コンピュータビジョンと
イメージメディア 130－7
（２００１．１１．８）

観測値としてウェーブレット係数を用いたHMMに基づく交通監視映像における移動物体分離手法

加藤ジェーン[1]　渡邉豊英[1]　長谷博之[2]

[1] 名古屋大学大学院工学研究科

[2] 富山大学工学部

【概要】本論文では、交通監視映像における各小領域が自動車、自動車の影、背景のいずれに属するかを精度よく分離するための、 HMMをベースに改良した分離手法を提案する。ある位置の一つの小領域での時間ステップごとの観測値が、一つのHMMとしてモデル化される。モデル化の時点では、近傍領域の状態に関連なく独立にモデルパラメータが学習がされる。これに対し、分離を実行する状態見積もりアルゴリズムは、それぞれの小領域に対して近傍領域に依存した分類を行うことができる。このアルゴリズムは、過去の観測値にのみ基づいているため、状態見積もりはリアルタイムに実行可能である。候補となる状態間の曖昧さは、観測値としての輝度値以外に、高周波領域のウェーブレット係数を第２観測値として導入することによって軽減される。未知のHMMパラメータは、Baum-Welshアルゴリズムに基づき、通常のビデオシーケンスからすべて自動的に学習される。実際に高速道路を撮影した映像を用いた実験では、本手法を用いることにより精度よく自動車を背景や影から分離できることが示された。

An HMM-based Segmentation Method with Observation of Wavelet Coefficients for Traffic Monitoring Movies

Jien Kato[1], Toyohide Watanabe[1] and Hiroyuki Hase[2]

[1]Dept.of Information Engineering, Nagoya University

[2]Faculty of Engineering, Toyama University

Abstract *This paper proposes an improved HMM-based segmentation method which is designed to classify each small region of images into vehicles, the shadows of vehicles and the background from a traffic monitoring movie. The observations over time for one small region location are modeled as a single HMM, independent of the neighboring regions. A state estimation algorithm is used to perform context-dependent classification of individual HMM regions. Because this algorithm is only based on the past observations, state estimation is possible to be performed in real time. The ambiguity among different categories is reduced by introducing high frequency wavelet coefficients as the second observation in addition to the intensities. All the unknown HMM parameters can be fully automatically learned from an ordinary video sequence based on Baum-Welsh algorithm. Results on real-world motorway sequences show that it is possible to accurately distinguish vehicles by this method.*

I Introduction

For enhancing the robustness to different lighting conditions of car tracking, we have proposed a segmentation method based on hidden Markov models that classifies each small region of a traffic monitoring movie into three different categories: foreground (vehicles), background and shadow[1]. Two important problems with respect to this method are choosing suitable observation that aims at describing the statistical properties of individual regions and performing

context-dependent classification that incorporates spatial information among the regions. This paper focuses on the improvement for the proposed method from this viewpoint.

Choosing observations is a critical issue in classification problems because observations often set the limits of classification performance. Since the distributions of intensity for different categories usually have a large overlap, it is impossible to construct a successful model which is purely based on intensity values. To classify each small region of an image sequence into the foreground, background and shadow, both a mean filter and a Sobel filter, defined according to the pixel intensities within a region, have been employed. Introducing the Sobel filter is based on the idea that the Sobel filter describes the space homogeneity property which should be different between the foreground and non-foreground categories. The Sobel filter cooperated with the mean filter has greatly improved the classification results. But, the improvement is not enough when the intensity differences between the foreground and non-foreground categories are small. This becomes the motivation to adopt high frequency wavelet coefficients as the new observation in this paper.

We consider the second problem. In our approach, observations over time for one specific region location are modeled as a single HMM, independent of the neighboring regions. As most block-based image classification algorithms such as BVQ[2], this approach leads to an issue of choosing region sizes. A too large region size obviously entails crude classification, while choosing a small region size means that only very local properties belonging to the small region are examined. The penalty then comes from losing information about surrounding regions. A well-known approach in signal processing to attack this type of problems is to incorporate context information[3]. For example, in our particular problem a foreground state is highly unlikely to be situated in isolation, surrounded by background regions. According to this important consideration, a new criterion for selection of optimal states by not making decisions independently for each region but performing the context-dependent classification of individual regions have been modeled in this paper.

In Section II, we give briefly a mathematical formulation of the model and the iterative re-estimation formulae for the model parameters according to Baum-Welsh algorithm. Section III addresses the new observation of high frequency wavelet coefficients. Section IV describes state estimation that takes the context-dependence among individual regions into account. In Section V, experimental results on real-world image sequences are presented. Finally, we draw conclusions and point out future directions in Section VI.

II The Hidden Markov Model

The theory of hidden Markov models was proposed in the 1960s by Baum et al.[5]-[8]. HMM's have earned their popularity in large part from successful application to speech recognition [9], [10], [11], [12]. Under an HMM is a basic Markov chain. At any discrete unit of time, the system is assumed to exist in one of a finite set of states. Transitions between the states take place according to a probability, depending only on the state of the system at the unit of time immediately preceding (one-step Markovian). In an HMM, at each unit of time, a single observation is generated from the current state according to a probability distribution, depending only on this state. HMM's owe both their name and modeling power to the fact that these states represent abstract quantities that are themselves never observed. They correspond to the clusters of contexts having similar probability distributions of the observation.

We apply the HMM to the problem of segmenting each field of a traffic monitoring movie into

three different categories: foreground (**F**), background (**B**) and shadow (**S**). To make our method robust, especially to single out foreground objects reliably, we use both grey-value intensities and high frequency wavelet coefficients as the observations. At any unit of time, the two values are observed and depend on an underlying unobservable process which explains the transitions among hidden categories **B**, **S**, and **F**. The distributions of both shadow and background are approximated by 2D Gaussian-mixture densities, and that of the foreground is modeled as a uniform probability density. The state transition probabilities impose the temporal continuity, which means a region belongs to a certain category for a period of time, on each category. All the unknown HMM parameters are estimated by using an EM algorithm[13].

We provide a mathematical formulation of the model and the iterative re-estimation formulae for the model parameters. Let $S = \{S_b,\ S_s,\ S_f\}$ be the states corresponding to the three categories. The parameters of the HMM, notated as $\lambda = \{A, B, \pi\}$, are specified as follows:

- Initial state distribution: $\pi = \{\ \pi_b,\quad \pi_s,\quad \pi_f\ \}$, $\pi_i = \Pr(S_i \text{ at } t = 1)$.
- State transition matrix: $A = \begin{pmatrix} a_{bb} & a_{bs} & a_{bf} \\ a_{sb} & a_{ss} & a_{sf} \\ a_{fb} & a_{fs} & a_{ff} \end{pmatrix}$, $a_{ij} = \Pr(S_j \text{ at } t+1 | S_i \text{ at } t)$.
- 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 observation probabilities of the background and shadow are characterized by only mean vector (μ_i) and covariance matrix ($\mathbf{\Sigma}_i$) instead of all the probabilities for different observation values, i.e.

$$b_i(v) = \frac{1}{\sqrt{(2\pi)^2 \det(\mathbf{\Sigma}_i)}} e^{-\frac{1}{2}(v-\mu_i)^t \mathbf{\Sigma}_i^{-1}(v-\mu_i)}, \quad i \in \{b, s\}. \tag{1}$$

A special case of the EM algorithm, Baum-Welsh algorithm[14], that performs maximum likelihood estimation is applied for learning the unknown model parameters. This algorithm produces a sequence of estimates for λ, given a set of observed data x such that each estimate has a greater value for $L(x, \lambda) = \log[p(x|\lambda)]$[13]. The re-estimation formulae for π, A and B are defined as

$$\overline{\pi}_i = \gamma_1(i), \tag{2}$$

$$\overline{\mu}_i = \frac{\sum_{t=1}^{T} u_t \gamma_t(i)}{\sum_{t=1}^{T} \gamma_t(i)}, \tag{3}$$

$$\overline{\mathbf{\Sigma}}_i = \frac{\sum_{t=1}^{T} \gamma_t(i)(u_t - \overline{\mu}_i)(u_t - \overline{\mu}_i)^t}{\sum_{t=1}^{T} \gamma_t(i)}, \tag{4}$$

$$\overline{a}_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)}, \tag{5}$$

where $\gamma_t(i) = \Pr(S_i \text{ at } t | u, \lambda)$ and $\xi_t(i,j) = \Pr(S_i \text{ at } t, S_j \text{ at } t+1 | u, \lambda)$ ($U = \{u_1, ..., u_T\}$ is a sequence of observation symbols) are auxiliary probabilities that can be efficiently calculated by the so-called forward-backward algorithm[16].

III Observations

We partition each field of an image sequence into non-overlap blocks with equal size ($k \times k$ pixels), called HMM region. In the learning phase, the model parameters for each HMM region are estimated based on the observations extracted from a learning sequence. On the other hand, in the testing phase, the optimal states associated with the past observations generated from a testing sequence are found over time for each HMM region, given the trained model.

We use the outputs of a $k \times k$ mean filter as the intensity observation instead of using the gray level directly to reduce the noise. Experiments on real-world motorway sequences reveal that since all the three intensity distributions (for **F**, **B**, and **S**) have a large amount of overlap, the HMM with only the intensity observations does not allow to classify vehicles in a robust way. To distinguish the foreground objects reliably, in addition to the intensities it is necessary to incorporate another type of observations so that the overlap between the distributions of different categories can be reduced.

Wavelet transformation is a tool that cuts up data into different frequency components, and then studies each component with a resolution matched to its frequency[4]. In a traffic monitoring sequence, vehicles are the objects of interest. They are usually sharply focused but background objects and shadows are not so. Sharply focused vehicles have more details within the objects than the background and shadows[15]. The details in the foreground objects result in larger high frequency energy in an image. If we measure the high frequency energy by the wavelet transformation, the vehicle regions should have more high value coefficients in high frequency bands than non-foreground regions. A Sobel filter is good at edge detection. The motivation for us to choose wavelet coefficients is to make our method more robust by analyzing the details of the regions rather than by only analyzing the edges.

Suppose for a $k \times k$ (pixel) region in the left up corner of a field specified as $\mathcal{F} = \{(m,n), m = 0, ..., M-1, n = 0, ..., N-1\}$, its wavelet coefficients are $\{\mathcal{W}_{m,n} = (m,n) \in \mathcal{F}\}$. The high frequency energy is calculated as the variance of the wavelet coefficients in LH, HL and HH bands, i.e., the variance of the three sets of

$$\{\mathcal{W}_{m,n}, m = M/2, ..., M/2 + k/2 - 1, n = 0, ..., k/2 - 1\},$$

$$\{\mathcal{W}_{m,n}, m = 0, ..., k/2 - 1, n = N/2, ..., N/2 + k/2 - 1\},$$

$$\text{and} \quad \{\mathcal{W}_{m,n}, m = M/2, ..., M/2 + k/2 - 1, n = N/2, ..., N/2 + k/2 - 1\}.$$

For other shifted HMM regions, their wavelet coefficient blocks for calculating the variance are shifted correspondingly. In our current implementation, Daubechies wavelet transformation ($N = 2$) is adopted[4].

IV State Estimation

Although several criteria for making choice of an "optimal" state sequence associated with the given observation sequence are possible, in view of tracking the basic requirement for the state estimation is working in real time. Namely, we cannot adopt a criterion that uses the whole sequence of observations such as Viterbi algorithm[17],[18]. One solution is to maximize the joint probability of the state at time t and the past observation $\{u_1, ..., u_t\}$ given the model, i.e.

$$\text{argmax}\{\alpha_t(k)\} = \text{argmax}\{\Pr(u_1, \ldots, u_t, S_k \text{ at } t|\lambda)\}. \tag{6}$$

However, a drawback to this method is that it does not incorporate the context-dependence among HMM regions. To take the spatial information among HMM regions into account, we estimate a state with

$$\text{argmax}\{\Pr(u_1,\ldots,u_t, S_k \text{ at } t|\lambda)\ \Pr(Q_{i,j}|\mathcal{Q}_{\mathcal{N}_{i,j}})\}, \tag{7}$$

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), given the probability of state set $\mathcal{Q}_{\mathcal{N}_{i,j}}$ at neighborhood $\mathcal{N}_{i,j}$ of (i,j). We define $\Pr(Q_{i,j}|\mathcal{Q}_{\mathcal{N}_{i,j}})$ as

$$\Pr(Q_{i,j}|\mathcal{Q}_{\mathcal{N}_{i,j}}) = \frac{1}{D}\exp(\kappa\vartheta(Q_{i,j})). \tag{8}$$

In Eq.(8), D and κ express a normalizer and a parameter that expresses the strongness of the context-dependence among HMM regions, respectively. The function $\vartheta(Q_{i,j})$ is simply selected as

$$\vartheta(Q_{i,j}) = \sum_{(s,r)\in\mathcal{N}^8_{i,j}} \frac{1}{16} I(i,j,s,r) + \sum_{(s',r')\in\mathcal{N}^{16}_{i,j}} \frac{1}{32} I(i,j,s',r'), \tag{9}$$

$$I(i,j,s,r) = \begin{cases} 1 & Q_{i,j} = Q_{s,r} \\ 0 & Q_{i,j} \neq Q_{s,r} \end{cases} \tag{10}$$

where $\mathcal{N}^8_{i,j}$ and $\mathcal{N}^{16}_{i,j}$ are the 8-neighbors of region (i,j) with distances 1 and 2, respectively. Notice that Eq.(6) and Eq.(7) can be solved by the forward procedure alone[16]. Since $\alpha_t(k)$ is defined recursively, it is possible to perform the state estimation with Eq.(6) and Eq.(7) in real time.

V Experimental Results

Several 30 second sequences are used for experiments. Although the traffic density and lighting condition of these sequences do not change too much, the typical time spent in **B**, **F** and **S** related with a test sequence might be very different from that of a learning sequence. The experimental results we are about to discuss are obtained with respect to a learned area located on the right lane where the shadow certainly exists. This area is composed of 18×28 HMM regions, each one has 4×4 pixel size.

Some results, all use the constrained model($a_{fs} = 0$)[1], are given in Fig.1. To make the explanation straightforward, we roughly divide the vehicles into light, dark and gray ones. First we consider light cars. The first row of Fig.1 shows six successive images of a light car at three-field intervals. The corresponding classification results, using two observations and adopting Eq.(6) as the optimality criterion, are given in the second row. It turns out that even if the context-dependence between HMM regions are not taken into account, the light car has completely distinguished from other categories. Namely, light cars stand out distinctly among background objects and shadows.

By "dark cars", we mean those whose intensity differences with the shadow are very small or the intensity distributions of them overlap each other. Dark cars are particularly noticeable since they are easily confused with the shadow. Actually, the HMM with only the intensity observation also allows to classify light cars in a robust way but not robust for dark cars. Because the distributions of different categories overlap and moreover the probability of the foreground is very low (1/256), when the gray-value of an HMM region that belongs to **F** (a dark car) also falls in the support of the shadow distribution at the same time, it is more likely classified as **S** than **F**. Introducing the second observation, the variance of wavelet coefficients

in high frequency bands, contributes to the robustness of foreground object recognition. As described before, the introduction of this observation is based on the idea that the variance of wavelet coefficients should be small for **S** and **B** but large for **F** because of the details inside a car. With a 2D feature vector, the area proportion where the densities of different categories overlap is less than the same proportion for 1D feature vector. The Bayes risk is thus reduced.

To confirm the effectiveness, we test a sequence at the same area using intensity alone and using wavelet coefficients together with intensity as the observations. The results with 1D and 2D features for a dark car (see the 3rd row) are shown in the 4th and 5th rows, respectively. A larger percentage of the dark car, not only the light portions such as the roof and lamps, stand out in the 5th row than in the 4th row. The 6th row is also related to the same images. The difference with the preceding rows is that we adopt Eq.(7) as the optimality criterion rather than Eq.(6). The state estimation based on Eq.(7) is applied to the interested area in raster order and repeated three times. By incorporating the measure of the context-dependence, the results are obviously improved.

Some results about a "gray car" are shown in the rest part of Fig.1: the images in the 7th row, the results based on individual HMM regions in the 8th row, and the results in view of the context-dependence among HMM regions in the last row. The same problem of misclassification because of overlapping between the distribution of the foreground and that of the background concerns gray cars. However, since the variance of the background is usually much smaller than that of the shadow, the risk a gray car is confused with the background is lower, as you can see from Fig.1.

The state estimation process has been implemented on an SGI O2 R5000 SC 180 entry-level desktop workstation and allowed to run at the field-rate of 50 Hz (real time).

VI Conclusions and Future Work

We have described an improved HMM-based segmentation method which is designed to model the vehicles, the shadows of vehicles and the background for a traffic monitoring movie. A considerable advantage of this model is that unlike other approaches, it is not necessary to select the training data. All the HMM parameters are fully automatically estimated from an ordinary video sequence. The gray-value intensities and high frequency wavelet coefficients over time for one specific region location are modeled as a single HMM, independent of the neighboring regions. A state estimation algorithm is used to perform context-dependent classification of individual regions. Because this algorithm is only based on the past observations, state estimation can be performed in real time. Since all three distributions of different categories have a large overlap, it is impossible to construct a model which is purely based on intensity values. Using high frequency wavelet coefficients has improved the results significantly. This method itself has proved to be a low-level car tracking approach by experimental results. Since it runs comfortably in real time, it also offers the possibility of being used as a low-level process for a high-level tracking approach[19]. As the future work, it will be useful to expand the model itself to deal with both of temporal and spatial feature information.

References

[1] 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).

[2] K.O.Perlmutter, S.M.Perlmutter, R.M.Gray *et al.,* "Bayes risk weighted vector quantization with posterior estimation for image compression and classification," *IEEE Trans. Image Processing,* Vol.5, pp.347-360, Feb. 1996.

[3] C.H.Fosgate, H.Krim, W.W.Irving *et al.,* "Multiscale segmentation and anomaly enhancement of SAR imagery," *IEEE Trans. Image Processing,* Vol.6, pp.7-20, Feb. 1997.

[4] I.Daubechies, *Ten Lectures on Wavelets.* Philadelphia: SIAM,1992.

[5] L.E.Baum and T.Petrie, "Statistical inference for probabilistic functions of finite state Markov chains," *Ann. Math. Stat.,* Vol.37, pp.1554-1563, 1966.

[6] L.E.Baum and J.A.Eagon, "An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology," *Bull. Amer. Math. Stat.,* Vol.37, pp.360-363, 1967.

[7] L.E.Baum and T.Petrie, G.Soules, and N.Weiss, "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains," *Ann. Math. Stat.,* Vol.41, No.1, pp.164-171, 1970.

[8] L.E.Baum, "An inequality and associated maximization technique in statistical estimation for probabilistic functions of finite state Markov chains," in *Inequalities 3*, New York: Academic, pp.1-8, 1972.

[9] J.K.Baker, "The dragon system—An overview," in *Proc. Int. Conf. Acoust., Speech, Signal Process.,* pp.24-29, Feb. 1975.

[10] X.D.Huang, Y.Ariki, and M.A.Jack, *Hidden Markov Models for Speech Recognition*. Edinburgh, U.K.: Edinburgh University Press, 1990.

[11] D.B.Paul, "Speech recognition using hidden Markov mesh," *Lincoln Lab. J.,* Vol.3, No.1, pp.41-62, 1990.

[12] R.Cole, L.Hirschman, L.Atlas, and M.Beckman *et al.,* "The challenge of spoken language system: Research directions for the nineties," *IEEE Trans. Speech Audio Processing*, Vol.3, pp.1-21, Jan. 1995.

[13] A.P.Dempster, N.M.Laird and D.R.Rubin, "Maximum likelihood from incomplete data via the EM Algorithm," *J. R. Stat. Soc.*, B 39:1-38, 1977.

[14] 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.Stat.*, Vol.41, No.1, pp.164-171, 1970.

[15] J.Z.Wang, J.Li, R.M.Gray and G.Wiederhold, "Unsupervised multiresolution segmentation for images with low depth of field," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol.23, No.1, pp.85-90, 2001.

[16] L.R.Rabiner, "A tutorial on Hidden Markov Models and selected applications in speech recognition," *In Proc.of the IEEE*, Vol.77, No.2, pp.257-286, Feb., 1989.

[17] A.J.Viterbi, "Error bounds for convolutional codes and an asymptotically optimal decoding algorithm," *IEEE Trans. Informat. Theory*, Vol.IT-13, pp.260-269, Apr., 1967.

[18] G.D.Forney, "The Viterbi algorithm," *In Proc.IEEE*, Vol.61, pp.268-278, Mar., 1973.

[19] M.Isard and A.Blake, "ICONDENSATION: Unifying low-level and high-level tracking in a stochastic framework," *In Proc. 5th European Conf. on Computer Vision*, pp.893-908, 1998.

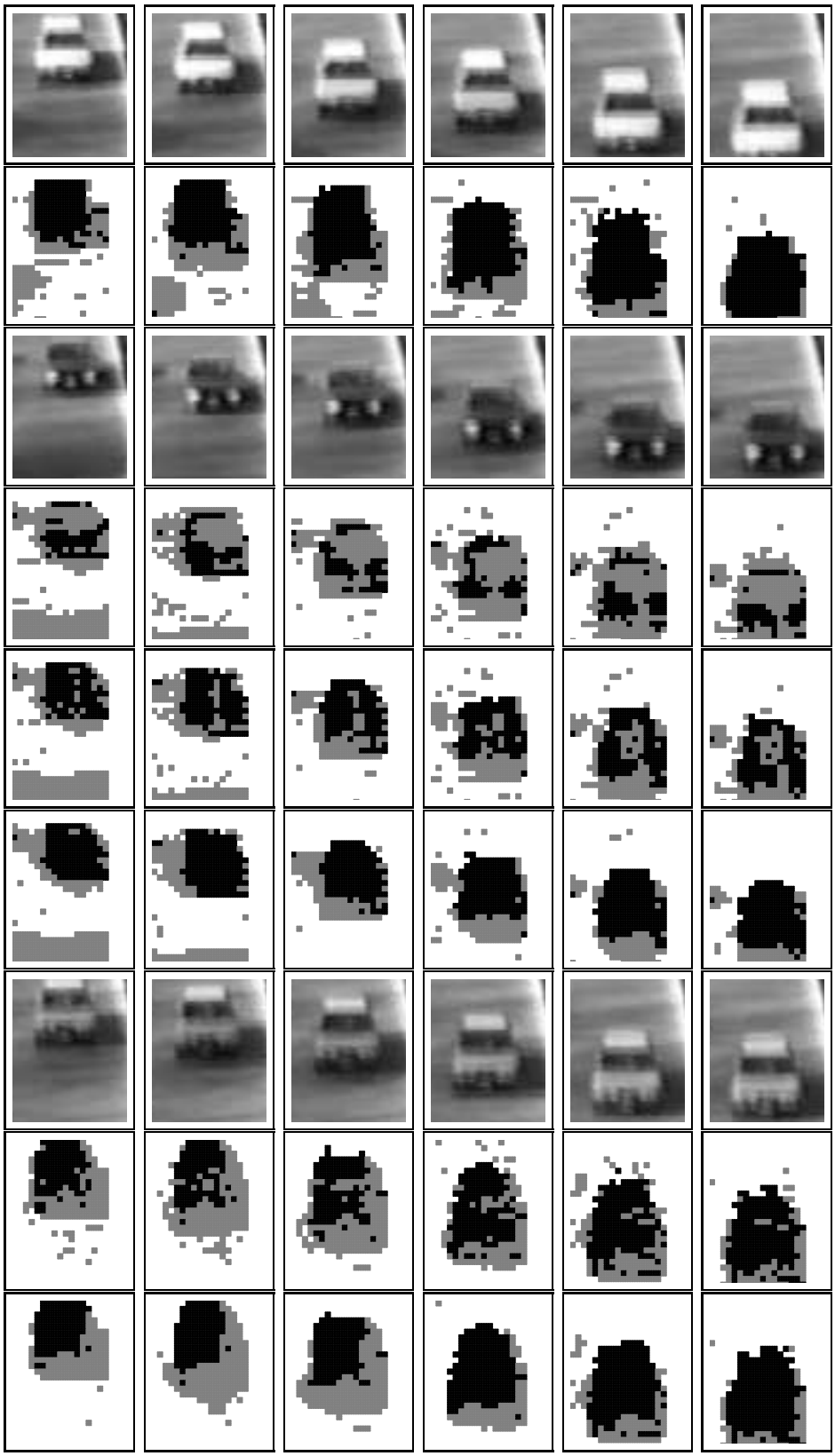

Figure 1: The visualization of the results of state estimation for an interested area. Foreground: black, shadow: gray and background: white. The images in a row are taken at three-field intervals from a test sequence.