

# An HMM-Based Segmentation Method for Traffic Monitoring Movies

Jien Kato, *Member, IEEE Computer Society*,  
 Toyohide Watanabe, *Member,*  
*IEEE Computer Society*,  
 Sébastien Joga, Jens Rittscher, and  
 Andrew Blake, *Member,*  
*IEEE Computer Society*

**Abstract**—Shadows of moving objects often obstruct robust visual tracking. We propose an HMM-based segmentation method which classifies in real time each pixel or region into three categories: shadows, foreground, and background objects. In the case of traffic monitoring movies, the effectiveness of the proposed method has been proven through experimental results.

**Index Terms**—Car tracking, hidden Markov model, image classification, image segmentation, wavelet coefficients.

## 1 INTRODUCTION

THE main obstacle to robust visual tracking is that distracting features, such as clutter in the background regions, compete for the attention of the tracker and may succeed in pulling the tracker away from foreground (target) objects [3]. To make the tracker reliable, it is a common practice to discriminate the foreground pixels from the background pixels. Earlier researchers have attempted to increase the robustness of the tracker by image differentiation techniques [11], [2], [10], [13]. However, as to applications such as traffic monitoring systems, typical troublesome features are the shadows of vehicles, which are not well-handled by traditional techniques.

This paper introduces a new segmentation method [9] based on Hidden Markov Models (HMMs) to deal with problems of the shadows of vehicles. The method is mainly composed of two phases: the learning phase and the segmentation phase. In the learning phase, the tracking process learns the unknown HMM parameters with an EM-type (Expectation-Maximization) algorithm [5] over several seconds of a video sequence. In the segmentation phase, the process classifies each small region in a field image of a movie into three different categories: foreground (F), background (B), and shadow (S) over time.

## 2 RELATED RESEARCH

A number of researchers have adopted image differentiation techniques such as background subtraction and interframe differentiation [11], [2], [10], [13] to make the trackers reliable. Background subtraction methods are based on the assumption that

the background is perfectly static. This assumption seems unrealistic for outdoor systems. Moreover, this method cannot remove the shadows of moving objects, such as vehicles. On the other hand, interframe differentiation has the advantage that the shadows only appear as outlines on the resulting image. However, the homogeneous regions inside the vehicle appear as background regions. Thus, it is difficult to distinguish vehicles from shadows by using image differentiation techniques.

Another approach is to enhance immunity to distracting background objects by modeling the background [12], [16], [7], [15]. Toyama, et al. [16] discussed the issue of background maintenance by using a multilayered approach. The intensity distribution over time is modeled as an autoregressive process of order 30, but the computation costs seem to be too expensive for real-time systems. Haritaoglu, et al. [7] composed a simple gray-value distribution model for the background. According to this method, the background is modeled by representing each pixel by three values: the pixel's minimum and maximum intensities and the largest interframe difference between two consecutive frames during the training period. With this distribution model, the system determines whether a part of the scene contains foreground objects. However, the obtained foreground image is not sufficiently clear, so it cannot be used for tracking. Rowe and Blake [15] proposed a statistical model for the distribution of intensities of the background pixels. They investigated the special case of a video camera mounted on a pan-tilt head. In this case, they could not use a global threshold for foreground-background separation because of the intensity variation caused by the moving camera. Practical tests show better results than the tests done by image subtraction methods, but the method did not solve the problem of the shadows of moving objects.

It is still necessary to design a method which is able to model shadows as well as foreground and background objects and also work in real time.

## 3 APPROACH TO ROBUST SEGMENTATION

For robust car tracking, our method should perform accurate segmentation of the foreground objects from background objects and shadows. We employ HMMs to deal with three different categories including the shadows of vehicles. The use of HMMs has two main advantages that were not found in previous methods. First, an HMM is a suitable model to incorporate temporal continuity. Temporal continuity here means that a pixel belongs to a certain category for a period of time. If a pixel belongs to the foreground (a vehicle) at a given moment, it is likely that the pixel will still belong to the foreground (the same vehicle) at the next time step. Traditional background models based on intensities perform poorly when intensity differences among categories are small. In consideration of the fact that the temporal continuity of each category is very important, especially in case of almost no intensity differences among categories, we utilize HMMs (1D HMMs) to incorporate the temporal continuity of each category along a time-axis. 1D HMMs need less model parameters (than 2D HMMs do), which reduces learning computation costs. This is desirable for a real-time tracking system. The context-dependence among neighboring pixels or regions is incorporated in the segmentation phase. Second, it is not necessary to provide specific data for learning. This is because HMMs are able to learn the observation distributions for different hidden states from an ordinary image sequence. This property is particularly important for traffic monitoring sequences. Since all the three intensity distributions (for F, B, and S) have a large amount of overlap, it is difficult to find a robust way to learn these distributions separately.

### Corporate

- J. Kato and T. Watanabe are with the Department of Information Engineering, Nagoya Univ., Furo-Cho, Chikusa-Ku, Nagoya, 464-8603, Japan. E-mail: {jien, watanabe}@nuie.nagoya-u.ac.jp.
- S. Joga is with France Telecom R&D-DMR/RMO, 38 rue du general Leclerc, 92794 Issy-Les-Moulineaux Cedex 9, France. E-mail: sebastien.joga@rd.france-telecom.com.
- J. Rittscher is with GE Corporate Research and Development, Visualization and Computer Vision Lab, PO Box 8, Schenectady, NY 12301. E-mail: rittsche@crd.ge.com.
- A. Blake is with Microsoft Research, 7 JJ Thompson Avenue, Cambridge CB3 0FB, UK. E-mail: ablake@microsoft.com. **NY 12301, USA**

Manuscript received 13 June 2000; revised 2 Feb. 2001; accepted 9 Jan. 2002. Recommended for acceptance by A. Kundu.

For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number 112276.

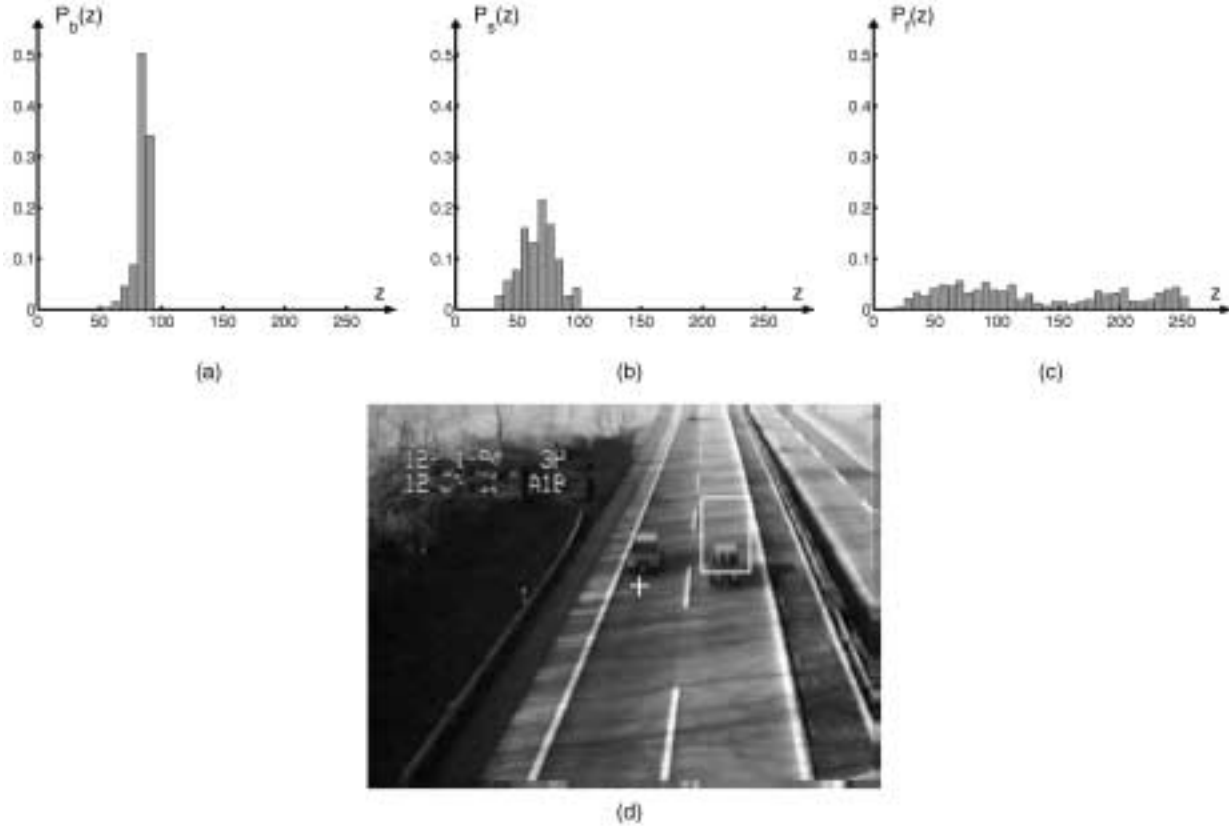


Fig. 1. Frequency profiles of the intensity of  $z \in \{B, S, F\}$ . (a) Background, (b) shadow, and (c) foreground. (d) A picture taken from a motorway sequence. The data related to (a)-(c) is observed at the pixel marked with a cross. The rectangle on the right lane is an area of interest.

### time, and these

### densities, and the

In addition to the adoption of the HMM technology, we also developed ways to make our method more robust, especially to extract the foreground objects more reliably. The first, is to use two kinds of observation symbols: intensities and high frequency wavelet coefficients. The two values are observed at any time these two series of observations depend on an underlying unobservable process which explains the transitions among hidden categories **B**, **S**, and **F**. Details are described in Section 4.5. The second, is to use a constrained HMM. Among all the category transitions, specific transition is prohibited. Details are described in Section 4.5. The third, is to take the context-dependence among nearby regions into account during the segmentation phase. Details are described in Section 4.6. Experimental results on real world motorway sequences show that this method makes it possible to accurately segment the image into three categories, as described in Section 5.

## Section 4.2

## 4 IMPLEMENTATION

### 4.1 The Hidden Markov Model

To design a suitable model for traffic monitoring, we investigated the intensity variations of three categories at a given pixel first (see Fig. 1). Except in the case of congested traffic, the background distribution occupies the main area of the distribution at a given pixel. The shadow has of course a lower intensity than the background. However, the distributions of **B**, **S**, and **F** are not independently separated but partially overlap each other. According to the Fig. 1, the distributions of the shadow and background can be approximated by Gaussian densities. As to the foreground, the simplest reasonable model is a uniform probability density since the density of vehicles covers a large range of gray-values and does not depend on a given location. We also use high frequency wavelet coefficients as the second observation. In the

same way that we deal with the intensities, we also approximate the distributions of the second observation for the background and shadow by Gaussian densities the distribution of the foreground as a uniform probability density. Moreover, for background and shadow, we treat two observations as a single two-dimensional feature vector. Thus, each distribution is actually modeled as a two-dimensional Gaussian-mixture density. These approximations lead to less learning computation time.

Let  $S = \{S_b, S_s, S_f\}$  be the states corresponding to three categories. The parameters of the HMM [14], notated as  $\lambda = \{A, B, \pi\}$ , are specified in our particular problem as follows:

- Initial state distribution:

$$\pi = \{\pi_b, \pi_s, \pi_f\}, \quad \pi_i = \Pr(S_i \text{ at } t = 1).$$

- State transition matrix:

$$A = \begin{pmatrix} a_{bb} & a_{bs} & a_{bf} \\ a_{sb} & a_{ss} & a_{sf} \\ a_{fb} & a_{fs} & a_{ff} \end{pmatrix}, \quad a_{ij} = \Pr(S_j \text{ at } t + 1 | S_i \text{ at } t).$$

- Observation probability distribution in state  $j$ :  $B = \{b_j(v)\}$ ,  $b_j(v) = \Pr(v \text{ at } t | S_j \text{ at } t)$ , where  $v$  is the feature vector.

The state transition which plays the role of modeling the temporal continuity of categories is a first-order Markovian. The observation probabilities of background and shadow are characterized by only mean vectors ( $\mu_i$ ) and covariance matrices ( $\Sigma_i$ ) instead of all the probabilities for different observation values, i.e.,

$$b_i(v) = \frac{1}{\sqrt{(2\pi)^2 \det(\Sigma_i)}} e^{-\frac{1}{2}(v-\mu_i)^T \Sigma_i^{-1} (v-\mu_i)}, \quad i \in \{b, s\}. \quad (1)$$

The mean vectors are denoted by  $\mu_i = (\mu_1^i, \mu_2^i)$  and the covariance matrices by

$$\Sigma_i = \begin{pmatrix} \sigma_{11}^i & \sigma_{12}^i \\ \sigma_{21}^i & \sigma_{22}^i \end{pmatrix},$$

where the subscripts 1 and 2 mean the first and second observations, respectively. On the other hand, the distribution for foreground has density

$$b_f(v) = \frac{1}{256 \times 512}. \quad (2)$$

## 4.2 Observations

The observation symbols are a series of the values observed during the target movie. The observations depend on an underlying unobservable process which explains the transitions among the hidden categories **B**, **S**, and **F**. In principle, the proposed method can allocate one set of HMM parameters to each pixel location. However, we assume that a nonoverlapped small block with equal size ( $k \times k$  pixels), called an HMM region, has a set of HMM parameters. In the learning phase, the model parameters for each HMM region are estimated from a learning sequence. While in the segmentation phase, one series of optimal states is found for each HMM region over time. This assumption leads to a reduction of the number of HMMs, which contributes to considerably less computation time.

For the intensity observations, we use the outputs of a  $k \times k$  mean filter instead of using the gray-level intensities directly in order to reduce noise. However, experiments show that the discrimination of dark-colored vehicles from shadows using only the intensity observations is difficult. To distinguish foreground objects from shadows reliably, we employ high frequency wavelet coefficients as the second observation. The introduction of this observation is based on the idea that the variance of wavelet coefficients in high frequency bands should be small for **S** and **B**, but large for **F**. This is because the foreground objects are generally sharply focused and have more details within the objects than the out-of-focus background and shadow regions [18]. The second observation is calculated as the variance of the wavelet coefficients in LH, HL, and HH bands. In our current implementation, Daubechies wavelet transformation ( $N = 2$ ) is adopted [4]. We treat two observations as a single two-dimensional feature vector, as described in Section 4.1.

## 4.3 Baum-Welsh Reestimation Formulae

We learn the unknown HMM parameters by use of an EM algorithm. EM algorithms perform an iterative computation of maximum likelihood estimation when the observed data are incomplete [5]. The aim of parameter learning is to find the model parameter  $\lambda$  which maximizes  $L(x, \lambda) = \log[p(x|\lambda)]$  for a given set  $x$  of observed data. A special case of the EM algorithm, Baum-Welsh algorithm [1], is applied to learn unknown model parameters. It produces a sequence of estimates for  $\lambda$ , given a set of observed data  $x$ , so that each estimate  $\lambda^i$  has a greater value of  $\log[p(x|\lambda)]$  than the preceding estimate  $\lambda^{i-1}$ . The reestimation formulae for  $\pi$ ,  $A$  and  $B$  are defined as follows:

$$\bar{\pi}_i = \gamma_1(i), \quad (3)$$

$$\bar{\mu}_i = \frac{\sum_{t=1}^T v_t \gamma_t(i)}{\sum_{t=1}^T \gamma_t(i)}, \quad (4)$$

$$\bar{\Sigma}_i = \frac{\sum_{t=1}^T \gamma_t(i) (v_t - \bar{\mu}_i)(v_t - \bar{\mu}_i)^T}{\sum_{t=1}^T \gamma_t(i)}, \quad (5)$$

$$\bar{a}_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i, j)}{\sum_{t=1}^{T-1} \gamma_t(i)}, \quad (6)$$

where  $\gamma_t(i) = \Pr(S_i \text{ at } t|V, \lambda)$  and

$$\xi_t(i, j) = \Pr(S_i \text{ at } t, S_j \text{ at } t+1|V, \lambda)$$

are auxiliary probabilities that can be efficiently calculated by the so-called forward-backward algorithm [14].  $V = \{v_1, \dots, v_T\}$  is a sequence of observation symbols. The model parameters are reestimated using (3)-(6) until the likelihood stabilizes. Experiments show that 10 reestimations seem to function well and allow a reasonable time of computation.

## 4.4 Parameter Initialization

As the learning algorithm leads to a local maximum, it is important to choose appropriate initial HMM parameters. To set initial parameters properly, time constants  $\tau_b$ ,  $\tau_s$ , and  $\tau_f$  are defined as the typical duration time in which a pixel belongs to **B**, **S**, and **F**. Also, let  $\lambda_b$ ,  $\lambda_s$ , and  $\lambda_f$  be the proportions of the time spent in **B**, **S**, and **F** with  $\lambda_b + \lambda_s + \lambda_f = 1$ . A reasonable set of initial parameters for the state transition matrix can be chosen as

$$A = \begin{pmatrix} 1 - \frac{1}{\tau_b} & \frac{1}{\tau_b} \Lambda_{sf} & \frac{1}{\tau_b} \Lambda_{fs} \\ \frac{1}{\tau_s} \Lambda_{bf} & 1 - \frac{1}{\tau_s} & \frac{1}{\tau_s} \Lambda_{fb} \\ \frac{1}{\tau_f} \Lambda_{sb} & \frac{1}{\tau_f} \Lambda_{sb} & 1 - \frac{1}{\tau_f} \end{pmatrix}, \quad \Lambda_{ij} = \lambda_i / (\lambda_i + \lambda_j). \quad (7)$$

← The initial probability is chosen to be

$$\pi = \{\lambda_b, \lambda_s, \lambda_f\}. \quad (8)$$

The initial parameter for  $\mu_b$  can be estimated by the mode of intensities or the wavelet coefficients at a given HMM region since  $\lambda_b \gg \lambda_s$  and  $\lambda_b \gg \lambda_f$ . The covariance matrix  $\Sigma_b$  is determined empirically, as estimating  $\Sigma_b$  from data is not a stable way even if the traffic is not very heavy. The initial parameters for  $\mu_1^s$  and  $\sigma_{11}^s$  are selected based on the assumption that the intensity of the shadow is lower than that of the background. Approximating the support of a Gaussian as  $[\mu - 2\sigma, \mu + 2\sigma]$ ,  $\mu_1^s$  and  $\sigma_{11}^s$  can be chosen so that the shadow distribution support goes from 0 to the upper limit of the background distribution support, i.e.,

$$\mu_1^s = \frac{\mu_1^b + 2\sqrt{\sigma_{11}^b}}{2}, \quad \sigma_{11}^s = \left(\frac{\mu_1^s}{2}\right)^2 \quad (9)$$

The remaining initial parameters for  $\mu_2^s$  and other elements of  $\Sigma_s$  are given the same values as the corresponding parameters for the background. Obviously, all these initial parameters meet the stochastic constraints for HMM parameters:  $\sum_i \pi_i = 1$ ,  $\sum_j a_{ij} = 1$ , and  $\sum_v b_i(v) = 1$ .

## 4.5 Improvement of the Model

The HMM is at first considered as an ergodic model in which every state can be reached in a single step from every other state. The validity of this assumption is examined by inspection of the learned parameters.

For investigations, we use a thirty-second motorway sequence (see a frame image in Fig. 1d) to learn the model parameters. The parameters are estimated at an HMM region at the center of the left lane, where there is almost no state transition between **F** and **S** during the whole sequence. For simplicity, only intensities with a  $3 \times 3$  mean filter are used for the observation the second observation is not used here. The learned moments are  $\mu_1^b = 80.143$ ,  $\sigma_{11}^b = 10.244$ ,  $\mu_1^s = 68.2294$ , and  $\sigma_{11}^s = 44.9762$ . The two learned Gaussian

observation, and the

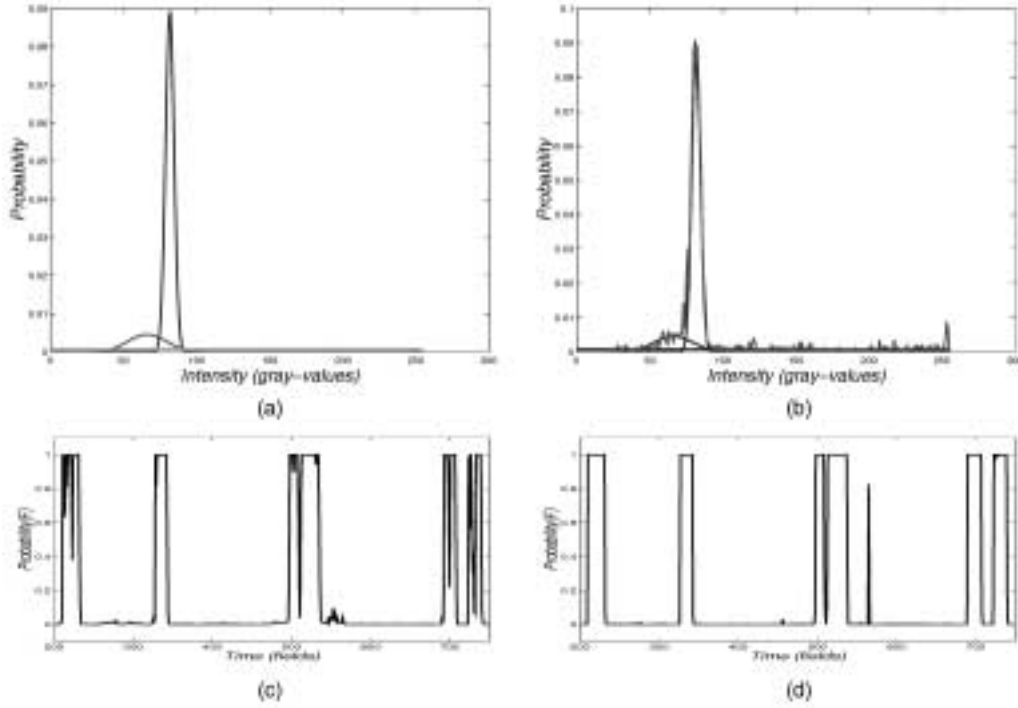


Fig. 2. Model parameter learning. (a) A set of distributions  $b_b$ ,  $b_s$  and  $b_f$  estimated for one HMM region. (b) Comparison of the learned model with raw data. (c) The probability of being in  $F$  at an HMM region, given learned parameters of the ergodic model. (d) The same probability as (c), given learned parameters of the constrained model.

intensities and one uniform density are shown in Fig. 2a. Fig. 2b shows that the raw data are well-matched by estimated  $\mu_1^b$ ,  $\sigma_{11}^b$ ,  $\mu_1^s$ , and  $\sigma_{11}^s$ . The learned state transition matrix

$$A = \begin{pmatrix} 0.986032 & 0.0129378 & 0.00102398 \\ 0.0138807 & 0.884321 & 0.101796 \\ 0.0333774 & 0.0253723 & 0.941252 \end{pmatrix}$$

implies that  $\tau_b \approx 72$ ,  $\tau_s \approx 9$ , and  $\tau_f \approx 17$ . The typical duration time  $\tau_b$  is not so different from the average time actually spent in  $B$  ( $\bar{\tau}_b = 75$ ). Thus, it seems there is no confusion between  $B$  and  $\{S, F\}$ . On the other hand, the shadows of vehicles almost do not appear at this position, but  $\tau_s \approx 9$  was obtained. However, this is not a serious problem because, in the learning sequence, there is always a dark part in front of a vehicle.

The problem is that  $\tau_f \approx 17$  is much shorter than the average time actually spent in  $F$  ( $\bar{\tau}_f = 31$ ). Wind-screens of cars are usually darker than the rest of the vehicle and are easily classified as shadow in the learning process. Since a car is learned as two or more separated foreground units,  $\tau_f$  is under-estimated. To solve this problem, we consider an HMM that is not fully connected: namely,  $a_{fs} = 0$ . Experiments show this assumption is reasonable for all used traffic monitoring sequences. The state transition matrix estimated with the constrained model in the same HMM region of the identical learning sequence becomes

$$A = \begin{pmatrix} 0.980455 & 0.0157114 & 0.00382949 \\ 0.0130607 & 0.897758 & 0.0891776 \\ 0.0479983 & 0 & 0.952003 \end{pmatrix}.$$

The probability  $a_{ff}$  (and, thus,  $\tau_f$ ) has been increased to 0.952003, which implies  $\tau_f \approx 21$ . Figs. 2c and 2d show the probabilities of being in state  $F$  using the ergodic and constrained models, respectively. The constrained model categorizes four cars out of five as whole entities which are all classified into separated units by the ergodic model.

Thus, it is necessary to modify the model according to assumption  $a_{fs} = 0$ .

#### 4.6 State Estimation

In the segmentation phase, an “optimal” state is found for each HMM region. Given the observation sequence, several criteria for selecting an optimal state sequence may be used. In view of tracking, the basic requirement for the state estimation is to work in real time. Namely, we cannot adopt a criterion that uses the whole sequence of observations as the Viterbi algorithm does [17], [6]. One solution is to maximize the joint probability of the state at time  $t$  and the past observation  $\{v_1, \dots, v_t\}$ , under the model  $\lambda$ , i.e.,

$$\operatorname{argmax}\{\alpha_t(k)\} = \operatorname{argmax}\{\Pr(v_1, \dots, v_t, S_k \text{ at } t | \lambda)\}, \quad (10)$$

However, this criterion does not incorporate context-dependence among HMM regions. For example, an  $F$  state is highly unlikely to exist in isolation surrounded by  $B$  regions. To take this important consideration into account, we estimate a state with a criterion

$$\operatorname{argmax}\{\Pr(v_1, \dots, v_t, S_k \text{ at } t | \lambda) \Pr(Q_{i,j} | \mathcal{Q}_{\mathcal{N}_{i,j}})\}, \quad (11)$$

where  $\Pr(Q_{i,j} | \mathcal{Q}_{\mathcal{N}_{i,j}})$  means the probability of the state being  $Q_{i,j}$  at region  $(i, j)$ , under the conditions that the state set  $\mathcal{Q}_{\mathcal{N}_{i,j}}$  is observed at neighborhood  $\mathcal{N}_{i,j}$  of region  $(i, j)$ . We define  $\Pr(Q_{i,j} | \mathcal{Q}_{\mathcal{N}_{i,j}})$  as

$$\Pr(Q_{i,j} | \mathcal{Q}_{\mathcal{N}_{i,j}}) = \frac{1}{D} \exp(\kappa \vartheta(Q_{i,j}; \mathcal{Q}_{\mathcal{N}_{i,j}})), \quad (12)$$

where  $D$  is a normalization constant and  $\kappa$  is a parameter that denotes the strength of the context-dependence between HMM regions. The function  $\vartheta(Q_{i,j}; \mathcal{Q}_{\mathcal{N}_{i,j}})$  is selected as

$$\vartheta(Q_{i,j}; \mathcal{Q}_{\mathcal{N}_{i,j}}) = \sum_{(s,r) \in \mathcal{N}_{i,j}^8} \frac{1}{16} I(i, j, s, r) + \sum_{(s',r') \in \mathcal{N}_{i,j}^{16}} \frac{1}{32} I(i, j, s', r'), \quad (13)$$

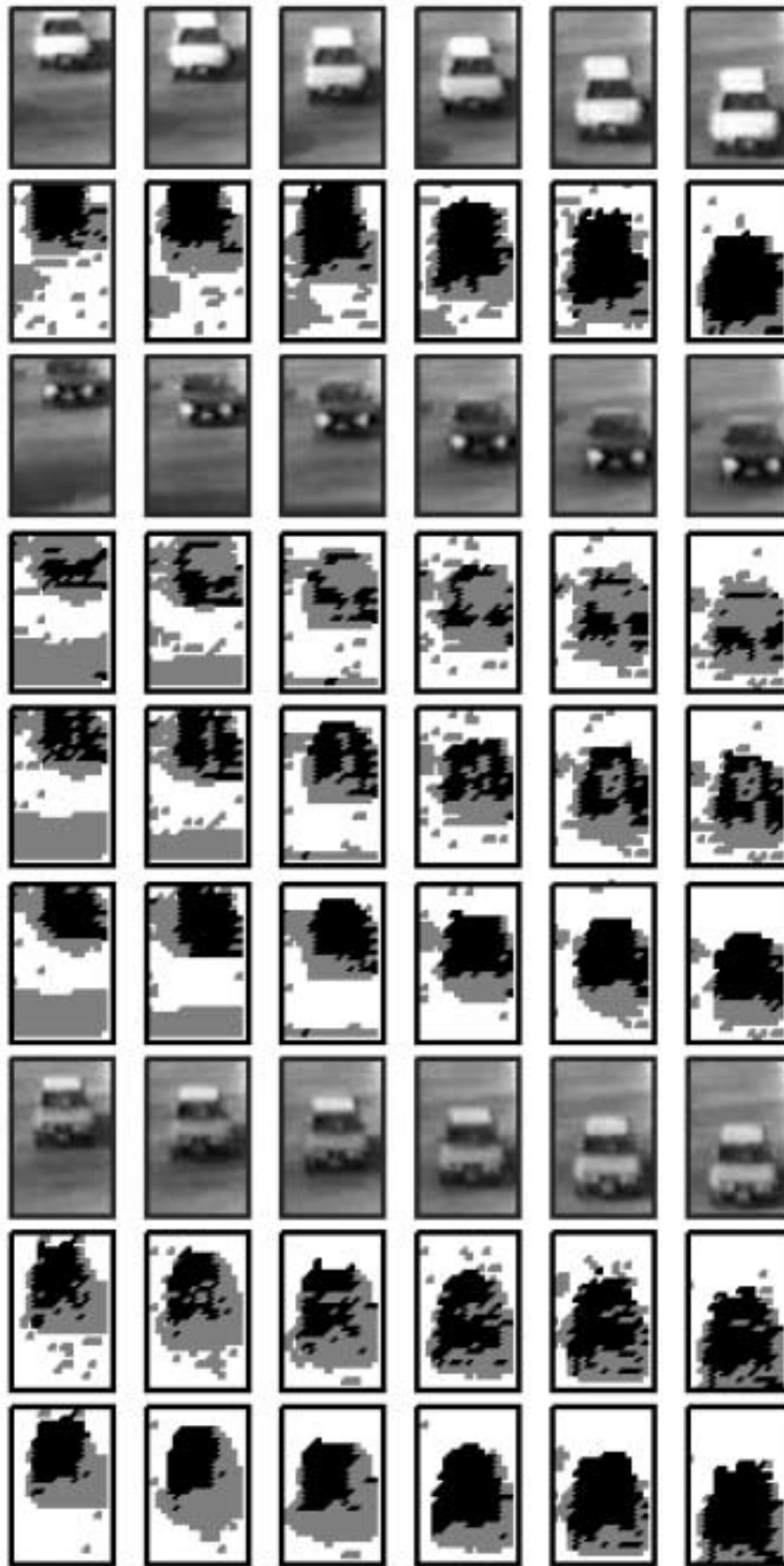


Fig. 3. The visualization of the results of state estimation under the constrained model with  $a_{fs} = 0$ , for the area shown in Fig 1d. Foreground: black, shadow: gray, and background: white.

$$I(i, j, s, r) = \begin{cases} 1 & Q_{i,j} = Q_{s,r} \\ 0 & Q_{i,j} \neq Q_{s,r} \end{cases} \quad (14)$$

where  $\mathcal{N}_{i,j}^8$  and  $\mathcal{N}_{i,j}^{16}$  are the 8-neighbors of region  $(i, j)$  with distances 1 and 2. Note that (10) and (11) can be solved by the forward procedure alone [1]. Since these criteria are defined recursively, it is possible to perform the state estimation in real time using (10) or (11).

## 5 EXPERIMENTAL RESULTS

In this section, we present the experimental results that are all obtained under the constrained model with  $a_{fs} = 0$ . Several 30 second sequences are used for the experiments. Although the traffic density and light conditions of these sequences do not change much, the typical duration times spent in **B**, **F**, and **S** might be quite different from each other. The results we discuss here are obtained with respect to an area in the right lane, where all categories are observed. Without loss of generality, the following descriptions are also true of other areas in the image. The area is composed of  $18 \times 28$  HMM regions and each region has  $4 \times 4$  pixel size (see Fig. 1d). Some results are given in Fig. 3. To make the explanation straightforward, we roughly divide the cars into light, dark, and gray ones.

First, we consider the light cars. The first row in Fig. 3 shows six successive images of a light car at three-field intervals. The corresponding classification results are given in the second row, where two observations are used and (10) is adopted as the optimization criterion. The light car is clearly distinguished from other categories, even if context-dependence is not used. This means light cars stand out distinctly from background objects and shadows.

By "dark cars," we mean the cars whose intensities are not so different from those of shadows. Dark cars are particularly problematic since they are easily confused with shadows by the tracker. The HMM using only the intensity observation allows definition of light cars in a robust way, but the model does not work well for dark cars. Part of a dark car is more likely to be classified as shadow. This is because the distributions of different categories overlap, namely, the gray-value of an HMM region that belongs to **F** (a dark car) also falls in the support of the shadow distribution and, moreover, the probability of the foreground is very low.

Introducing the second observation contributes to the robustness of foreground object recognition. With a 2D feature vector composed of the two observation symbols, the area proportion where the densities of different categories overlap becomes less than the area proportion with a 1D feature vector. Thus, the Bayes risk is reduced. To confirm effectiveness, we test a sequence in the same area by using intensity alone and by using wavelet coefficients together with intensities as the observations. The results with 1D and 2D features for a dark car (see the third row) are shown in the fourth and fifth rows, respectively. Equation (10) is used in both cases for state estimation. In the fourth row, only the light portions such as the roof and lamps are classified as foreground. A larger percentage of the dark car stands out more in the fifth row than in the fourth row. The sixth row is also the result of the same images. The difference from the fifth row is that we adopt (11) as the optimization criterion rather than (10). The state estimation based on (11) is applied to the area of interest in raster order and repeated three times. By incorporating the measure of the context-dependence, the results are improved.

Some results concerning the "gray car" are shown in the remaining images of Fig. 3. The images are shown in the seventh row. The results based on individual HMM regions are shown in the eighth row the results with the context-dependence among HMM regions are shown in the last row. A similar

misclassification problem also occurs with gray cars because the distribution of gray cars overlaps with the distribution of the background. However, since the variance of the background is usually much smaller than that of the shadow, the risk of a gray car being confused with the background is lower, as seen in Fig. 3.

The state estimation process has been implemented on an SGI O2 R5000 SC 180 entry-level desktop workstation and is able to run at the field-rate of 50 Hz (real time). Some video clips including these results are provided on our web page at URL <http://www.watanabe.nuie.nagoya-u.ac.jp/member/jien/demo.htm>.

## 6 CONCLUSION

We proposed a new HMM-based segmentation method which is able to model shadows as well as foreground and background regions. A considerable advantage of this model is that this model performs accurate segmentation of foreground objects out of background objects and shadows, as seen in Fig. 3. The model also enables the tracker to perform the state estimation in real time because the state estimation is based only on the past observation and the estimation criterion is defined recursively. Another advantage is that, unlike other approaches, it is no longer necessary to provide specific data for training. All the HMM parameters are estimated by the EM algorithm from an ordinary video sequence.

Our contrived techniques are especially useful for reliable extraction of the foreground objects. First, the state estimation algorithm is developed to perform context-dependent classification. By incorporating the measure of the context-dependence, the results are improved significantly. Second, introducing the second observation contributes to the robustness of foreground object recognition. Though the high frequency wavelet coefficients are effective for traffic monitoring movies, the choice of filters is open to further discussion. Finally, the constrained model imposes temporal continuity constraints on the foreground objects and contributes to the robustness of foreground object detection.

Based on the experimental results, we can see our proposed method works as a low-level car tracker. Since this low-level tracker runs comfortably in real time, it also offers the possibility of being used as a low-level component for a high-level tracking approach [8].

## ACKNOWLEDGMENTS

The authors would like to thank Professor Karen Fedderholdt for her valuable assistance.

## REFERENCES

- [1] L.E. Baum, T. Petrie, G. Soules, and N. Weiss, "A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains," *Ann. Math. Statistics*, vol. 41, no. 1, pp. 164-171, 1970.
- [2] A. Baumberg and D. Hogg, "Learning Flexible Models from Image Sequences," *Computer Vision-ECCV '94*, J.-O. Eklundh, ed., Springer-Verlag, vol. 1, pp. 299-308, 1994.
- [3] A. Blake and M. Isard, *Active Contours*. London: Springer, p. 352, 1998.
- [4] I. Daubechies, *Ten Lectures on Wavelets*. Philadelphia: SIAM, 1992.
- [5] A.P. Dempster, N.M. Laird, and D.R. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *J.R. Statistics Soc.*, vol. B, no 39, pp. 1-38, 1977.
- [6] G.D. Forney, "The Viterbi Algorithm," *Proc. IEEE*, vol. 61, pp. 268-278, Mar. 1973.
- [7] I. Haritaoglu, D. Harwood, and L.S. Davis, "W4 - a Real Time System for Detection and Tracking People and their Parts," *Proc. Third Face and Gesture Recognition Conf.*, pp. 222-227, 1998.
- [8] M. Isard and A. Blake, "CONDENSATION: Unifying Low-Level and High-Level Tracking in a Stochastic Framework," *Proc. Fifth European Conf. Computer Vision*, pp. 893-908, 1998.
- [9] J. Kato, T. Watanabe, and M. Yoneda, "HMM-Based Background-Object-Shadow Separation for Traffic Monitoring Movies," *Trans. Information Processing Society of Japan*, vol. 42, no. 1, pp. 1-15, 2001, (in Japanese).

- [10] D. Koller, J. Weber, and J. Malik, "Robust Multiple Car Tracking with Occlusion Reasoning," *Computer Vision-ECCV '94*, J.-O. Eklundh, ed., vol. 1, pp. 189-196, 1994.
- [11] N. Mine, Y. Yagi, and M. Yachida, "Detection of Change Region by Integrating Subtracted Image and Edge Boundary Image," *Trans. IEICE*, vol. J77-D-II, no. 3, pp. 631-634, 1994.
- [12] H. Nakai, "Non-Parameterized Bayes Decision Method for Moving Object Detection," *Proc. Second Asian Conf. Computer Vision*, pp. 447-451, Dec. 1995.
- [13] N. Paragiso and R. Deriche, "A PDE-Based Level Set Approach for Detection and Tracking of Moving Objects," Technical Report 3173, INRIA Sophia Antipolis, 1997.
- [14] L.R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257-286, Feb. 1989.
- [15] S. Rowe and A. Blake, "Statistical Mosaics for Tracking," *Image and Vision Computing*, vol. 14, pp. 549-564, 1996.
- [16] K. Toyama, J. Krumm, B. Brumitt, and B. Meyers, "Wallflower: Principles and Practice of Background Maintenance," *Proc. Seventh Int'l Conf. Computer Vision*, pp. 255-261, 1999.
- [17] A.J. Viterbi, "Error Bounds for Convolutional Codes and an Asymptotically Optimal Decoding Algorithm," *IEEE Trans. Information Theory*, vol. 13, pp. 260-269, Apr. 1967.
- [18] J.Z. Wang, J. Li, R.M. Gray, and G. Wiederhold, "Unsupervised Multi-resolution Segmentation for Images with Low Depth of Field," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 23, no. 1, pp. 85-90, Jan. 2001.

► For more information on this or any computing topic, please visit our Digital Library at <http://computer.org/publications/dilib>.