An HMM/MRF-Based Stochastic Framework for Robust Vehicle Tracking

Jien Kato, *Member, IEEE*, Toyohide Watanabe, *Member, IEEE*, Sébastien Joga, Ying Liu, and Hiroyuki Hase, *Associate Member, IEEE*

***Abstract*—Shadows of moving objects often obstruct robust visual tracking. In this paper, we present a car tracker based on a hidden Markov model/Markov random field (HMM/MRF)-based segmentation method that is capable of classifying each small region of an image into three different categories: vehicles, shadows of vehicles, and background from a traffic-monitoring movie. The temporal continuity of the different categories for one small region location is modeled as a single HMM along the time axis, independently of the neighboring regions. In order to incorporate spatial-dependent information among neighboring regions into the tracking process, at the state-estimation stage, the output from the HMMs is regarded as an MRF and the maximum *a posteriori* criterion is employed in conjunction with the MRF for optimization. At each time step, the state estimation for the image is equivalent to the optimal configuration of the MRF generated through a stochastic relaxation process. Experimental results show that, using this method, foreground (vehicles) and nonforeground regions including the shadows of moving vehicles can be discriminated with high accuracy.**

***Index Terms*—Hidden Markov model (HMM), image classification, image segmentation, Markov random field (MRF), traffic surveillance, vehicle tracking.**

I. Introduction

OVER THE last decade, applications to intelligent transportation systems (ITS) of visual motion analysis have emerged as one of the principal areas of research within the computer-vision community. This increasing interest is due in part to the falling cost of computing power and the storage that is necessary to process image sequences and, perhaps more primarily, because of a deepening recognition of the importance of grasping traffic situations by using visual information. For example, autonomous guided vehicles for driving on roads must track the features of the road [1], [2] and other moving vehicles around them [3]. Static surveillance systems are required to not only collect traffic data [4], [5], but also to analyze complicated environments; for instance, detecting abnormal events from intersection scenes [6], [7]. These tasks seem too difficult to be achieved only by counting on spot sensors such as loop detectors or supersonic wave sensors, because the acquired information by such sensors is very localized and limited to one specific kind. Many research efforts on visual motion analysis [1]–[3], [5]–[7], however, have proven to be useful and promising for these advanced and extensive applications in the ITS field.

Visual motion analysis from the view of ITS attempts to use tracking information to infer the behaviors of moving objects of interest (mainly vehicles) within a traffic scene. Therefore, reliable vehicle-tracking techniques are vital to this goal. In contrast with spot sensors, approaches based on computer-vision techniques are able to acquire various kinds of information from a wider scope, but, on the other hand, they are relatively lacking in reliability. The robustness of visual-tracking techniques has been an important and difficult problem that precludes them from being in practical use over a long period.

One of the main obstacles to the robustness of vehicle tracking is recognized as an illumination issue. Many previous attempts such as [8]–[11] have been made at addressing this problem. In an outdoor environment, it is not surprising that illumination varies within even a few seconds. The resultant variation of intensities of image pixels easily reaches such an extent that the tracking process fails to catch hold of the target objects. To resolve this problem, one idea, which is quite simple and straightforward to implement, is using illumination-invariant features such as differentiation of the image intensities among neighboring pixels [11]. Other frameworks for computing geometric and illumination invariants have been proposed for object recognition. These allow for changes of position and number of light sources, brightness and contrast, and even hue [8]–[10], [12] and also seem to be applicable to the vehicle-tracking problem. The methodologies adopted in these frameworks, however, mainly deal with illumination variation in top-down or high-angle view images, where the top-down or high-angle views alleviate the effects of shadows of moving vehicles and lead to limited success in one aspect of the illumination problem.

Actually, to acquire necessary information for ITS, front-view or low-angle view images also often need to be analyzed in many practical systems. That means that there is another important aspect of the illumination problem: shadows of moving objects. Previous work [13], [14] has revealed that in front-view or low-angle images, shadows of moving objects are one of the main factors that disturb robust visual tracking and, thus, need special attention. As a typical example, in active contour-based tracking with a high-level approach [15], [16], shadows of moving objects compete for the attention of

Manuscript received July 31, 2003; revised February 23, 2004. The Associate Editor was N. Papanikolopoulos.

J. Kato and T. Watanabe are with the Department of Systems and Social Informatics, Nagoya University, Nagoya 464-8603, Japan (e-mail: jien@is.nagoya-u.ac.jp; watanabe@is.nagoya-u.ac.jp).

S. Joga is with France Telecom R&D, 92794 Issy-Les-Moulineaux Cedex 9, France (e-mail: sebastien.joga@rd.francetelecom.com).

Y. Liu is with the Department of Intellectual Information Systems Engineering, Toyama University, Toyama 930-8555, Japan (e-mail: liuying@sysk.iis.toyama-u.ac.jp).

H. Hase is with the Department of Information Science, Fukui University, Fukui 910-8507, Japan (e-mail: hase@fuis.fuis.fukui-u.ac.jp).

Digital Object Identifier 10.1109/TITS.2004.833791

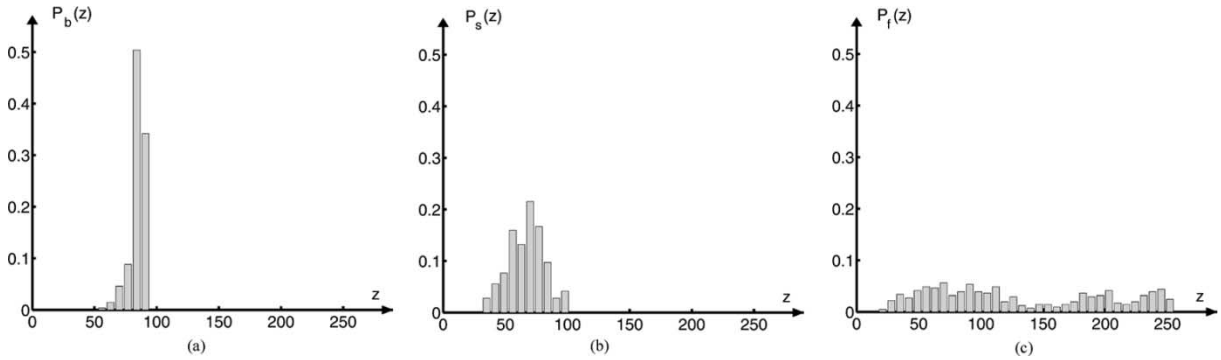

Fig. 1. Intensity profiles. Intensity values for a single region location are collected from a 30-s video sequence and are classified by hand into three categories: (a) background, (b) shadow, and (c) foreground. These histograms show a large amount of overlap among different categories.

the tracker and may succeed in pulling the tracker away from foreground (target) objects, since the shadows appear to be similar to the target objects and behave in the same way as they do. This type of illumination problem is more intractable than that in the aforementioned case, where disturbance of shadows originates from only the static background objects. The illumination-invariant techniques [8]–[12] are, of course, powerless against this kind of problem.

Earlier researchers have adopted image-differentiation techniques including both background subtraction and interframe differentiation [5], [17], [18] to repress background clutters and make the trackers reliable. Background subtraction is based on the assumption that the background is basically static. Obviously, this method cannot remove the shadows of moving objects. On the other hand, computing the interframe differentiation means approximating the derivative of the image with respect to time. By using this technique, the shadows of moving objects appear on the resulting image as outlines, but at the cost of removing the homogeneous regions inside foreground objects from the resulting image because the interframe differentiation of these regions is small. As the simple background subtraction or interframe differentiation schemes are known to perform poorly in repressing background clutters, a number of researchers have attempted to construct a probabilistic background model [5], [14], [19], [20] for the same purpose. For example, Haritaoglu *et al.* [19] modeled the background by simply learning the minimal and maximal gray-value intensity for every pixel location. Rowe *et al.* [14] established an intensity distribution of background pixels for the special case of a pan-tilt head camera, where there is no global threshold available for foreground–background separation. Toyama *et al.* [20] addressed the problem of background maintenance by using a multilayered approach. Unfortunately, none of the above models are able to deal with the shadows of moving objects.

To enhance the robustness of visual tracking against the shadows of moving objects, this paper proposes a hidden Markov model/Markov random field (HMM/MRF)-based stochastic framework for modeling not only foreground/background objects, but also the shadows in traffic-monitoring image sequences. This approach is rooted in the idea of enhancing immunity to the distracting objects by modeling them, so that when foreground events happen against a cluttered

Fig. 2. Scene taken from a highway sequence; the vehicles running on the right lane are shadowed by the vehicle running on the left lane. The data shown in Fig. 1 are collected at the position of the cross.

environment, the features that appear to match the distracting objects can be ignored by the tracker. The proposed framework is able to work as a low-level tracker itself and also as a probabilistic background model for a high-level tracking process. Here, we use the term "low-level" for the tracking methods that use only very weak assumptions about the object of interest in the image processing stage [19], [21]–[23] and "high-level" for the methods that make use of strong modeling constraints to guide even the lowest level image processing stages [15], [16].

II. Approach

Since we would like to deal with shadows as well as foreground and background objects, we consider shadow, foreground, and background as three independent categories and abbreviate them as $\mathbf{S}$, $\mathbf{F}$, and $\mathbf{B}$. It is useful to investigate the intensity variation of these categories at a given pixel first. Fig. 1 shows the intensity profiles for $\mathbf{B}$, $\mathbf{S}$, and $\mathbf{F}$ collected from a 30-s video sequence at the pixel position of images indicated in Fig. 2. These profiles are classified by hand.

The background has a quite narrow and symmetric distribution. Except for congested traffic, the background usually occupies the main area of the distribution at a given pixel. Shadows are those pixels that receive less light, so they have lower intensity than the background. Both background and shadow distri-

butions partially overlap each other. The distributions of $\mathbf{B}$ and $\mathbf{S}$ seem to be suitable to be approximated by Gaussian densities. On the other hand, the distribution of foreground (vehicles) usually covers a much larger range of gray values. Therefore, a reasonable model of the intensity variation for $\mathbf{F}$ could be a uniform distribution.

There exist two reasons that rule out the consideration of constructing a traditional background model as described in [5], [14], [19], and [20]. First, because the distributions for three different categories overlap each other, it is difficult to find a robust fitting method for learning the necessary parameters of distribution functions from data. In other words, learning the three distributions separately is not a robust way. Second, compared with the magnitude of the probability that a pixel belongs to background, the probability that a pixel belongs to foreground is often insignificant. This probability, however, grows if the fact that the intensities at a given pixel from two consecutive images are dependent is taken into account. If at a given time a pixel belongs to foreground (a vehicle), it is likely that at the next time step it will still belong to foreground (the same vehicle). This temporal continuity, namely, occurrence of any category for a period of time, calls for a model that can incorporate temporal information well.

An HMM is particularly suitable for this task because it is a double stochastic process with an underlying stochastic process that is not observable (hidden) and can only be observed through another set of stochastic processes that produces the sequence of observed symbols [24]. Consider a pixel in the image. At any time, an intensity value can be observed. Observed intensity values depend on an underlying process that is not observed. This underlying process explains the transitions between hidden categories ($\mathbf{B}$, $\mathbf{S}$, and $\mathbf{F}$) at each time step. As mentioned above, it is, for example, more likely to go from $\mathbf{B}$ to $\mathbf{B}$ than from $\mathbf{B}$ to $\mathbf{S}$ or from $\mathbf{B}$ to $\mathbf{F}$. This property can be appropriately characterized by a first-order Markov chain embedded in an HMM. From this viewpoint, the problem of segmentation of an image sequence into three categories can be solved by building an HMM that explains and characterizes the occurrence of the observed intensities from the image sequence. The probability distributions for each category or, in the usual notation of an HMM, the state, are allowed to be automatically learned from an ordinary (state mixed) image sequence.

We divide each image of a traffic-monitoring movie into nonoverlapping small regions (termed HMM regions) and model the observations over time at one region location as a single HMM along the time axis, independent of the neighboring region. One important issue with respect to this approach is how to take spatial dependence among neighboring regions into account. This is because, in general, a scene is understood in not only the temporal but also the spatial context of objects within this scene. An easy example to understand is that a foreground region is highly unlikely to exist in isolation surrounded by background regions. Such spatial context constraints also need to be entertained in interpretation of visual information at each time step, as well as temporal context information. Our previous work [25] and [26] has attempted to realize spatio-temporal-dependent modeling for traffic monitoring movies, but in either [25] or [26], the issue of modeling spatial context of image sequences remains essentially unsolved.

The MRF theory provides a convenient way to model context-dependent entities such as image pixels and correlated features through characterizing mutual influences among the entities using conditional MRF distributions [27]. Furthermore, MRF used in conjunction with statistical criteria for optimization, e.g., maximum *a posteriori* (MAP), enables us to formulate an objective function in terms of an MAP optimization principle. We incorporate this MAP–MRF schema into our HMM-based approach to achieve context-dependent classification in both a temporal and spatial sense. At each time step, instead of estimating the state for each region individually, the outputs of HMMs corresponding to all region locations are modeled as an MRF; this MRF together with the MAP principle is employed to obtain an optimal state combination for the image. By integrating an MRF model into the state-estimation phase of HMMs in this way, we achieve spatiotemporal-dependent modeling for traffic-monitoring movies.

III. Classification of Regions Using HMMs

The mechanism of an HMM is summarized as follows. There is a finite number of hidden states in the model that represent abstract quantities (e.g., categories $\mathbf{F}$, $\mathbf{B}$, and $\mathbf{S}$) corresponding to the "clusters" of contexts that have similar probability distributions of the observations. At each discrete unit of time, a new state is entered based upon a transition probability distribution that depends on the immediately previous state (Markovian property). After each transition takes place, an observation output is produced according to a probability distribution that is held fixed only for the current state. An HMM is used in two stages: parameter learning and state estimation. In the first stage, given a sequence of observation (learning sequence), the parameters are learned to maximize the probability of the observations given the model. In the second stage, given another sequence of observations (test sequence), a state sequence that is optimal in some meaningful sense is chosen.

In this section, we first define the model for our particular purpose, formulate unknown model parameters by Baum–Welsh reestimation formulas, and give the reestimation algorithm. Then, we describe how to initialize various parameters and discuss the problem of state estimation with a view to achieving real-time vehicle tracking by classifying each HMM region into different categories over time.

A. Definition of the Model

We start with the following notations:

$T =$ length of the observation sequence (number of image fields);

$Q =$ $\{q_{\mathbf{B}} = \mathbf{B}, q_{\mathbf{S}} = \mathbf{S}, q_{\mathbf{F}} = \mathbf{F}\}$, states;

$M_i =$ number of possible observations;

$\boldsymbol{x} =$ $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T\}$, observation sequence;

$$A = \begin{bmatrix} a_{\mathbf{BB}} & a_{\mathbf{BS}} & a_{\mathbf{BF}} \\ a_{\mathbf{SB}} & a_{\mathbf{SS}} & a_{\mathbf{SF}} \\ a_{\mathbf{FB}} & a_{\mathbf{FS}} & a_{\mathbf{FF}} \end{bmatrix},$$

$a_{ij} = \Pr(q_j \text{ at } t+1 | q_i \text{ at } t)$, state transition probability matrix;

The Forward-backward Procedure

1) Initialization

$$\begin{aligned} \alpha_1(i) &= \pi_i p_i(\boldsymbol{x}_1), \quad i \in \{\mathbf{B},\mathbf{S},\mathbf{F}\}, \\ \beta_T(i) &= 1, \quad i \in \{\mathbf{B},\mathbf{S},\mathbf{F}\}. \end{aligned}$$

2) Induction

$$\begin{aligned} \alpha_{t+1}(j) &= \left(\sum_i \alpha_t(i)a_{ij}\right) p_j(\boldsymbol{x}_{t+1}), \quad t = 1,\ldots,T-1, \quad j \in \{\mathbf{B},\mathbf{S},\mathbf{F}\}, \\ \beta_t(i) &= \sum_j a_{ij} p_j(\boldsymbol{x}_{t+1})\beta_{t+1}(j), \quad t = T-1,\ldots,1, \quad i \in \{\mathbf{B},\mathbf{S},\mathbf{F}\}. \end{aligned}$$

3) Termination

$$\Pr(X|\omega) = \sum_i \alpha_T(i).$$

Fig. 3. Algorithm to calculate the auxiliary probabilities.

$P =$ $\{p_i(\boldsymbol{y})\}$, probability of observing the observation $\boldsymbol{y}$ in state i;

$\Pi =$ $\{\pi_{\mathbf{B}}, \pi_{\mathbf{S}}, \pi_{\mathbf{F}}\}$, $\pi_i = \Pr(q_i \text{ at } t = 1)$, initial state probability.

As described in Section II, the background and shadow observation probabilities are modeled by Gaussian densities and the foreground observation probability by a uniform density. Hence

$$p_i(\boldsymbol{x}) = \frac{1}{2\pi\sqrt{\det\boldsymbol{\Sigma}_i}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)\right\}, i \in \{\mathbf{B},\mathbf{S}\} \tag{1}$$

$$p_{\mathbf{F}}(\boldsymbol{x}) = \frac{1}{M_1 \times M_2} \tag{2}$$

where $\boldsymbol{x}$ stands for a pattern vector consisting of two kinds of observations (the issue regarding observations will be discussed later) and, consequently, $M_1 \times M_2$ means the domain of the pattern vectors. The approximation using Gaussian densities leads to the advantage of characterizing background and shadow only by mean vectors ($\boldsymbol{\mu}_i$) and covariance matrices ($\boldsymbol{\Sigma}_i$).

B. Baum–Welsh Reestimation Formulas

Let $\omega = \{A, P, \Pi\}$ denote the model. Adjusting parameters ω to fit the observations is equivalent to the problem of finding such that ω maximizes the likelihood $L(\boldsymbol{x}, \omega) = \log[\Pr(\boldsymbol{x}|\omega)]$ for a given set $\boldsymbol{x}$ of observed data. Due to the existence of hidden variables in the model, there is no known way to solve this problem analytically. An expectation–maximization (EM) algorithm is usually used [24]. EM algorithms are iterative procedures that produce a sequence of estimates for ω, given data $\boldsymbol{x}$, so that each estimate $\omega^{(m+1)}$ has a greater value of the likelihood L than the preceding estimate $\omega^{(m)}$. We utilize the Baum–Welsh reestimation formulas, a special case of the EM algorithms, to learn the unknown model parameters.

To introduce the Baum–Welsh reestimation formulas, we first define two important auxiliary probabilities, the so-called forward and backward probabilities. They are

$$\alpha_t(i) = \Pr(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t, q_t = i|\omega) \tag{3}$$

and

$$\beta_t(i) = \Pr(\boldsymbol{x}_{t+1}, \ldots, \boldsymbol{x}_T|q_t = i, \omega) \tag{4}$$

where q_t means the state at time step t. The probabilities can be solved inductively using the algorithm shown in Fig. 3 and are used to calculate other auxiliary probabilities

$$\gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{\Pr(\boldsymbol{x}|\omega)} \tag{5}$$

and

$$\xi_t(i,j) = \frac{\alpha_t(i)a_{ij}p_j(\boldsymbol{x}_{t+1})\beta_{t+1}(j)}{\Pr(\boldsymbol{x}|\omega)} \tag{6}$$

which need to define the Baum–Welsh reestimation formulas.

The Baum–Welsh reestimation formulas for calculating A, P, and Π in our problem are given in Fig. 4, where a notion of the expected frequency of event occurrence is taken into account. To reestimate the means and covariances for $\mathbf{B}$ and $\mathbf{S}$, usual estimators are used, but the sums are weighted by the probabilities of being in the corresponding states.

C. Initialization

An EM-type algorithm achieves local maximization. The convergence rate of reestimation could be different depending on the initial estimates for the parameters and, also, inappropriate initial parameters could lead to an implausible solution. Thus, parameter initialization is important.

To set initial parameters properly, we define some typical values relative to time. Let $\tau_{\mathbf{B}}$ be the typical length of time a pixel belongs to background and $\tau_{\mathbf{S}}$ and $\tau_{\mathbf{F}}$ be the same time constants for shadow and foreground. Also, assume that $\lambda_{\mathbf{B}}$, $\lambda_{\mathbf{S}}$, and $\lambda_{\mathbf{F}}$ are the proportions of the time spent in $\mathbf{B}$, $\mathbf{S}$, and $\mathbf{F}$ ($\lambda_{\mathbf{B}} + \lambda_{\mathbf{S}} + \lambda_{\mathbf{F}} = 1$). Using these definitions, the initial state transition matrix is chosen as

$$A = \begin{bmatrix} 1-\frac{1}{\tau_{\mathbf{B}}} & \frac{1}{\tau_{\mathbf{B}}}\Lambda_{\mathbf{SF}} & \frac{1}{\tau_{\mathbf{B}}}\Lambda_{\mathbf{FS}} \\ \frac{1}{\tau_{\mathbf{S}}}\Lambda_{\mathbf{BF}} & 1-\frac{1}{\tau_{\mathbf{S}}} & \frac{1}{\tau_{\mathbf{S}}}\Lambda_{\mathbf{FB}} \\ \frac{1}{\tau_{\mathbf{F}}}\Lambda_{\mathbf{SB}} & \frac{1}{\tau_{\mathbf{F}}}\Lambda_{\mathbf{SB}} & 1-\frac{1}{\tau_{\mathbf{F}}} \end{bmatrix}$$

$$\Lambda_{ij} = \frac{\lambda_i}{(\lambda_i + \lambda_j)} \tag{7}$$

and the initial state distribution is chosen to be

$$\Pi = \{\lambda_{\mathbf{B}}, \lambda_{\mathbf{S}}, \lambda_{\mathbf{F}}\}. \tag{8}$$

As to initial parameters of the observation densities such as intensity, the mean for background state $\mu_{\mathbf{B}}^{\text{Int}}$ is estimated to be

The Baum-Welsh Reestimation Formulas

1) Initial state probability

$$\overline{\pi}_i = \gamma_1(i), \quad i \in \{\mathbf{B}, \mathbf{S}, \mathbf{F}\}.$$

2) State transition probability matrix

$$\overline{a}_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)}, \quad i,j \in \{\mathbf{B}, \mathbf{S}, \mathbf{F}\}.$$

3) Observation probabilities

$$\overline{\boldsymbol{\mu}}_i = \frac{\sum_{t=1}^{T} \boldsymbol{x}_t \gamma_t(i)}{\sum_{t=1}^{T} \gamma_t(i)}, \quad i \in \{\mathbf{B}, \mathbf{S}\},$$

$$\overline{\boldsymbol{\Sigma}}_i = \frac{\sum_{t=1}^{T} \gamma_t(i)(\boldsymbol{x}_t - \boldsymbol{\mu}_i)(\boldsymbol{x}_t - \boldsymbol{\mu}_i)^T}{\sum_{t=1}^{T} \gamma_t(i)}, \quad i \in \{\mathbf{B}, \mathbf{S}\}.$$

Fig. 4. Algorithm to reestimate the parameters of the HMM.

the mode of the intensities at a given region, since $\lambda_{\mathbf{B}} \gg \lambda_{\mathbf{S}}$ and $\lambda_{\mathbf{B}} \gg \lambda_{\mathbf{F}}$. The variance $\sigma_{\mathbf{B}}^{\mathrm{Int}}$ is determined empirically. The initial parameters of the observation density for shadow are based upon the assumption that the shadow is darker than the background. Let the support of a Gaussian be $[\mu - 2\sigma, \mu + 2\sigma]$; $\mu_{\mathbf{S}}^{\mathrm{Int}}$ and $\sigma_{\mathbf{S}}^{\mathrm{Int}}$ are chosen as

$$\mu_{\mathbf{S}}^{\mathrm{Int}} = \frac{\mu_{\mathbf{B}}^{\mathrm{Int}} + 2\sigma_{\mathbf{B}}^{\mathrm{Int}}}{2}, \quad \sigma_{\mathbf{S}}^{\mathrm{Int}} = \frac{\mu_{\mathbf{S}}^{\mathrm{Int}}}{2} \tag{9}$$

so that the shadow distribution support goes from 0 to the upper limit of the background distribution support. This ensures that $\mu_{\mathbf{S}}^{\mathrm{Int}} < \mu_{\mathbf{B}}^{\mathrm{Int}}$, when $\mu_{\mathbf{B}}^{\mathrm{Int}} > 2\sigma_{\mathbf{S}}^{\mathrm{Int}}$, i.e., the background intensities are not as low as the intensities in shadow regions. Obviously, all these initial parameters meet the stochastic constraints for HMM parameters $\sum_i \Pi_i = 1$, $\sum_j a_{ij} = 1$, and $\sum_{\boldsymbol{y}} p_i(\boldsymbol{y}) = 1$.

D. State Estimation for the Individual Region

We have discussed the issue of parameter learning using an available observation sequence for training. The major concern of this work is vehicle tracking or, in other words, segmentation of the images into different categories over time. This means that an "optimal" state sequence associated with another observation sequence (test sequence) has to be chosen, given the model and estimated parameters. This issue is called state estimation.

Several optimization criteria are possible. One is the Bayesian classifier [28], which maximizes $\Pr(q_t|\boldsymbol{x}_t) = \Pr(\boldsymbol{x}_t|q_t)\Pr(q_t)/\Pr(q_t, \boldsymbol{x}_t)$. However, since this method does not take the temporal information of a test sequence into account, the probability of being in the foreground state is usually underestimated. On the other hand, because the basic requirement for state estimation is that it must work in real time, we cannot adopt criteria that need the whole sequence of observations (including the future observed data), such as the Viterbi algorithm [24]. Furthermore, in the face of reducing processing time, an algorithm that can be defined recursively is preferable.

A solution is to maximize the probability of the state at time step t given only past observations $\Pr(q_t|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t, \omega)$. This probability can be implemented recursively using the forward probabilities alone, i.e.,

$$\Pr(q_t|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t, \omega) = \frac{\alpha_t(i)}{\sum_{i \in \{\mathbf{B},\mathbf{S},\mathbf{F}\}} \alpha_t(i)}. \tag{10}$$

Denote $\mathcal{P}_i^j = \Pr(q_t = i|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t, \omega)$. The state for each individual HMM region j can, thus, be estimated as

$$i^* = \arg\max_{i \in \{\mathbf{B},\mathbf{S},\mathbf{F}\}} \left\{\mathcal{P}_i^j\right\}. \tag{11}$$

Processing based on this criterion can be achieved in real time.

How to perform spatial-dependent state estimation given the output from the HMMs will be described in the next section.

IV. Spatial-Dependent State Estimation

It was mentioned in Section III-D that, at each time step, an output from the HMM for each small region location j is $\left\{\mathcal{P}_{\mathbf{B}}^j, \mathcal{P}_{\mathbf{S}}^j, \mathcal{P}_{\mathbf{F}}^j\right\}$, a set of probabilities for this region that belongs to $\mathbf{B}$, $\mathbf{S}$, and $\mathbf{F}$, respectively. Because of overlapping among learned observation distributions for different states, especially between foreground and shadow, foreground regions contrasting less with shadow regions may have $\mathcal{P}_{\mathbf{F}}^j \leq \mathcal{P}_{\mathbf{S}}^j$. Obviously, simply using (11) to do the state estimation for each individual region independently will fail in such a case.

In this section, we propose a state-estimation method that is possible to segment out foreground regions according to even poor results from the HMMs, through an MRF, to take the spatial-dependent information into account. In Sections IV-A–D, we first give basic concepts and notations of MRFs, then describe two important aspects of this work—deriving the MRF model and estimating involved parameters—and, finally, discuss how to find the MAP solution for our problem.

A. MRF

We first introduce some necessary notations and basic concepts about MRFs.

$\boldsymbol{\mathcal{S}} = \{1, \ldots, N\}$ set of sites denoting a lattice for a two-dimensional (2-D) image of size $m \times n = N$;

$\mathcal{N}_i =$ set of neighbors of site i;
$\mathcal{N} = \{\mathcal{N}_i | \forall\, i \in \mathcal{S}\}$ neighborhood system for $\mathcal{S}$ that maintains the interrelationship between sites;
$c \subseteq \mathcal{S}$ clique of $\mathcal{S}$ consisting either of a single site or of several sites that are all neighbors of each other;
$\mathcal{C} = \mathcal{C}_1 \cup \mathcal{C}_2 \ldots$ collection of all cliques for $(\mathcal{S}, \mathcal{N})$.

Let $\boldsymbol{u} = \{u_i, i \in \mathcal{S}\}$ be a collection of random variables defined on $\mathcal{S}$. Then, $\boldsymbol{u}$ is said to be an MRF if: 1) all of its realizations (in MRF terms, configurations) have nonzero probabilities and 2) its conditional distribution satisfies the Markov property

$$p(u_i | \boldsymbol{u}_{\mathcal{S}-i}) = p\left(u_i | \boldsymbol{u}_{\mathcal{N}_i}\right). \tag{12}$$

If $\boldsymbol{u}$ is an MRF, it is well known that the joint probability distribution of $\boldsymbol{u}$ is a Gibbs distribution [29], given by

$$p(\boldsymbol{u}) = Z^{-1} \exp\left\{\frac{-U(\boldsymbol{u})}{T}\right\} \tag{13}$$

where $U(\boldsymbol{u})$ is called an energy function with

$$U(\boldsymbol{u}) = \sum_{c \in \mathcal{C}} V_c(\boldsymbol{u}). \tag{14}$$

$V_c(\boldsymbol{u})$'s are the clique potentials that only depend upon the random variables defined on the sites in given clique c and Z is the normalizing constant, the so-called partition function, with

$$Z = \sum_{\boldsymbol{u} \in \mathbb{F}} \exp\left\{\frac{-U(\boldsymbol{u})}{T}\right\} \tag{15}$$

where $\mathbb{F}$ means the configuration space.

B. Posterior Distribution Modeling

The outputs from all the region locations of the image at time step t obviously constitute a Markov random field defined on a label set $\mathcal{L} = \{\mathbf{B}, \mathbf{S}, \mathbf{F}\}$. Let $\boldsymbol{D} = \left\{\left(\mathcal{P}^j_{\mathbf{B}}, \mathcal{P}^j_{\mathbf{S}}, \mathcal{P}^j_{\mathbf{F}}\right), j \in \mathcal{S}\right\}$ be the observed data and $\boldsymbol{u}$ be a configuration of the MRF. The method we adopt to solve our problem within an MAP–MRF schema finds

$$\boldsymbol{u}^* = \arg\max_{\boldsymbol{u} \in \mathbb{F}} \{p(\boldsymbol{D}|\boldsymbol{u}) p(\boldsymbol{u})\}. \tag{16}$$

This section focuses on the issue of modeling the posterior distribution $p(\boldsymbol{u}|\boldsymbol{D}) \propto p(\boldsymbol{D}|\boldsymbol{u}) p(\boldsymbol{u})$.

First, let us consider the prior distribution $p(\boldsymbol{u})$. Define $\mathcal{S}$ as a 2-D lattice with such a second-order neighborhood system, as shown in Fig. 5. At the state-estimation phase, since it is no longer necessary to discriminate between shadow and background, the category of $\mathbf{S}$ is once merged into the background for simplicity (denote the enlarged $\mathbf{B}$ as $\mathbf{B}^+$). This advantage of taking values on a binary label set $\mathcal{L} = \{\mathbf{F}, \mathbf{B}^+\}$ (equivalently on the label set $\mathcal{L} = \{0, 1\}$) enables us to adopt a simple MRF, called the autologistic model [30], as the prior distribution for $\boldsymbol{u}$. In Gibbs form, its energy function can be written as

$$U(\boldsymbol{u}|\Phi_{\boldsymbol{u}}) = \sum_{\langle i \rangle \in \mathcal{C}_1} \alpha_i u_i + \sum_{\langle i, i' \rangle \in \mathcal{C}_2} \beta_{i,i'} u_i u_{i'} \tag{17}$$

where $\Phi_{\boldsymbol{u}} = (\{\alpha_i, \beta_{i,i'}\}, i, i' \in \mathcal{S})$ denotes the set of parameters for the distribution of $\boldsymbol{u}$. This model conveys spatial-context information with its simple form and low computational cost:

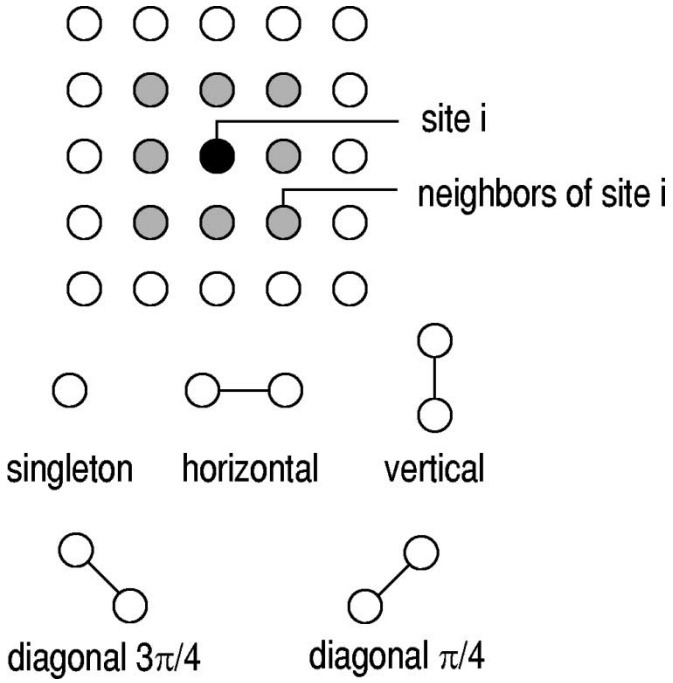

Fig. 5. 2-D lattice with a second-order neighborhood system, where the neighbors of site i are the eight sites that immediately surround it. Five types of cliques including a single-site clique and four pair-site cliques are taken into account.

only single- and pair-site cliques (totally five types, as shown in Fig. 5) are taken into account. The parameters α_i and $\beta_{i,i'}$ measure the external field and bonding strengths, respectively, among the sites neighboring each other. Learning these parameters from the training data will be discussed in Section V.

Second, let us consider the likelihood model $p(\boldsymbol{D}|\boldsymbol{u})$. We construct this model based on the idea of random replacement. Let $\boldsymbol{D} = \{d_j, j \in \mathcal{S}\}$, where $d_j \in \{\mathbf{F}, \mathbf{B}^+\}$ means the value every region takes. The initial value of d_j can be obtained immediately from the state estimation for individual region location, namely, $d = i^* = \arg\max_{i \in \{\mathbf{F}, \mathbf{B}^+\}} \{\mathcal{P}^j_i\}$. A d_j is transformed into u_j according to the following likelihood probabilities:

$$p(d_j = \vartheta | u_j = \vartheta') = \begin{cases} \mathcal{P}^j_{\mathbf{A}}, & \text{if } \vartheta = \vartheta' = \mathbf{A},\ \mathbf{A} \in \{\mathbf{F}, \mathbf{B}^+\} \\ \mathcal{P}^j_{\mathbf{F}}, & \text{if } \vartheta = \mathbf{B}^+,\ \vartheta' = \mathbf{F} \\ \mathcal{P}^j_{\mathbf{B}^+}, & \text{if } \vartheta = \mathbf{F},\ \vartheta' = \mathbf{B}^+. \end{cases} \tag{18}$$

This means that a label value remains unchanged with the probability of the state indicated by the label and changes to any other label value with the probability of the state specified by the new label. This describes the transition from one state to another at a region location. Transition probabilities $\mathcal{P}^j_{\mathbf{A}}$ are in inverse proportion to the likelihood energy $U = -\ln \mathcal{P}^j_{\mathbf{A}}$. Thus, the likelihood model ensures that a configuration fitter for the data is more optimal and more likely to be accepted.

In consideration of the fact that the probabilities used in the likelihood model are obtained for each region location independently, we assume that the variables d_j associated with the sites in the neighborhood $\mathcal{N}_j$ are mutually independent. This gives the following simple product for the likelihood density, i.e.,

$$p(\boldsymbol{D}|\boldsymbol{u}) = \prod_{j \in \mathcal{S}} p(d_j | u_j). \tag{19}$$

The remaining work is to learn the unknown parameters and find the configuration $\boldsymbol{u}$ of the MRF model that maximizes the posterior distribution for a fixed $\boldsymbol{D}$ or to minimize

$$U^{\mathrm{Pos}}(\boldsymbol{u}) = U(\boldsymbol{u}|\Phi_{\boldsymbol{u}}) + U(\boldsymbol{D}|\boldsymbol{u}), \tag{20}$$

where the prior and likelihood energy functions are specified by (17) and (18), respectively.

C. Parameter Learning for the MRF

The Gibbs distribution of the MRF is further supposed to be both homogeneous and isotropic. Thus, we have $\alpha_i = \alpha$ and $\beta_{i,i'} = \beta$, regardless of either positions of the sites or directions of the cliques. This simplification leads to

$$U(\boldsymbol{u}|\Phi_{\boldsymbol{u}}) = \alpha \sum_{\langle i\rangle \in \boldsymbol{\mathcal{C}}_1} u_i + \beta \sum_{\langle i,i'\rangle \in \boldsymbol{\mathcal{C}}_2} u_i u_{i'} \tag{21}$$

the advantage of reducing the number of unknown model parameters to two.

Parameter learning can be regarded as the problem to resolve

$$\Phi^*_{\boldsymbol{u}} = \arg\max_{\Phi_{\boldsymbol{u}}} \{p(\boldsymbol{u}|\Phi_{\boldsymbol{u}})\} \tag{22}$$

by maximum-likelihood (ML) estimation. However, the ML estimation is rarely achievable just as the definition is. This is because, in the Gibbs form of $p(\boldsymbol{u}|\Phi_{\boldsymbol{u}})$, the partition function Z is also a function of $\Phi_{\boldsymbol{u}}$ and, moreover, has to be calculated by summing over all possible realizations in the configuration space. This difficulty obliges us to use an approximate scheme called the coding method [29].

We partition $\boldsymbol{\mathcal{S}}$ into four disjointed sets $\boldsymbol{\mathcal{S}}^{(k)} (k = 1, 2, 3, 4)$, called codings, so that no two sites in one coding are neighbors (see Fig. 6). Under the Markov assumption (12), the variables associated with the sites in one $\boldsymbol{\mathcal{S}}^{(k)}$ are mutually independent, given the labels at all other sites.

According to (13) and (17), the conditional probability for the prior model with $\mathcal{L} = \{0, 1\}$ can be written as

$$p(u_i|\boldsymbol{u}_{\boldsymbol{\mathcal{N}}_i}) = \frac{\exp\left\{-\left(\alpha u_i + \beta u_i \sum_{i' \in \boldsymbol{\mathcal{N}}_i} u_{i'}\right)/T\right\}}{1 + \exp\left\{-\left(\alpha + \beta \sum_{i' \in \boldsymbol{\mathcal{N}}_i} u_{i'}\right)/T\right\}}. \tag{23}$$

Taking the probability independence within a coding into account, the likelihood for this model can be calculated by a simple product, i.e.,

$$p^{(k)}(\boldsymbol{u}|\Phi_{\boldsymbol{u}}) = \prod_{i \in \boldsymbol{\mathcal{S}}^{(k)}} \frac{\exp\left\{-\left(\alpha u_i + \beta u_i \sum_{i' \in \boldsymbol{\mathcal{N}}_i} u_{i'}\right)/T\right\}}{1 + \exp\left\{-\left(\alpha + \beta \sum_{i' \in \boldsymbol{\mathcal{N}}_i} u_{i'}\right)/T\right\}} \tag{24}$$

which dispenses with the need to evaluate the partition function Z. Its log likelihood takes the form

$$\ln p^{(k)}(\boldsymbol{u}|\Phi_{\boldsymbol{u}}) = -\sum_{i \in \boldsymbol{\mathcal{S}}^{(k)}} \left\{ \alpha u_i + \beta u_i \sum_{i' \in \boldsymbol{\mathcal{N}}_i} u_{i'} + \ln\left[1 + \exp\left(-\alpha - \beta \sum_{i' \in \boldsymbol{\mathcal{N}}_i} u_{i'}\right)\right]\right\}. \tag{25}$$

Four sets of the ML estimates for $\{\hat{\alpha}^{(k)}, \hat{\beta}^{(k)}\}$ can be obtained by solving

$$\frac{\partial \ln p^{(k)}(\boldsymbol{u}|\alpha, \beta)}{\partial \alpha} = 0, \quad \frac{\partial \ln p^{(k)}(\boldsymbol{u}|\alpha, \beta)}{\partial \beta} = 0 \tag{26}$$

and any set can be used for state estimation.

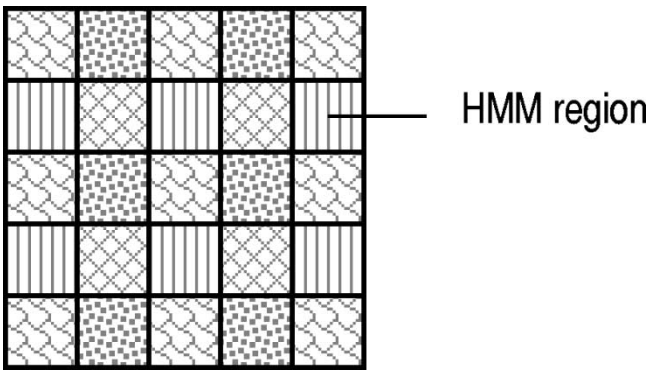

Fig. 6. Dividing the image into four codings for learning MRF parameters.

D. State Estimation for Images

Given the model and the results from the HMMs, we search the optimal configuration through a stochastic relaxation process at each time step. The stochastic relaxation process includes a Gibbs sampler that works under a simulated annealing algorithm [31]. A clear and comprehensive treatment of the stochastic relaxation process can be found in much of the literature, such as [27] or [30]. In the following, we focus only on how to find the MAP solution for state estimation by use of this method.

In our approach, the Gibbs sampler generates a Markov chain $\{\boldsymbol{u}\langle 0\rangle, \ldots, \boldsymbol{u}\langle t\rangle, \ldots\}$ that converges to the Gibbs distribution $p(\boldsymbol{u})$ as $t \to \infty$, regardless of the initial configuration $\boldsymbol{u}\langle 0\rangle$. Here, $\langle\cdot\rangle$ represents the time in the stochastic relaxation process. We select the starting configuration as $\boldsymbol{D} = \{d_j, j \in \boldsymbol{\mathcal{S}}\}$, where $d_j = i^* = \arg\max_{i \in \{\mathbf{F}, \mathbf{B}^+\}} \left\{\mathcal{P}_i^j\right\}$, the results from the HMMs.

At each time step, only one site undergoes a possible change so that $\boldsymbol{u}\langle t-1\rangle$ and $\boldsymbol{u}\langle t\rangle$ can differ in, at most, one region location on an image. Let $n_1, \ldots, n_t, \ldots$ be the sequence in which the sites are visited for possible replacement and $\boldsymbol{u}\langle t\rangle = \{u_j\langle t\rangle, j = n_1, \ldots, n_t, \ldots, n_N\}$ be the configuration. At time t. We choose a state x following the conditional distribution specified by (23), given the observed states of the neighboring sites $u_r\langle t-1\rangle$, $r \in \boldsymbol{\mathcal{N}}_{n_t}$. The new configuration $\boldsymbol{u}\langle t\rangle$ has $u_{n_t}\langle t\rangle = x$ and $u_m\langle t\rangle = u_m\langle t-1\rangle$, $m \neq n_t$. The sequence $\{n_t\}$ we actually adopt is simply the same as the raster order.

The temperature T is lowered according to an annealing schedule defined as

$$T(y) = \frac{C}{\ln(1+y)}, \quad 1 \leq y \leq Y \tag{27}$$

where C is a constant, $T(y)$ indicates the temperature during the yth iteration of the annealing algorithm, and Y is the total number of iterations. This schedule controls T so that it is decreased very slowly, particularly near the freezing point, so that a configuration with the global minimum of Gibbs *posteriori* energy [see (20)] can finally be found. This configuration gives the state estimation for an image at the current time step.

V. Experiments

This section describes several experiments performed to: 1) estimate and evaluate the model parameters and 2) observe the performance of the spatio-temporal-dependent modeling

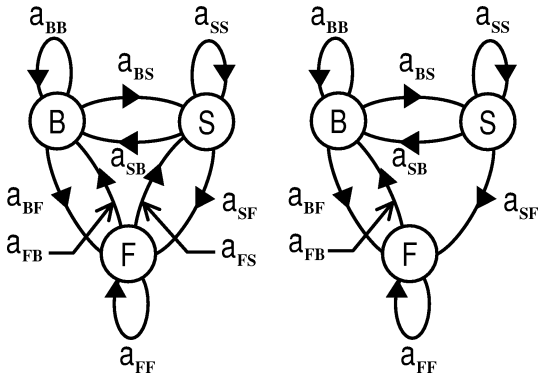

Fig. 7. (a) Ergodic HMM and (b) constrained HMM are shown.

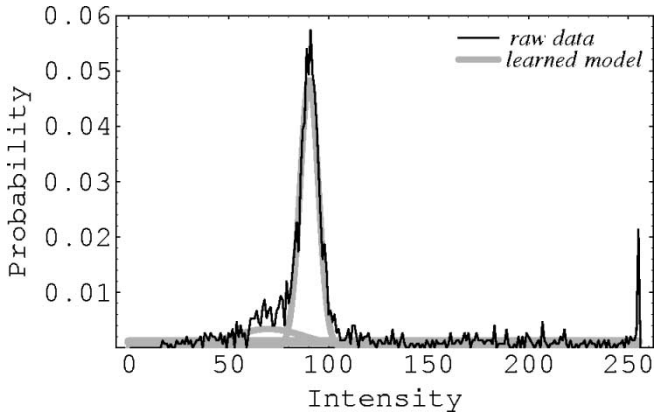

Fig. 8. HMM parameter learning. The estimated observation densities fit the raw data well.

approach to segmentation (state estimation) of real-world traffic-monitoring movies.

A. Parameter Learning and Evaluation

Before applying the models to a test sequence, it is advantageous to observe the model parameters estimated from a learning sequence and compare these parameters with the training data.

1) The HMMs: We use a 30-s highway sequence (see a scene in Fig. 2) to learn the HMM parameters. As an example, the parameters are learned at a region location on the left lane (Fig. 2). The HMM is first considered as an ergodic model, as shown in Fig. 7(a), and only intensity calculated by a 4×4 mean filter is used for observation values. The main parameters estimated are listed as

$$\hat{\mu}_{\mathbf{B}}^{\mathrm{Int}} = 80.14, \quad \hat{\sigma}_{\mathbf{B}}^{\mathrm{Int}} = 10.24$$
$$\hat{\mu}_{\mathbf{S}}^{\mathrm{Int}} = 68.23, \quad \hat{\sigma}_{\mathbf{S}}^{\mathrm{Int}} = 44.98$$
$$\hat{A} = \begin{bmatrix} 0.9860_{(\mathbf{BB})} & 0.0129_{(\mathbf{BS})} & 0.0010_{(\mathbf{BF})} \\ 0.0139_{(\mathbf{SB})} & 0.8843_{(\mathbf{SS})} & 0.1018_{(\mathbf{SF})} \\ 0.0334_{(\mathbf{FB})} & 0.0254_{(\mathbf{FS})} & 0.9413_{(\mathbf{FF})} \end{bmatrix}.$$

First, let us consider the moments for observation densities. To see how the learned densities overlap and how they fit the histogram obtained from the raw data, we weight the estimated density of each state by the steady state probability of transition matrix $\hat{A}$ and show them in Fig. 8. We also give the histogram of the observations of the raw data normalized by T, the length of the observation sequence, in the same figure. From Fig. 8, it is clear that $\left(\hat{\mu}_{\mathbf{B}}^{\mathrm{Int}}, \hat{\sigma}_{\mathbf{B}}^{\mathrm{Int}}\right)$, $\left(\hat{\mu}_{\mathbf{S}}^{\mathrm{Int}}, \hat{\sigma}_{\mathbf{S}}^{\mathrm{Int}}\right)$ fit the raw data well.

Second, let us consider the transition matrix $\hat{A}$. From $\hat{A}$, the typical values relative to time $\hat{\tau}_{\mathbf{B}} \approx 72$, $\hat{\tau}_{\mathbf{S}} \approx 9$, and $\hat{\tau}_{\mathbf{F}} \approx 17$ can be derived [cf., (7)]. $\hat{\tau}_{\mathbf{B}}$ coincides closely with $\tau_{\mathbf{B}} = 75$, the true average time spent in $\mathbf{B}$. On the other hand, although the shadows of vehicles almost do not appear at the region location being studied, $\hat{\tau}_{\mathbf{S}} \approx 9$ is observed. This is because, in the learning sequence, there always is a dark area in front of vehicles that is likely classified as shadow. This can be also explained by the learned parameter $\hat{a}_{\mathbf{BF}} = 0.0010 \approx 0$.

The problem here is that $\hat{\tau}_{\mathbf{F}} \approx 17$ is much smaller than $\tau_{\mathbf{F}} = 31$; namely, the foreground is underestimated. It has been found that windscreens of vehicles are usually darker than other parts of the vehicles. This probably accounts for the misclassification of $\mathbf{F}$ into $\mathbf{S}$ in the learning process. Since a vehicle is learned as separated foreground units, $\hat{\tau}_{\mathbf{F}}$ is underestimated.

One way to alleviate this problem is assuming that the HMM is not fully connected but constrained by $a_{\mathbf{FS}} = 0$, as shown in Fig. 7(b). Learning the state-transition matrix based on this idea results in

$$\hat{A} = \begin{bmatrix} 0.9805_{(\mathbf{BB})} & 0.0157_{(\mathbf{BS})} & 0.0038_{(\mathbf{BF})} \\ 0.0131_{(\mathbf{SB})} & 0.8978_{(\mathbf{SS})} & 0.0892_{(\mathbf{SF})} \\ 0.0480_{(\mathbf{FB})} & 0.0000_{(\mathbf{FS})} & 0.9520_{(\mathbf{FF})} \end{bmatrix}$$

where probability $\hat{a}_{\mathbf{FF}}$ increased and $\hat{\tau}_{\mathbf{F}}$ became larger ($\hat{\tau}_{\mathbf{F}} \approx 21$), as expected. The constraint of $a_{\mathbf{FS}} = 0$ is reasonable and proved to be satisfactorily effective in situations in which there is almost no transition between $\mathbf{F}$ and $\mathbf{S}$, such as the left lane in Fig. 2. However, in situations where the transition from $\mathbf{F}$ to $\mathbf{S}$ actually exists, for example, the shadow of a vehicle that is behind the vehicle, it will lead to delay of the transition from $\mathbf{F}$ to $\mathbf{S}$.

To avoid the underestimation of foreground and distinguish foreground objects from shadows more reliably, we think it is necessary to introduce other observations besides intensity. We employ high-frequency wavelet coefficients for 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 $\mathbf{S}$ and $\mathbf{B}$, but large for $\mathbf{F}$, because the foreground objects are generally more spatially complicated than background and shadow regions. Suppose for a $k \times k$ HMM region in the upper left-hand corner of an image specified as $\mathcal{I} = \{(m,n), m = 0, \ldots, M-1, n = 0, \ldots, N-1\}$, its wavelet coefficients are $\{\mathcal{W}_{m,n} = (m,n) \in \mathcal{I}\}$. The second observation for the HMM region is calculated as the variance of the following three wavelet coefficient sets:

$$w_{i,j}^{LH} = \left\{ \mathcal{W}_{m,n}, m = \frac{M}{2}, \ldots, \frac{M}{2} + \frac{k}{2} - 1 \right.$$
$$\left. n = 0, \ldots, \frac{k}{2} - 1 \right\}$$
$$w_{i,j}^{HL} = \left\{ \mathcal{W}_{m,n}, m = 0, \ldots, \frac{k}{2} - 1 \right.$$
$$\left. n = \frac{N}{2}, \ldots, \frac{N}{2} + \frac{k}{2} - 1 \right\}$$

and

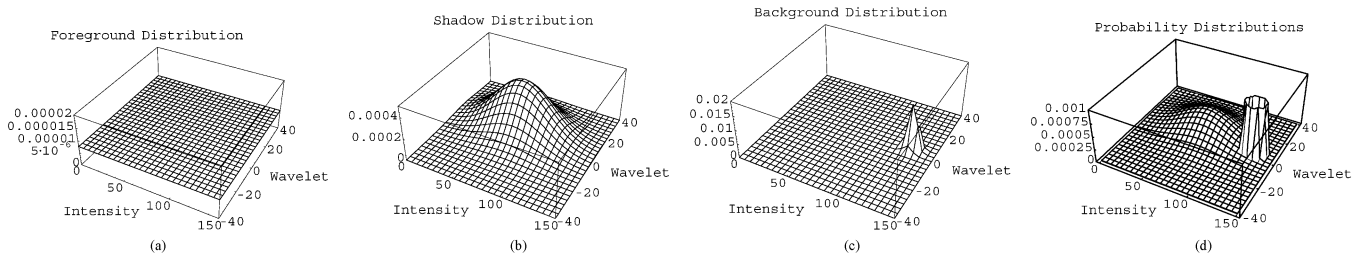

Fig. 9. 2-D observation densities. The observation densities are learned at one region location for (a) background, (b) shadow, and (c) foreground and shown together in (d). In (a)–(d), the first observation, intensity, uses the output of a mean filter and the second observation, wavelet, is calculated as the variance of the wavelet coefficients in high-frequency bands.

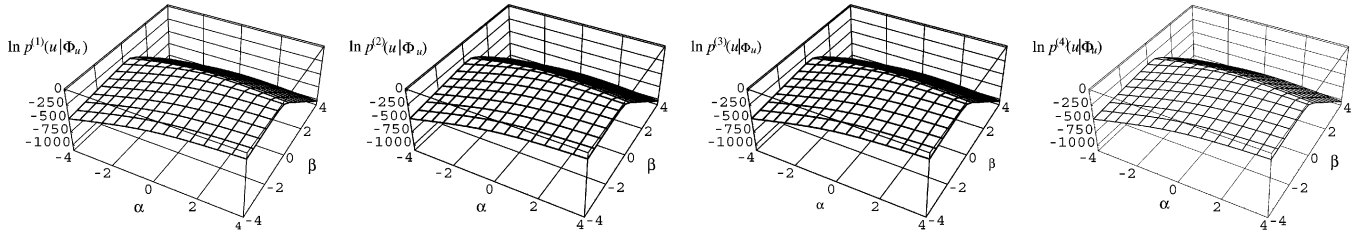

Fig. 10. MRF parameter learning. Likelihood distributions for four codings are calculated from the training data.

$$w_{i,j}^{HH} = \left\{ \mathcal{W}_{m,n}, m = \frac{M}{2}, \ldots, \frac{M}{2} + \frac{k}{2} - 1 \right. \\ \left. n = \frac{N}{2}, \ldots, \frac{N}{2} + \frac{k}{2} - 1 \right\}$$

i.e.,

$$\text{Var} = \sum_{i=0}^{k-1} \sum_{j=0}^{k-1} \Big(\left(m_{i,j} - w_{i,j}^{LH} \right)^2 + \left(m_{i,j} - w_{i,j}^{HL} \right)^2 + \left(m_{i,j} - w_{i,j}^{HH} \right)^2 \Big) \quad (28)$$

where

$$m_{i,j} = \frac{1}{3} \left(w_{i,j}^{LH} + w_{i,j}^{HL} + w_{i,j}^{HH} \right). \quad (29)$$

For other shifted HMM regions, their wavelet coefficient sets for calculating the variance are shifted correspondingly. In our current implementation, Daubechies wavelet transformation ($N = 2$) is adopted [32]. The two observations are treated as a single 2-D feature vector, as shown in Fig. 9.

Parameter learning using the Baum–Welsh reestimation formulas converges quickly. Experiments show that a fixed number of reestimations equal to ten gives satisfactory results. The model parameters used in this section are all obtained by ten times of reestimation.

2) MRF: A good result of state estimation with the optimization criterion of (11) is used for learning α and β, the parameters involved in the prior of the MRF model. We plot the log-coding likelihood given by (25) in Fig. 10. Fortunately, the function is convex. The parameters are estimated by solving (25) in a numerical procedure. As shown in Table I, a different data sample set $\boldsymbol{\mathcal{S}}^{(k)}$ gives rise to a different estimate $(\hat{\alpha}^{(k)}, \hat{\beta}^{(k)})$, but our experiments show that the parameter variation does not lead to any significant difference in the performance of the model.

TABLE I
MRF PARAMETER ESTIMATION

Coding	$\hat{\alpha}^{(k)}$	$\hat{\beta}^{(k)}$
$\boldsymbol{\mathcal{S}}^{(1)}$	6.283	−1.578
$\boldsymbol{\mathcal{S}}^{(2)}$	8.397	−1.931
$\boldsymbol{\mathcal{S}}^{(3)}$	4.475	−1.393
$\boldsymbol{\mathcal{S}}^{(4)}$	9.475	−2.452
Average	7.158	−1.838

TABLE II
STATE–ESTIMATION RESULTS

Learning Sequence	Missing	Partially Missing	Completely Detected	Completely Detected Ratio
1	0	4	84	95.5
2	0	2	86	97.7
3	0	0	88	100
4	0	2	86	97.7
5	0	5	83	94.3

B. Results on State Estimation

Our method has been tested on a number of real-world traffic-monitoring movies. Table II shows some results obtained by using five sets of HMM parameters, learned from five different 30-s sequences, and applying the model with these parameters to a 2.5-min highway sequence including 88 vehicles. Between the test and the learning sequences, light conditions do not change very much; however, the typical times spent in $\mathbf{B}$, $\mathbf{F}$, and $\mathbf{S}$ differ from each other. Without a loss of generality, parameter learning and state estimation are performed with respect to an interested area that covers two lanes: in the left lane, there is almost no transition between $\mathbf{F}$ and $\mathbf{S}$, while in the right lane all categories $\mathbf{B}$, $\mathbf{F}$, and $\mathbf{S}$ can be observed (Fig. 2). The area consists of 43×28 HMM regions and each region has a 4×4 pixel size. The HMM is the ergodic model and the resolution of images is 768×576 pixels.

From Table II, we can see that there are no vehicles missing in the tracking process, although some (five vehicles in total, for different sets of parameters) are "partially missing," which

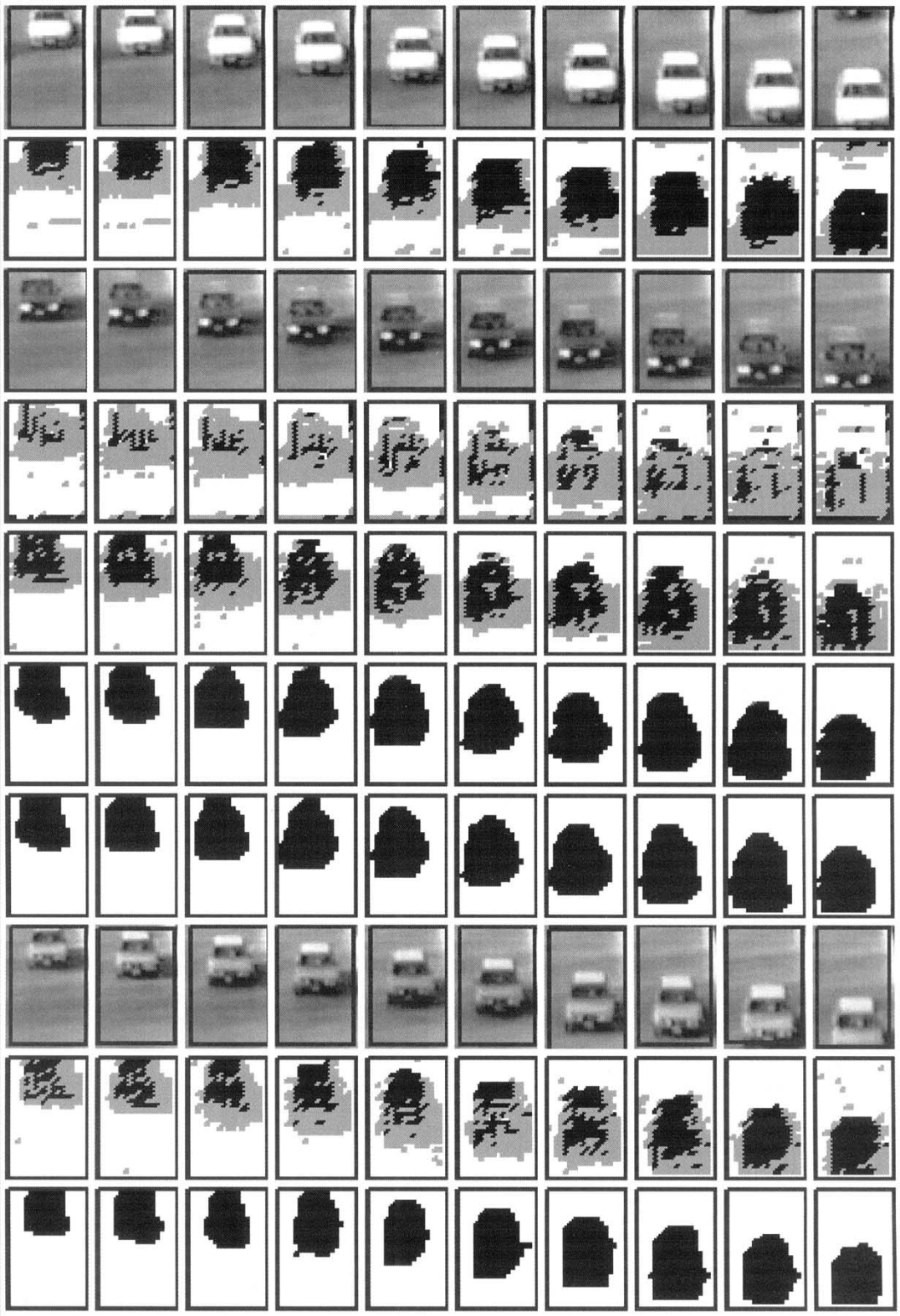

Fig. 11. State-estimation results for light cars (rows 1–2), dark cars (rows 3–7), and gray cars (rows 8–10). Foreground: black, shadow: gray and (enlarged) background: white.

means that the vehicles only partially detected or detected as separate regions. Let us look at these results in depth. Fig. 11 summarizes some of the experimental results via images. To make the following explanation straightforward, we divide roughly the situations regarding these results into three types according to the contrast of the foreground objects with the background and shadow regions.

First, we consider "light cars," the situation in which foreground objects have higher intensities than those of the background. The first row in Fig. 11 gives ten such successive images

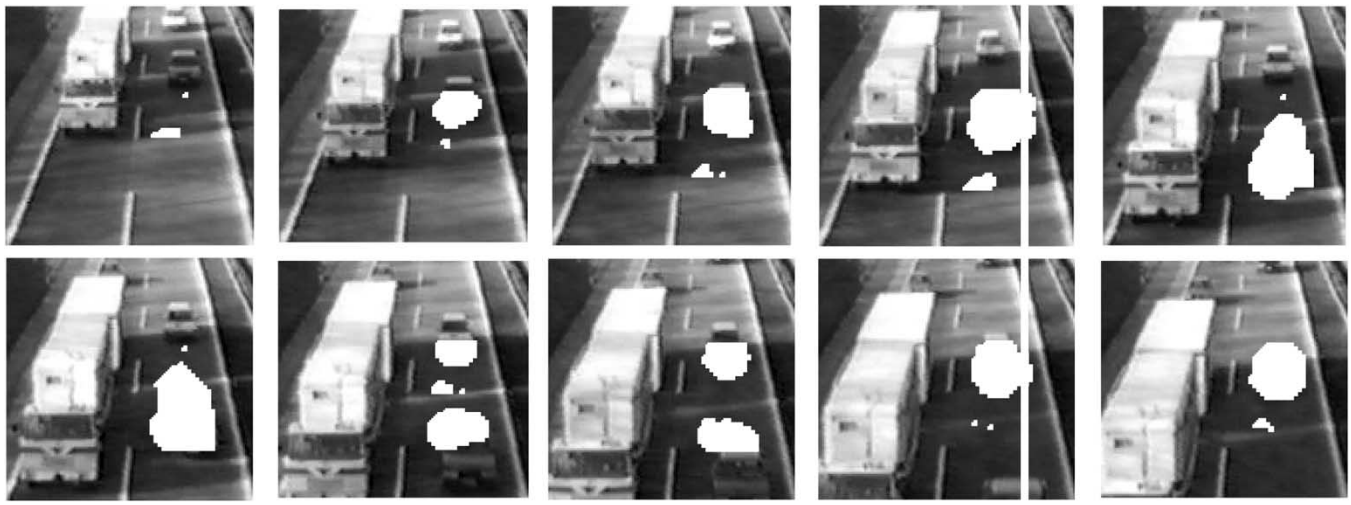

Fig. 12. Detecting and tracking the vehicles running in the shadow cast by a truck. Foreground: white.

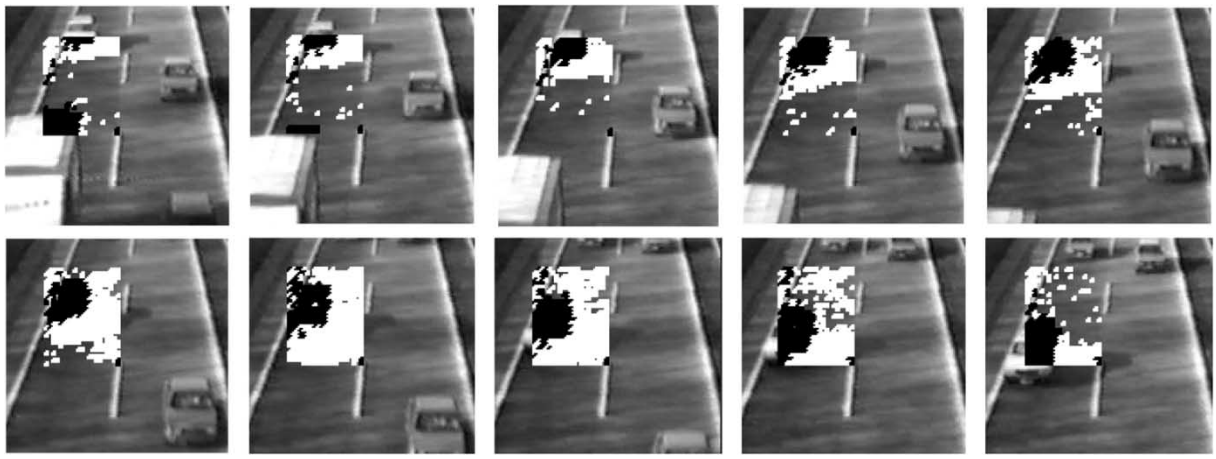

Fig. 13. Reaction to camera's autogain control. The target vehicle follows a light-colored truck. At the moment the truck passes through, the background suddenly becomes dark and, thus, is classified as shadow, but the tracker still detects and keeps holding on the target vehicle. To show the shadow, we use (11) for state estimation. Foreground: black and shadow: white.

at one-frame intervals. According to the observation distribution in Fig. 9, we can predict that the light cars are at low risk of being confused with other categories, either $\mathbf{B}$ or $\mathbf{S}$, because in this area $\mathbf{B}$ and $\mathbf{S}$ provide negligible probabilities. The state-estimation results shown in the second row in Fig. 11 confirm this prediction: the foreground objects are completely distinguished from other categories. In effect, to procure these results, we have not employed the MRF model yet, but only the forward probabilities, i.e., the optimization criterion of (11). We conclude that the light cars stand out from other categories *per se*.

Second, we consider "dark cars," the situation in which foreground objects have intensities similar to those of shadow (see examples in the third row). The partially missing vehicles in Table II are all dark cars, which are prone to being confused with shadow, because when the gray values of the foreground region of a dark car also fall into the support of the shadow distribution, the probability that this region is classified as $\mathbf{S}$ is obviously much larger than the probability that this region is classified as $\mathbf{F}$. Introducing the second observation, the variance of wavelet coefficients in high-frequency bands, contributes to the alleviation of this problem. This is because the overlapping among the observation densities of different categories is reduced. To verify this idea, we test the sequence by using intensity alone and by using wavelet coefficients in conjunction with intensity as the observations and give the results in the fourth and fifth rows, respectively. In the fourth row, only the light portions of a dark car, such as the roof and lamps, are classified as foreground. On the other hand, in the fifth row a larger percentage of the car is detected as expected. The sixth row also provides the results for the same images, but using the state estimation of the MRF model rather than simple forward probabilities, as in the fourth and fifth rows. Because we have taken the spatial-dependent information among neighboring regions into account, the whole region of the car is segmented from the enlarged background $\mathbf{B}^{+}$. For comparison, the 20th iteration of the simulated annealing algorithm is taken for the results in the sixth row and the 100th iteration for the results in the seventh row. No remarkable difference between them can be observed, meaning that the stochastic relaxation process converges quickly.

Finally, we consider "gray cars," the situation where foreground objects have intensities similar to those of background. Some images of gray cars are shown in the eighth row. The

state-estimation results based on the forward probabilities of individual HMM regions are shown in the ninth row and those with the context dependence among HMM regions modeled by the MRF are shown in the last row. Due to the overlapping between the distributions of gray cars and the background, a similar misclassification problem also occurs with gray cars. However, since the variance of the background is usually much smaller than that of shadow, the risk of a gray car being confused with the background is lower.

Since our method is developed to deal with shadows of moving objects, we give an example in Fig. 12 to show how the model detects the vehicles running in the shadow of other vehicles. Also, in Fig. 13 we give an example to show how the model keeps holding on the target vehicle when the camera's autogain control functions.

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). Several movie clips, including the above results, are provided on a web page at URL http://www.watanabe.nuie.nagoya-u.ac.jp/member/jien/

VI. Conclusion

In the first half of this paper, we proposed an HMM-based segmentation method that classifies each small region of images into three categories: vehicles, shadows of vehicles, and background from a traffic-monitoring movie. The temporal continuity of each category for one small region location is modeled by an HMM along the time axis and the state estimation at each time step is performed by recursively computing forward probabilities, but independently of neighboring regions. In order to incorporate spatial-dependent information among neighboring regions into the model, in the second half of this paper we described a new state-estimation method that models the output from the HMMs as an MRF and employs MAP criterion in conjunction with the MRF to find an optimal configuration for the image, through a stochastic relaxation process. By integrating the MAP–MRF schema with the HMM-based segmentation method, we achieved context-dependent classification in both a temporal and spatial sense.

The proposed method provides a way to successfully model the shadows of moving objects as well as background and foreground regions and, thus, contributes to enhancing the robustness of the tracking process against the moving shadows. The experimental results show that, using this method, foreground (vehicles) and nonforeground regions can be discriminated with high accuracy. In particular, this method proves to be effective at reducing the misestimates for dark cars or dark regions inside a vehicle, such as windscreens, by taking the spatial-dependent information into account. Another advantage is that, unlike many other background models, the method does not need any specific data for training. All the HMM parameters are estimated by an EM algorithm from ordinary image sequences.

When illumination conditions or traffic density change a lot, the HMM parameters need to be updated to fit into the new situation. Since the illumination conditions and traffic density vary throughout the day, an important problem for future work is deriving a criterion when the parameters of the model need to be updated through a relearning process.

Finally, it should be noted that there are no constraints on the shape or movement of foreground objects imposed on the proposed model. Although this work has focused on the segmentation of traffic-monitoring movies, it also suggests application to other segmentation problems of image sequences. However, careful investigation has to be carried out to find a suitable model and suitable observations for a particular task.

References

[1] E. Dickmanns, "Expectation-based dynamic scene understanding," in *Active Vision*, A. Blake and A. Yuille, Eds. Cambridge, MA: MIT Press, 1992, pp. 303–336.

[2] J. D. Crisman, "Color region tracking for vehicle guidance," in *Active Vision*, A. Blake and A. Yuille, Eds. Cambridge, MA: MIT Press, 1992, ch. 7.

[3] S. Smith, "Asset-2: Real-time motion segmentation and shape tracking," in *Proc. 5th Int. Conf. Computer Vision (ICCV'95)*, 1995, pp. 237–244.

[4] N. Ferrier, S. Rowe, and A. Blake, "Real-time traffic monitoring," in *Proc. 2nd IEEE Workshop Applications of Computer Vision*, 1994, pp. 81–88.

[5] D. Koller, J. Weber, and J. Malik, "Robust multiple car tracking with occlusion reasoning," in *Proc. 3rd Eur. Conf. Computer Vision (ECCV'94)*, 1994, pp. 189–196.

[6] S. Kamijyo, Y. Matsushita, and M. Sakauchi, "Traffic monitoring and accident detection at intersections," *IEEE Trans. Intell. Transport. Syst.*, vol. 1, pp. 108–119, June 2000.

[7] G. D. Sullivan and A. Blake, "Visual interpretation of known objects in constrained scenes," *Phil. Trans. R. Soc. Lond. B*, vol. 337, pp. 361–370, 1992.

[8] P. Fieguth and S. Wesolkowski, "Highlight and shading invariant color image segmentation," in *Proc. 3rd Int. Workshop Energy Minimization Methods in Computer Vision and Pattern Recognition (EMMCVPR'0l)*, Sept. 2001, pp. 314–327.

[9] T. Gevers and A. Smeulders, "Pictoseek: Combining color and shape invariant features for image retrieval," *IEEE Trans. Image Processing*, vol. 9, pp. 102–119, Jan. 2000.

[10] R. Alferez and Y. Wang, "Geometric and illumination invariants for object recognition," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 21, pp. 505–536, June 1999.

[11] S. Kamijo, T. Nishida, and M. Sakauchi, "Occlusion robust and illumination invariant vehicle tracking for acquiring detailed statistics from traffic images," *IEICE Trans. Inform. Syst.*, vol. E85-D, no. 11, pp. 1753–1766, 2002.

[12] D. Forsyth, "A novel algorithm for color constancy," *Int. J. Comput. Vis.*, vol. 5, pp. 5–36, 1990.

[13] J. Rittscher, "First year report," Dept. Eng. Sci., Univ. Oxford, Oxford, U.K., 1999.

[14] S. Rowe and A. Blake, "Statistical mosaics for tracking," *Image Vis. Comput.*, vol. 14, pp. 549–564, 1996.

[15] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," in *Proc. 1st Int. Conf. Computer Vision (ICCV'87)*, 1987, pp. 259–268.

[16] A. Blake and M. lsard, *Active Contours*. London, U.K.: Springer-Verlag, 1998.

[17] A. Baumberg and D. Hogg, "Learning flexible models from image sequences," in *Proc. 3rd Eur. Conf. Computer Vision (ECCV'94)*, vol. 1, 1994, pp. 299–308.

[18] N. Mine, Y. Yagi, and M. Yachida, "Detection of change region by integrating subtracted image and edge boundary image," *Trans. Inst. Electron., Inform., Commun. Eng. (IEICE)*, vol. J77-DX-ll, no. 3, pp. 631–634, 1994.

[19] I. Haritaoglu, D. Harwood, and L. Davis, "w^4s: A real-time system for detecting and tracking people in 2.5d," in *Proc. 5th Eur. Conf. Computer Vision (ECCV'98)*, vol. 1, Freiburg, Germany, 1998, pp. 877–892.

[20] K. Toyama, J. Krumm, B. Brumitt, and B. Meyers, "Wallflower: Principles and practice of background maintenance," in *Proc. 7th Int. Conf. Computer Vision (ICCV'99)*, 1999, pp. 255–261.

[21] S. lntille, J. Davis, and A. Bobick, "Real-time closed-world tracking," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 1997, pp. 93–101.

[22] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland, "Pfinder: Real-time tracking of the human body," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 19, pp. 780–785, July 1997.

[23] C. Bregler, "Learning and recognition human dynamics in video sequences," in *Proc. IEEE Conf. Computer Vision Pattern Recognition*, San Juan, Puerto Rico, June 1997, pp. 568–574.
[24] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," in *Proc. 7th Int. Conf. Computer Vision (ICCV'99)*, vol. 77, Feb. 1989, pp. 257–286.
[25] J. Kato, T. Watanabe, S. Joga, J. Rittscher, and A. Blake, "An HMM-based segmentation method for tracking monitoring movies," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 24, pp. 1291–1296, Sept. 2002.
[26] J. Rittscher, J. Kato, S. Joga, and A. Blake, "A probabilistic background model for tracking," in *Proc. 6th Eur. Conf. Computer Vision (ECCV'00)*, Dublin, U.K., June 2000, pp. 336–350.
[27] G. Winkler, "Image analysis," in *Random Fields and Markov Chain Monte Carlo Methods: A Mathematical Introduction*, 2nd ed, ser. Applicat. Math. 27. Berlin, Germany: Springer-Verlag, 2003.
[28] A. R. Webb, *Statistical Pattern Recognition*, 2nd, Ed. London, U.K.: Wiley, 2002.
[29] J. Besag, "Spatial interaction and the statistical analysis of lattice systems," *J. R. Statist. Soc.*, ser. B, vol. 36, pp. 192–236, 1974.
[30] S. Z. Li, *Markov Random Field Modeling in Image Analysis*, ser. Comput. Sci. Workbench. Tokyo, Japan: Springer-Verlag, 2001.
[31] S. Geman and D. Geman, "Stochastic relaxation, gibbs distributions, and the bayesian restoration of image," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-6, pp. 721–741, June 1984.
[32] I. Daubechies, *Ten Lectures on Wavelets*. Philadelphia, PA: SIAM, 1992.

Jien Kato (M'02) received the M.E. and Ph.D. degrees in information engineering from Nagoya University, Nagoya, Japan, in 1990 and 1993, respectively.

In 1999, after research positions with Toyama University, Toyama, Japan, she joined the Robotics Research Group, Department of Engineering Science, University of Oxford, Oxford, U.K. She currently is an Associate Professor with the Department of Systems and Social Informatics, Nagoya University. Her research interests include statistical computer vision, contextual models of object recognition, image processing, and applications.

Dr. Kato is a Member of the Information Processing Society of Japan, the Institute of Electronics, Information, and Communication Engineers of Japan, and the IEEE Computer Society.

Toyohide Watanabe (A'88–M'02) received the B.S., M.S., and Ph.D. degrees from Kyoto University, Kyoto, Japan, in 1972, 1974, and 1983, respectively.

In 1987, he was an Associate Professor in Department of Information Engineering, Nagoya University, Nagoya, Japan, where he currently is a Professor in the Department of Information Engineering. His research interests include knowledge/data engineering, computer-supported collaborative learning, parallel and distributed process interaction, document understanding, and drawing interpretation.

Dr. Watanabe is a Member of the Information Processing Society of Japan, the Institute of Electronics, Information and Communication Engineers of Japan, the Japan Society for Software Science, the Japan Society of Artificial Intelligence, the Japanese Society for Information and Systems in Education, the ACM, AAAI, AACE, and the IEEE Computer Society.

Sébastien Joga received the degree from the Ecole Polytechnique, Palaiseau, France, in 1999 and the degree from the Ecole Nationale Supérieure des Télécommunications, Paris, France, in 2001.

He worked on computer-vision systems for car tracking in the Department of Engineering Science, University of Oxford, Oxford, U.K., in 1999. He currently is with France Telecom R&D. His research interests include the performance of packet services over UMTS and the performance of radio resource-management algorithms.

Ying Liu received the M.E. degree in electronics and computer science and the Ph.D. degree in systems science and engineering from Toyama University, Toyama, Japan, in 2001 and 2004, respectively.

Her research interests include the analysis of acoustic field by finite element method and visual tracking.

Hiroyuki Hase (A'97) received the B.E. degree in electrical engineering from Toyama University Toyama, Japan, in 1971 and the Ph.D. degree from Tohoku University Sendai, Japan, in 1989.

From 1975 to 1989, he was an Assistant in the Department of Electronics and Information Engineering, from 1989 to 1993, he was a Lecturer in the same department, and from 1993 to 2003, he was an Associate Professor in the Department of Intellectual Information Systems Engineering, all with Toyama University. Since April 2003, he has been a Professor in the Department of Information Science, University of Fukui, Fukui, Japan. His current research interests include document image analysis, character recognition, and facial expression analysis.

Dr. Hase is a Member of the Information Processing Society of Japan, the Institute of Electronics, Information and Communication Engineers, the Institute of Image Information and Television Engineers, the Institute of Image Electronics Engineers of Japan, and the IEEE Computer Society.