{\mu}_k(t-1) - \underline{\mathbf{W}}_k \hat{\mu}_k^T)] \quad (10)$$

where $\underline{\mathbf{W}}_k$ is a matrix of size $M \times 5$ whose entries represent the dependency of a feature vector component at time $t-1$ with respect to its values in the previous M frames. The features that are tracked over the frames correspond to the position of the object, their graylevel (representing changes in luminance) and their associated optical flow. The relationship (10) can be generalized for tracking the characteristics of a feature vector entry with respect to all its entries in the previous frames. The components of the matrix $\underline{\mathbf{W}}_k$ can be found by using the Least Mean Squares (LMS) algorithm [6]. LMS algorithm can ensure the tracking of the feature points over several frames and has been used successfully in many applications. The change in the moving object representative vectors are reflected in the segmentation of the moving objects. In order to maximize the probability in (4), we should maximize simultaneously (5,8,10). The relationship (5) provides the initial estimate, while (10) provides an estimate of the feature vectors based on their previous values, which must be also consistent with an accurate frame reconstruction (8).

4 Prediction of a future frame

A prediction function provides an estimate of the moving object segmentation and optical flow estimation in a future frame based on the data extracted from the previous frames. A prediction function can be determined from a relationship similar to (10). Let us denote by $\pi_t(X_k(t+1))$ and $\pi_t(\mathcal{M}_k(t))$ the prediction of the moving region X_k segmentation and of its motion vector from the frame t into the frame $t+1$. The prediction function for the motion vector can be expressed based on the matrix $\underline{\mathbf{W}}_k$ derived from the maximization of the probability from (10). We neglect the dependency of the motion on the other feature vector entries. The optical flow prediction of a certain object is found in each consecutive frame as :

$$\pi_t(\hat{\mathcal{M}}_{k,x}(t)) = \mathbf{W}_{xx}\hat{\mathcal{M}}_{k,x} + \mathbf{W}_{yx}\hat{\mathcal{M}}_{k,y} \quad (11)$$

$$\pi_t(\hat{\mathcal{M}}_{k,y}(t)) = \mathbf{W}_{xy}\hat{\mathcal{M}}_{k,x} + \mathbf{W}_{yy}\hat{\mathcal{M}}_{k,y} \quad (12)$$

where $\hat{\mathcal{M}}_{x,k}$, $\hat{\mathcal{M}}_{k,y}$ represent the motion vectors on x and y directions associated with the k-th moving object for the last M frames and $\mathbf{W}_{xy}$, $\mathbf{W}_{xx}$ $\mathbf{W}_{yx}$, $\mathbf{W}_{yy}$ are their corresponding weighting vectors as found by the LMS algorithm. This prediction function can easily model complex movements such as rotation and acceleration. The number of frames M to be taken into account for the prediction system must be larger when the motion is smooth and smaller when the motion is fast changing. Similarly with the relationship (11) or (12) we can derive a prediction system for the change in the luminance by tracking the change in the average graylevel of a certain moving object.

The location of a moving object in a future frame is given by the segmentation in the actual frame and the prediction of its associated optical flow :

$$\pi_t(X_k(t+1)) = X_k(t) \oplus \pi_t(\hat{\mathcal{M}}_k(t)) \quad (13)$$

where we consider the displacement for all the pixels composing the moving object k, and where $\pi_t(\hat{\mathcal{M}}_k(t))$ components are derived in (11,12). Given a prediction function for the optical flow associated with the moving object k we can predict the frame $t+1$ considering the segmentation of the individual objects :

$$g(t+1) = \cup_{k=1}^{N} \pi_t(X_k(t+1)) \quad (14)$$

where $g(t+1)$ is the predicted image. As was shown in the previous Section, certain regions do not have a clear assignment. Their value is assigned based on an overlapping priority assumption. For example, if the background is known, it will get the lowest priority and it will be covered in the case of moving objects pointing to the same region or it will occupy the regions remained uncovered. The values to be used in the undetermined regions are taken from one of the previous frames, by considering the optical flow as well.

We consider a measure in order to attest the efficiency of the segmentation and of correctly evaluating the entire model. This measure compares the predicted future frame $g(t+1)$, reconstructed based on (14), and the actual frame (if it is available) :

$$E = |g(t+1) - f(t+1)| \quad (15)$$

This energy function is similarly with that employed for reconstructing each frame from the tracked sequence in (8). If E is above a certain threshold, then the model is not valid at the respective frame. Usually, this is caused because a new moving object appears in the sequence or one of the existing moving objects gets out of the scene. In such a case, the RBF network is

Figure 1. The first frame of the "Hamburg Taxi" image sequence

Figure 2. The 20th frame of the "Hamburg Taxi" image sequence

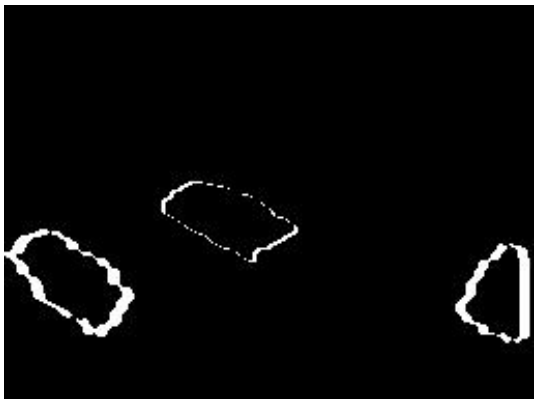

Figure 3. The undetermined regions.

Figure 4. The predicted 20th frame.

retrained in order to obtain the appropriate moving object segmentation and optical flow (5). The new model is tracked over the following frames as described in the previous Section.

5 Simulation results

We provide results when the proposed algorithm is applied in the "Hamburg Taxi" image sequence. Three main moving objects and the stationary background have to be identifyed and segmented in this image sequence. The first and the 20th frames are displayed in Figures 1 and 2. The moving object segmentation as provided by MRBF network for the first frame is given in Figure 5. Its correspondent optical flow is shown in Figure 6. The undetermined regions for the first frame are shown in Figure 3. They are located in between the boundaries of the same moving object in the two frames, as it can be observed from this Figure. After tracking the moving objects as described in Section 3 we obtain the segmentation of the 20th frame, as provided in Figure 7. Six past frames ($M = 6$) have been used in the tracking model. It can be observed that the segmentation of the white taxi in the center of the frame is quite well defined despite the fact that the taxi, due to the 3-D perspective view, changes its shape when turning around the corner. The optical flow corresponding to the tracked objects in the 20th frame is represented in Figure 8. The 20th frame, reconstructed from the predicted segmentation and moving object velocities as described in Section 4, is provided in Figure 4. The MRBF network training took 33.3 seconds. The trained network, can be used for those successive frames which match the model according to a criterion [3]. The network is applied for every feature vector of the image. In this case 95 seconds are required for segmenting the moving objects and the optical flow for 20 frames. When employing tracking as described in this study, only 68 seconds are necessary for the same frames. The processing times have been evaluated on a SiliconGraphics Indy Workstation. The segmentation provided by the tracking algorithm is quite good as it can be observed from the experimental results and provides a good basis for prediction based reconstruction.