Prediction and Tracking of Moving Objects in Image Sequences

Adrian G. Borş*and Ioannis Pitas†

Abstract

We employ a prediction model for moving object velocity and location estimation derived from Bayesian theory. The optical flow of a certain moving object depends on the history of its previous values. A joint optical flow estimation and moving object segmentation algorithm is used for the initialization of the tracking algorithm. The segmentation of the moving objects is determined by appropriately classifying the unlabeled and the occluding regions. Segmentation and optical flow tracking is used for predicting future frames.

1 Introduction

Tracking of moving objects is important for video surveillance while future frame prediction is used in video coding. A Bayesian approach shows that we can estimate the location of moving objects and their associated velocity based on a set of initial estimates. Occluding and unlabeled regions are identified and classified in the context of a tracking algorithm. A few approaches have been adopted for solving these problems. In [1] an occlusion adaptive mesh is used for tracking moving features over several frames. In other approaches, features are extracted from a set of frames and afterwards they are tracked over the sequence. Kalman filters have been used for tracking in [2, 3, 4]. Objects are segmented based on clustering in [3, 5]. Simultaneous optical flow estimation and moving object segmentation has been employed in [6]. In this approach the moving scene is modeled based on the Median Radial Basis Function (MRBF) network [8]. Each output unit of

*Adrian G. Borş was with the Dept. of Informatics, University of Thessaloniki, Thessaloniki 540 06, Greece. He is now with the Dept. of Computer Science, University of York, York YO10 5DD, U.K., (e-mail: Adrian.Bors@cs.york.ac.uk)

†I. Pitas is with the Dept. of Informatics, University of Thessaloniki, Thessaloniki 540 06, Greece (e-mail: pitas@zeus.csd.auth.gr).

the neural network corresponds to a moving object. The results provided by the MRBF modeling are used for the initialization of a tracking algorithm. The unlabeled regions in each frame are identified and classified appropriately based on the MRBF model. When new objects enter in the scene or when some objects leave the scene, retraining is necessary. In between two MRBF retraining stages, tracking is employed for following object movement. Using tracking we predict the moving object optical flow and segmentation. A future frame is represented as the union of its predicted moving objects.

The Bayesian model for motion and segmentation estimation over the entire image sequence is provided in Section 2. Tracking the moving objects over a set of frames is described in Section 3 and frame reconstruction based on estimating the moving object location and optical flow is described in Section 4. Simulation results are presented in Section 5 and the conclusions are drawn in Section 6.

2 Motion and segmentation estimation

Let us consider that each frame of an image sequence $f(t)$, $t = 1, \ldots, K$ is made up of a set of moving regions $\{X_i(t), i = 1, \ldots, N\}$ with the properties :

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

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

A 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 $\mathcal{S}_k$ of a certain moving region [6]. The still image feature vector $\mathcal{S}_k$ contains the location and the characteristic graylevel of the moving region. $\mathcal{S}_k$ is directly related to the segmentation label of the moving region k, while $\mathcal{M}_k = [\mathcal{M}_{k,x}, \mathcal{M}_{k,y}]$ is the velocity vector of the respective moving region. The classification of the image sequence in moving objects is done according to the maximization of the *a posteriori* probability :

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

where each probability corresponds to the segmentation of a moving object and its optical flow in the entire image sequence. After repeatedly applying the Bayes theorem and after expressing the probabilities from one frame with respect to those corresponding to the previous frames, we obtain :

$$P(\hat{\mu}_k(t), t=1,\ldots,K-1|f(t), t=1,\ldots,K) = \Pi_{t=p}^{K-1}[P(f(t+1)|\hat{\mu}_k(j), f(j), j=1,\ldots,t)]$$
$$\Pi_{t=p}^{K-1}[P(\hat{\mu}_k(t)|\hat{\mu}_k(j), f(j), j=1,\ldots,t-1, f(t))]\frac{P(\hat{\mu}_k(j), j=1,\ldots,p-1|f(j), j=1,\ldots,p)}{\Pi_{t=p}^{K-1}P(f(t+1)|f(j), j=1,\ldots,t)} \quad (4)$$

where K is the number of frames and p is a given frame $p < K$. A component of the first probability product from the right side of this relationship is associated to the reconstruction of a frame from the previous frames using the moving object feature vectors. A component of the second probability product corresponds to the feature vector tracking over several frames. The third probability factor models the moving object characteristics derived from the first p frames. The probabilities from the denominator denote the dependence of a frame on the previous ones and it can be neglected in the following considerations.

The first two frames are split in blocks and a feature vector denoted as $\mathbf{u}_{IJ} = [I, J, l]$ containing the site location $[I, J]$, the graylevel l and the motion vector is associated with each block. For $p = 2$, after using the Bayesian theorem, the third probability factor in (4), can be further described as :

$$P(\hat{\mathcal{M}}_j, \hat{\mathcal{S}}_j|f(2), f(1)) = P(f(2)|\hat{\mathcal{M}}_j, \hat{\mathcal{S}}_j, f(1))\ \frac{P(\hat{\mathcal{M}}_j|\hat{\mathcal{S}}_j, f(1))P(\hat{\mathcal{S}}_j|f(1))}{P(f(2)|f(1))} \quad (5)$$

where $P(\hat{\mathcal{S}}_j|f(1))$ represents the *a priori* probability of the segmentation and $P(\hat{\mathcal{M}}_j|\hat{\mathcal{S}}_j, f(1))$ is the probability of the optical flow estimation depending on the segmentation map and image [7]. After expressing each probability as an energy function, we model them with Gaussian functions. The Gaussian function associated with a moving region and implemented by a hidden unit of the MRBF network is given by :

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

where $\hat{\mu}_j$ and $\hat{\mathbf{\Sigma}}_j$ are the center vector and covariance matrix estimates and $WDFD(\hat{\mathcal{M}}_j)$ represents the weighted displaced frame difference (a measure of confidence in the motion estimation algorithm) [6]. An unsupervised training algorithm provides the estimates of the MRBF network parameters while modeling the probabilities from (5) [6, 8].

3 Moving object tracking

Let us neglect the dependence on all the frames excepting the previous. In this case we can express each probability in the first product of (4) as an energy function measuring the accuracy of reconstructing the frame $f(t+1)$ from the displaced moving objects which had been segmented in the frame $f(t)$:

$$P(f(t+1)|\hat{\mu}(t), f(t)) \simeq \frac{1}{Z} \exp\{-\delta[\cup_i^N (X_i(t) \oplus \hat{\mathcal{M}}_i(t)), f(t+1)]\{ \quad (7)$$

where Z is a normalizing constant, $X_i(t) \oplus \hat{\mathcal{M}}_i(t)$ represents the translation of the moving region $X_i(t)$ resulted from the segmentation of the frame $f(t)$ with its corresponding motion vector $\hat{\mathcal{M}}_i(t)$ and $\delta[f(t), g(t)]$ represents a function which counts in how many locations $f(t)$ and $g(t)$ are different. The maximization of this probability represents the minimization of the difference between the given frame and its prediction based on the previous frame segmentation and its estimated optical flow. It can be observed that by displacing the set of pixels $X_i(t)$ representing a moving region in the frame $f(t)$, certain pixels from $X_i(t) \oplus \hat{\mathcal{M}}_i(t)$ have uncertain assignment. When regions from one frame do not have a correspondent in the next frame (uncovered regions), (1) is not respected any more. When two or more different objects project in the same region of the next frame (occluding regions), (2) is not valid. Both situations occur in regions located at the margins of the moving objects and can be easily identified as providing a probability equal or smaller than $\exp(-1)/Z$ in (7). If we have a one-to-one correspondence between the frames $f(t)$ and $g(t)$ based on the given model then the probability from (7)is equal to $1/Z$. After detecting the unlabeled regions, we estimate their feature vectors $\mathbf{u}_{IJ}$ considering only the possible correlations given by the motion vectors of the neighboring moving objects. The trained MRBF network, can be applied in a multiresolution approach where the network parameters obtained from the initial block-based segmentation are used for image segmentation at pixel resolution [6]. We apply the already trained MRBF network only in the regions decided as uncertain according to (7).

The components of the second product from the expression (4) representing the dependency of a feature vector on the values of the same feature vector in the previous frames, can be expressed as in maximum-likelihood regression estimation [9] :

$$P(\hat{\mu}_k(t)|\hat{\mu}_k(j), f(j), j = 1, \ldots, t-1, f(t)) = \frac{1}{Z} \exp[-|\hat{\mu}_k(t) - \sum_{i=1}^{H} W_i \psi_i(\hat{\mu}_k)|] \tag{8}$$

where $\psi(\hat{\mu}_k)$ are a set of functions modeling the variation of the k-th object feature vector in time, W_i are their associated weights and Z is a normalizing constant. However, in most of the cases, moving objects have slow changing motion, which can be modeled by a linear system. Under this assumption, the model (8) can be simplified :

$$P(\hat{\mu}_k(t)|\hat{\mu}_k(j), f(j), j = 1, \ldots, t-1, f(t)) = \frac{1}{Z} \exp[-|\hat{\mu}_k(t) - \underline{\mathbf{W}}_k \mathbf{\Phi}_k^T|] \tag{9}$$

where $\mathbf{\Phi}_k$ consists of the feature vectors from the last M frames:

$$\mathbf{\Phi}_k = [\hat{\mu}_k(t-1)\hat{\mu}_k(t-2)\ldots\hat{\mu}_k(t-M)] \tag{10}$$

and $\underline{\mathbf{W}}_k$ is a matrix of size $5M \times 5$ whose entries represent the dependency of a feature vector component at time t with respect to all feature entries in the previous M frames. The features that are tracked over time correspond to object location, graylevel changes and optical flow. The components of the matrix $\underline{\mathbf{W}}_k$ can be found by using the Least Mean Squares (LMS) algorithm [10]. LMS is a fast on-line algorithm which can ensure feature tracking over several frames based on minimizing the prediction mean square error. Kalman filters can be seen as an extension of the LMS algorithm which however requires a much larger computational complexity. Changes in the moving object representative vectors are reflected in the moving object segmentation. In order to maximize the probability in (4), we should maximize its components from (5), (7) and (9). The relationship (5) provides the initial estimate, while (9) gives an estimate of the moving object feature vector from its previous values. This estimate must be consistent with an accurate frame reconstruction as given by (7).

4 Frame reconstruction from moving object prediction

A prediction function provides an estimate of the moving object segmentation and its corresponding optical flow in a future frame based on the data extracted from the previous frames. Let us denote by $\pi_t(X_k(t+1))$ and $\pi_t(\hat{\mathcal{M}}_k(t+1))$ the prediction of the location for the moving region k and the prediction of its optical flow respectively, from the frame t into the frame $t+1$. The prediction function for the velocity uses the matrix $\underline{\mathbf{W}}_k$, derived from the maximization of the probability from (9). The optical flow for a certain moving object is predicted for each consecutive frame by using the dependency on its previous values :

$$\pi_t(\hat{\mathcal{M}}_{k,x}(t+1)) = \mathbf{W}_{xx}\hat{\mathcal{M}}_{k,x} + \mathbf{W}_{yx}\hat{\mathcal{M}}_{k,y} \tag{11}$$

$$\pi_t(\hat{\mathcal{M}}_{k,y}(t+1)) = \mathbf{W}_{xy}\hat{\mathcal{M}}_{k,x} + \mathbf{W}_{yy}\hat{\mathcal{M}}_{k,y} \tag{12}$$

where $\hat{\mathcal{M}}_{k,x}$, $\hat{\mathcal{M}}_{k,y}$ represent the motion vector components 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 found by the LMS algorithm [10] as in (9). 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 to (11) or (12) we can derive a prediction system for 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 :

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

where we consider the displacement for all the pixels composing the moving object k, and where $\pi_t(\hat{\mathcal{M}}_k(t+1))$ 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 :

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

where $f_p(t+1)$ is the predicted image. As it was shown in the previous Section, certain regions do not have a clear assignment. Such regions are classified 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 fill in the regions which remained uncovered. The values to be used in the unlabeled regions of the predicted frame are taken from one of the previous frames by considering the optical flow.

The PSNR between the predicted frame $f_p(t+1)$ and the real one $f(t+1)$, when it is available, is considered for checking the validation of the assumed model:

$$PSNR = 20 \log_{10} \left(\frac{255RS}{\sqrt{\sum_{i=1,j=1}^{R,S} (f_{p,ij}(t+1) - f_{ij}(t+1))^2}} \right) \tag{15}$$

where $R \times S$ is the size of the image. If the PSNR between the two images is below a certain threshold, then the model is not valid at the respective frame. Usually, this is caused because a moving object enters or leaves the scene. In such a case, the MRBF network is retrained in order to obtain the appropriate moving object segmentation and optical flow (5), [6]. The new model is tracked over the following frames as described in the previous Section.

5 Simulation results

We provide simulation results when the proposed algorithm is applied in the "Hamburg Taxi" image sequence. The first and the 20th frames are displayed in Figures 1 and 2. In the center of a frame from this image sequence a white taxi turns around the corner, a black car moves from left to right while a van moves from right to left. The moving object segmentation as provided by the MRBF network for the first frame is shown in Figure 3. Its corresponding optical flow is provided in Figure 4. The segmentation and optical flow parameters are used for the initialization of the tracking algorithm. The occluding and unlabeled regions for the first frame are shown in Figure 5. They are located at the moving object boundaries according to a small probability in (7). The pixels of these regions are classified using the MRBF network parameters. The moving object segmentation resulted after this classification is displayed in Figure 6. 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 for tracking. It can be observed that the segmentation of the white taxi in the center of the frame is quite good despite the fact that, due to the 3-D perspective view, its projection changes while turning around the corner. The optical flow corresponding to the tracked objects in the 20th frame is represented in Figure 8. The predicted 20th frame, reconstructed from the predicted segmentation and moving object velocities, is provided in Figure 9. The difference between the predicted and the real 20th frame is shown in

Figure 10. We can see from this Figure that many errors in the prediction of the 20th frame are due to changes in illumination. In Figure 11, the PSNR of the predicted image when tracking the moving objects is ploted for a set of frames. The MRBF network training took 33.3 seconds when using a Silicon Graphics Indy Workstation. The trained network, can be used for those successive frames which match the model according to a criterion [6]. In this case 95 seconds are required for segmenting the moving objects and the optical flow for 20 frames when using 4×4 pixel blocks. When employing tracking as described in this study, only 68 seconds are necessary for the same frames using pixel resolution segmentation. In the first case only 3040 vectors had been processed while in the second case their number was 48640. 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 frame reconstruction. The prediction PSNR of the tracking algorithm is better than when considering the initial MRBF model for segmenting all the frames and assuming just the previous moving object features for reconstruction, as it can be observed from Figure 11.

6 Conclusions

We propose a moving object tracking algorithm derived from the Bayesian theory. The optical flow and the segmentation features are jointly modeled by the MRBF network in the initial stage. The occluding and unlabeled regions are detected and classified appropriately. The proposed algorithm provides good moving object tracking capabilities. Such capabilities are used for segmenting and estimating the moving object velocity and segmentation in a future frame. The proposed algorithm is employed for frame prediction.

References

[1] Y. Altunbasak, A. M. Tekalp, "Occlusion-adaptive, content-based mesh design and forward tracking," *IEEE Trans. on Image Processing,* vol. 6, no. 9, pp. 1270 - 1280, Sep. 1997.

[2] K. J. Bradshaw, I. D. Reid, D. W. Murray, "The active recovery of 3-D motion trajectories and their use in prediction," *IEEE Trans. on Pat. Anal. Mach. Intel.,* vol. 19, no. 3, pp. 219-224,

Mar. 1997.

[3] S.M. Smith, J. M. Brady, "ASSET-2: Real-time motion segmentation and shape tracking," *IEEE Trans. on Pattern Analysis and Machine Intelligence,* vol. 17, no. 8, pp. 814-820, Aug. 1995.

[4] J. Weber, J. Malik, "Rigid body segmentation and shape description from dense optical flow under weak perspective," *IEEE Trans. on Pattern Analysis and Machine Intelligence,* vol. 19, no. 2, pp. 139-143, Feb. 1997.

[5] A. Kumar, Y. B.-Shalom, E. Oron, "Precision Tracking based on segmentation with optimal layering for imaging sensors," *IEEE Trans. on Pattern Analysis and Machine Intelligence,* vol. 17, no. 2, pp. 182 - 188, Feb. 1995.

[6] A. G. Borş, I. Pitas, "Optical flow estimation and moving object segmentation based on median radial basis function network," *IEEE Trans. on Image Processing,* vol. 7, no. 5, pp. 693-702, May 1998.

[7] J. Konrad, E. Dubois, "Bayesian estimation of motion vector fields," *IEEE Trans. on Pat. Anal. Mach. Intel.,* vol. 14, No. 9, pp. 910-927, Sep. 1992.

[8] A. G. Borş, I. Pitas, "Median radial basis function neural network," *IEEE Trans. on Neural Networks,* vol. 7, no. 6 , pp. 1351-1364, Nov. 1996.

[9] V. Vapnik, *Estimation of dependences based on empirical data.* New York: Springer-Verlag, 1982.

[10] B. Widrow, S.D. Stearns, *Adaptive Signal Processing.* Englewood Cliffs, NJ: Prentice Hall 1985.

List of Figures

Figure 1: The first frame of the “Hamburg Taxi” image sequence.

Figure 2: The 20th frame of the “Hamburg Taxi” image sequence.

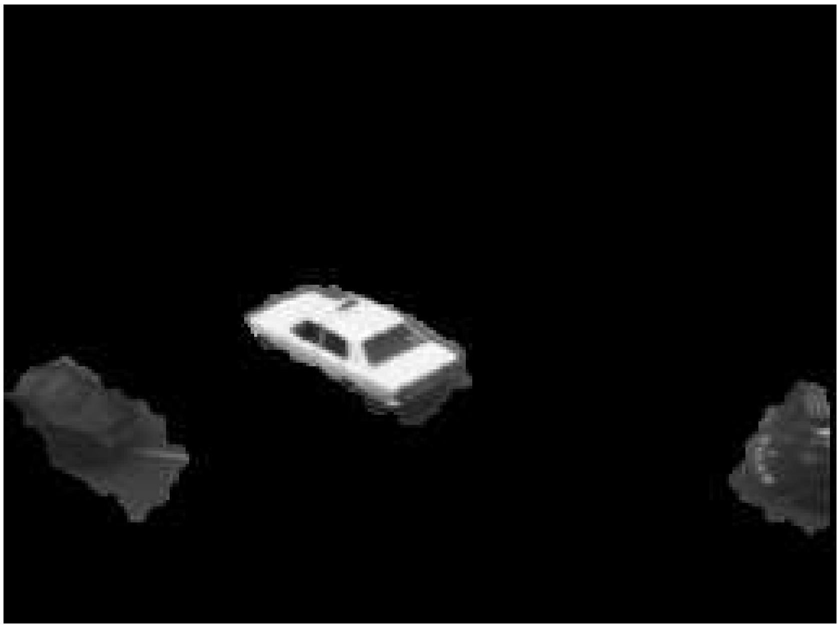

Figure 3: Moving object segmentation.

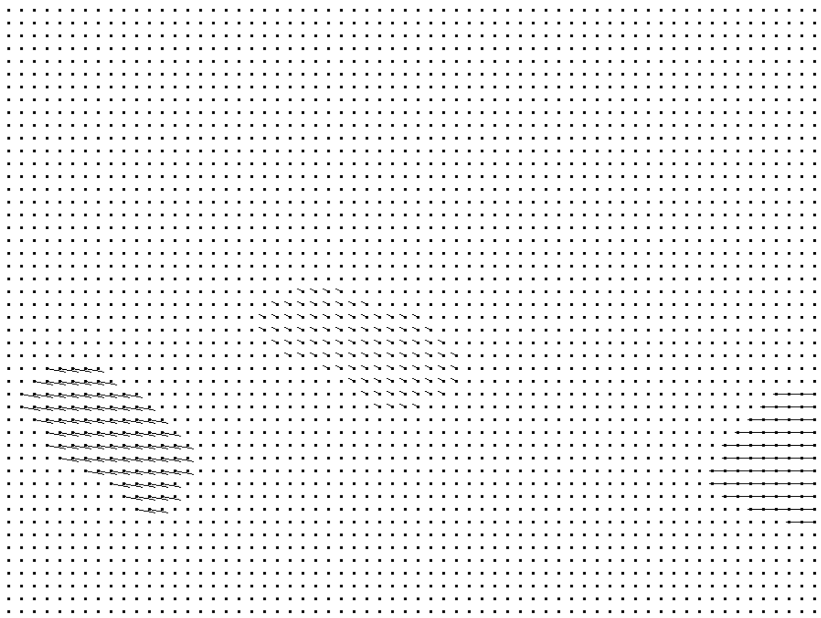

Figure 4: The optical flow of the first frame from the "Hamburg Taxi" image sequence.

Figure 5: Occluding and unlabeled regions.

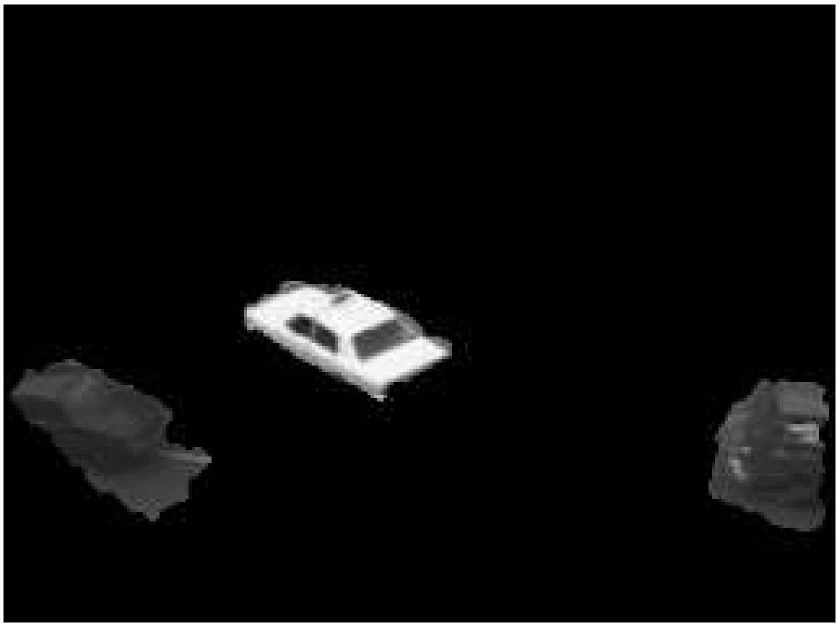

Figure 6: Segmentation of the moving objects after appropriately classifying the occluding regions.

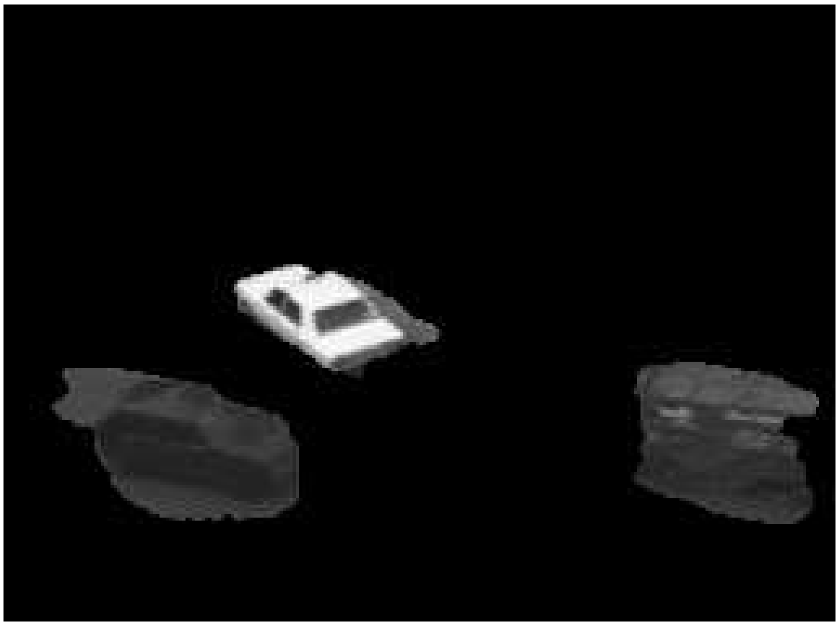

Figure 7: Moving object segmentation resulted after tracking them 20 frames.

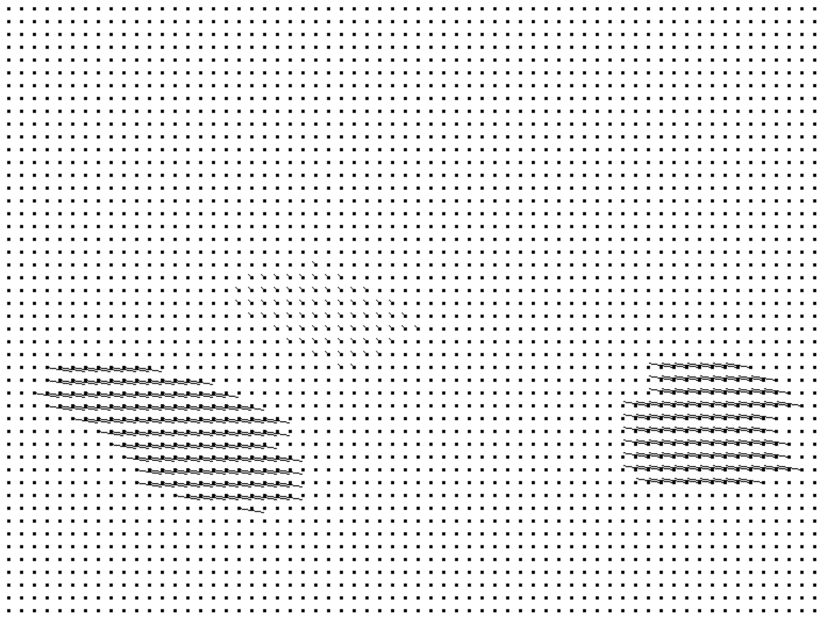

Figure 8: Estimated optical flow of the 20th frame.

Figure 9: The predicted 20th frame.

Figure 10: The difference between the predicted and the real 20th frame from the "Hamburg Taxi" image sequence.

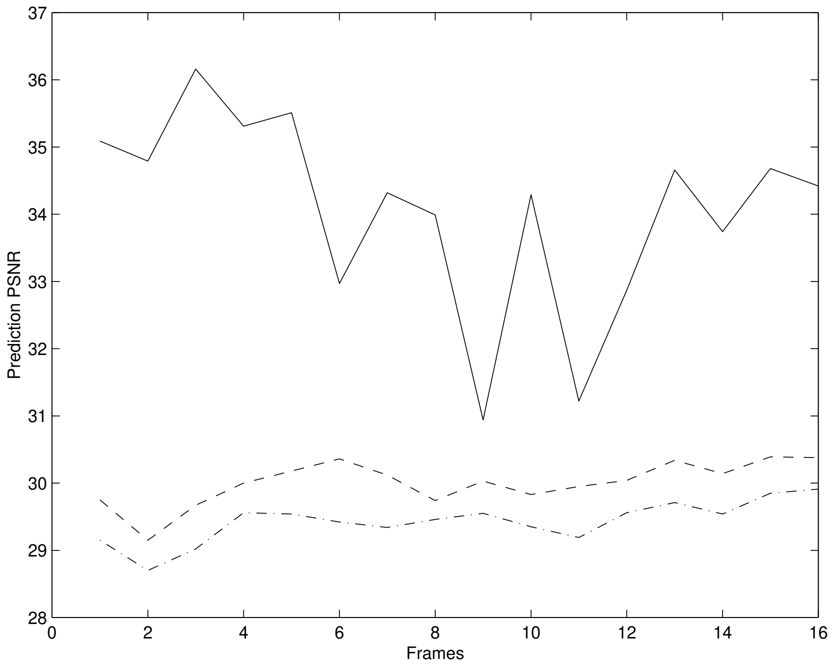

Figure 11: The PSNR of the predicted frame in the "Hamburg Taxi" image sequence. '–' denotes the PSNR of the proposed tracking algorithm. '- -' represents the PSNR prediction considering the initial MRBF model on 4×4 pixel blocks. '-.' denotes the PSNR between the actual frame and that used for prediction.