PAPER

Object Type Classification Using Structure-based Feature Representation

Tomoyuki NAGAHASHI[†a)], ***Nonmember***, **Hironobu FUJIYOSHI**[†b)], ***and*** **Takeo KANADE**[††c)], ***Members***

SUMMARY Current feature-based object type methods classify on texture and shape by using information derived from image patches. Generally, input features, such as the aspect ratio, are derived from the rough characteristics of an entire object. However, we derive input features from a parts-based representation of an object. We have developed a method that distinguishes object types using structure-based features described by a Gaussian mixture model. This approach uses Gaussian fitting of foreground pixels detected by background subtraction to segment an image patch into several sub-regions, each of which is related to a physical part of an object. The object is modeled as a graph, where the nodes contain SIFT (scale invariant feature transform) information obtained from the corresponding segmented regions and the arcs contain information on the distance between two connected regions. By calculating the distance between the reference and input graphs, we can use a k-NN-based classifier to classify an object as: single human, human group, bike, or vehicle. We found that higher classification performance can be obtained using both the conventional and structure-based features together compared with using either alone.

key words: *SIFT, feature representation, object classification.*

1. Introduction

Feature-based methods are commonly used for object recognition and type classification in visual surveillance [1]. For robustness, we need features that are invariant to changes caused by the environment, scaling, viewpoint, and lighting.

Previous work in this area has focused on producing descriptors and a classification method that is invariant to the scaling and viewpoint of detected objects. Lipton et al. [2] proposed a binary classification method that uses two feature vectors, dispersedness and area, to distinguish an image blob detected by adaptive background subtraction. The automated video surveillance system, called VSAM [1][3], uses a classification method based on an artificial neural network that enables classification robust to size changes (by using information about the zoom parameter of a camera). Since both of these features are only shape-based, the performance is not high. Texture-based features, such as histograms of oriented gradients for human detection, have been proposed [4]. This method computes high-dimensional features based on arcs and uses SVM (binary classification) to detect human regions. Viola and Jones have proposed a pedestrian detection system that integrates intensity and motion information [5]. In general, input features, which are used in conventional approaches for object type classification, are derived from the rough characteristics of an entire object. However, we derive input features from the parts-based representation of an object.

In this paper, we propose a method that distinguishes object types using structure-based features described by a Gaussian mixture model. Our method uses Gaussian fitting of an object image to segment it into several sub-regions, each of which is related to a physical part of the object. We model the object as a graph. The nodes contain the vector quantization histograms of SIFT (scale invariant feature transform) obtained from the corresponding segmented regions, and the arcs contain information on distances between two connected regions. By calculating the distance between the reference and input graphs, we can use a k-NN-based classifier to classify an object into one of the following categories: single human, human group, bike, or vehicle. We found that higher classification performance can be obtained using both the conventional and structure-based features together compared with using either set of features alone. We also found that the proposed method is robust to rotation changes compared with the bag-of-keypoints approach.

2. Structure-based Feature Representation

Our approach uses Gaussian fitting of the foreground pixels to segment an image patch into several sub-regions, each of which is related to a physical part of the object. We model the object as a graph. The nodes contain the vector quantization histograms of SIFT obtained from the corresponding segmented regions, and the arcs contain information about distances between two connected regions.

2.1 GMM-based Segmentation

Seki et al. [6][7] have proposed a method for modeling a class of objects. They use the Gaussian mixture model (GMM) to describe topological structures of an

[†]Dept. of Computer Science Chubu University
[††]The Robotics Institute Carnegie Mellon University
a) E-mail: kida@vison.cs.chubu.ac.jp
b) E-mail: hf@cs.chubu.ac.jp
c) E-mail: tk@cs.cmu.edu

object's internal patterns. This approach also eliminates influences caused by individual pattern differences. We thus applied the GMM in order to segment a detected object into several regions. Let $\boldsymbol{x} = \{u, v, I\}^T$ denote coordinate (u, v) and intensity I in the image, let $\boldsymbol{\Phi} = \{\alpha_j, \boldsymbol{\phi}_j = (\boldsymbol{\mu}_j,$ and $\boldsymbol{\Sigma}_j)\}_{j=1}^c$ denote the GMM parameter and c denote the number of Gaussian components. To fit the GMM, we use the deterministic annealing EM (DAEM) algorithm [8] to estimate the parameters $\boldsymbol{\Phi}_{ML}$ with the following equation:

$$\begin{aligned} \boldsymbol{\Phi}_{ML} &= \arg\max_{\boldsymbol{\Phi}} \sum_{j=1}^{c} (\alpha_j \cdot p_j(\boldsymbol{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j))^{\beta} \\ p(\boldsymbol{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) &= \frac{1}{\sqrt{(2\pi)^3 |\boldsymbol{\Sigma}_j|}} \cdot \\ & \exp\left\{ -\frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}_j^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_j) \right\}, \end{aligned} \tag{1}$$

where $\boldsymbol{\mu}_j$ is the average, $\boldsymbol{\Sigma}_j$ is the covariance matrix, $\boldsymbol{\phi}_j = \{\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\}$ is each Gaussian parameter, β is the annealing parameter, and α_j is the mixture ratio ($\alpha_j > 0$, $\sum_{j=1}^{c} \alpha_j = 1$). Figure 1 shows an example of GMM fitting using a three-dimensional Gaussian model expressed as $\boldsymbol{\Phi}_{ML}$ projected onto the (u, v) plane. We see that each Gaussian distribution corresponds to the internal pattern of an object.

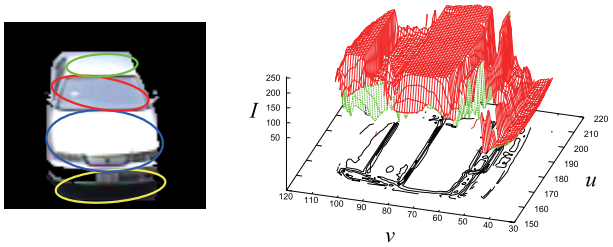


Fig. 1 Example of GMM fitting for detected pixels

2.1.1 Region Segmentation by Mixture of Gaussian Distribution

We developed a region segmentation method using Gaussian distribution parameter $\boldsymbol{\phi}$. A detected pixel $\boldsymbol{x}$ can be distinguished from the sub-region C_j using the following equation:

$$C_j = \arg\max_j p_j(\boldsymbol{x}|\boldsymbol{\phi}_j). \tag{2}$$

Figure 2 shows examples of GMM-based segmentation. We found that each Gaussian distribution corresponds to the physical part of an object. Figure 3 shows a comparisones between the proposed and conventional methods (mean-shift clustering [9]) for region segmentation. We found that dividing the side and back of the vehicle is difficult using mean-shift clustering. However, the proposed method can divide the sub-regions into a useful because thise method clusters the region in the $\{u, v, I\}^T$ space.

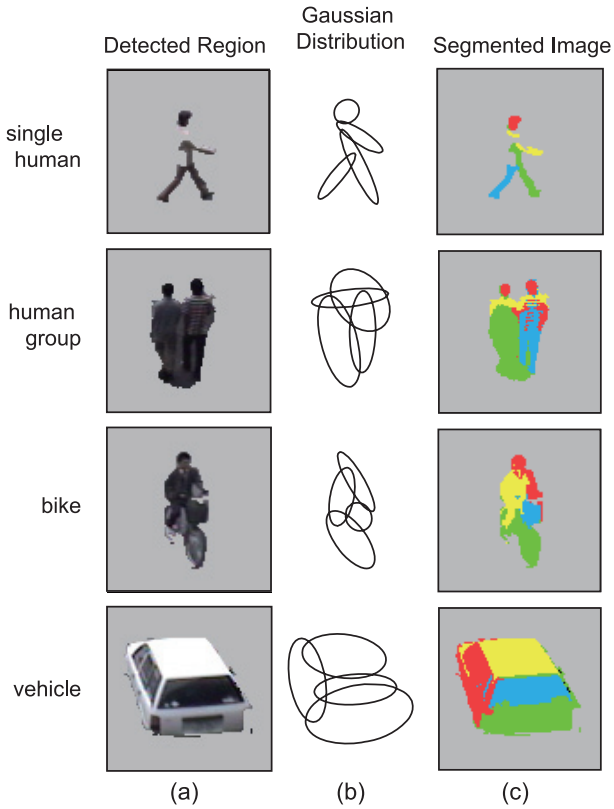


Fig. 2 Examples of GMM-based segmentation

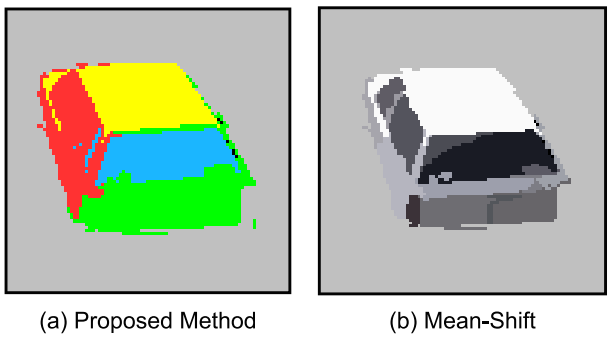


Fig. 3 Example of segmentation results. (a) Segmentation results with proposed method, and (b) segmentation results with mean-shift clustering. Proposed method represents structural information better than mean-shift clustering.

2.2 Feature Extraction

At each pixel, SIFT features are extracted. Vector quantization is performed to make a histogram for each segmented region. The SIFT descriptor is depicted as a 128-dimensional vector from a normalized gradient orientation histogram.

2.2.1 SIFT Descriptor

The SIFT descriptors are computed for normalized im-

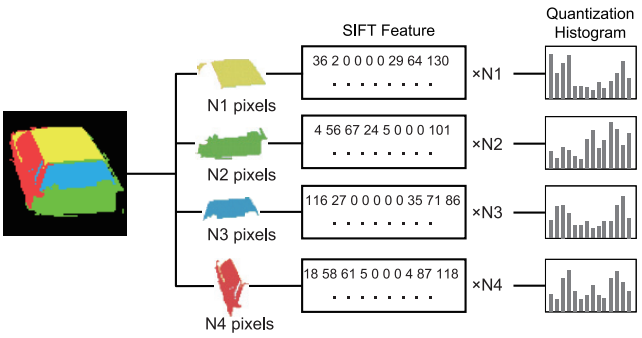


Fig. 4 Feature extraction

age patches with a code provided by Lowe [10]. A gradient orientation $\theta(x, y)$ and magnitude $m(x, y)$ of image $L(x, y)$ is computed as:

$$m(x,y) = \sqrt{f_x(x,y)^2 + f_y(x,y)^2} \tag{3}$$

$$\theta(x,y) = \tan^{-1}\left(\frac{f_y(x,y)}{f_x(x,y)}\right), \tag{4}$$

where $f_x(x, y) = L(x+1, y) - L(x-1, y)$ and $f_y(x, y) = L(x, y+1) - L(x, y-1)$. A gradient orientation histogram is given by:

$$h_\theta = \sum_x \sum_y w(x,y) \cdot \delta[\theta, \theta(x,y)] \tag{5}$$

$$w(x,y) = G(x,y,\sigma) \cdot m(x,y) \tag{6}$$

$$\delta[\theta, \theta(x,y)] = \begin{cases} 1 & \text{if} \quad \theta = \theta(x,y) \\ 0 & \text{otherwise} \end{cases}, \tag{7}$$

where $G(x, y, \sigma)$ is the Gaussian distribution, and θ is 36 bins covering the 360° range of orientations. The SIFT features are local histograms of edge directions computed over different parts of the region of interest. Using eight orientation directions and a 4×4-grid gives the best results, leading to a descriptor size of 128.

2.2.2 Vector Quantization Histogram

We cluster a SIFT descriptor to make a codebook. The codebook is the center of the cluster. More specifically, we apply the LBG algorithm [11] to a set of local descriptors extracted from training images, and continue using the SIFT descriptor. We used Euclidean distance in the clustering and quantization processes.

Finally, a vector quantization histogram representation is constructed from local descriptors in accordance with:

$$\boldsymbol{n} = \{n(C, v_1), \ldots, n(C, v_N)\}, \tag{8}$$

where $n(C, v)$ denotes the number of occurrences of SIFT descriptor v_i in sub-region C, and N denotes the number of clusters.

3. Object Type Classification of Graph Matching

3.1 Graph Representation For Structure-based Features

We modeled the object as a graph. The nodes contain the vector quantization histogram based on the SIFT features obtained from the corresponding segmented regions, and the arcs contain information on distances between two connected regions. By calculating the distance between the reference and input graphs, we can use a k-NN based classifier to classify an object (Figure 5).

3.2 Graph Matching

We constructed a complete graph. The nodes contain the vector quantization histograms for each segmented region, and the arcs contain the Euclidean distance between two connected regions. The arc feature is given by:

$$\boldsymbol{e}_{ij} = \sqrt{(m_u^i - m_u^j)^2 + (m_v^i - m_v^j)^2}, \tag{9}$$

where m is the mean of the Gaussian distribution. Let $\boldsymbol{N} = \{\boldsymbol{n}_1, \ldots, \boldsymbol{n}_c\}^T$ denote a set of nodes, and $\boldsymbol{E} = \{\boldsymbol{e}_{12}, \ldots, \boldsymbol{e}_n\}^T$ denote a set of arcs. The distance between reference graph $\boldsymbol{T} = \{\boldsymbol{N}^t, \boldsymbol{E}^t\}^T$ and input graph $\boldsymbol{X} = \{\boldsymbol{N}^x, \boldsymbol{E}^x\}^T$ is given by

$$cost(\boldsymbol{T}, \boldsymbol{X}) = \frac{w_n}{c} \sum_{j=1}^{c} ||\boldsymbol{n}_j^t - \boldsymbol{n}_j^x|| + \frac{w_e}{n} \sum_{k=1}^{n} ||\boldsymbol{e}_k^t - \boldsymbol{e}_k^x|| \tag{10}$$

$$w_n + w_e = 1, \tag{11}$$

where w_n and w_e are weight parameters. Since the correspondence of the nodes between $\boldsymbol{T}$ and $\boldsymbol{X}$ is unknown, the cost of all combinations of $\boldsymbol{T}$ and $\boldsymbol{X}$ nodes are calculated (Figure 6). The minimum cost is then selected from all combinations of $\boldsymbol{T}$ and $\boldsymbol{X}$ as

$$Cost(\boldsymbol{T}, \boldsymbol{X}) = \min_{i \in {}_cP_c} \{cost(\boldsymbol{T}, \boldsymbol{X}_i)\}. \tag{12}$$

A final matching score is calculated by the following equation:

$$Cost = \alpha \cdot Cost_l + (1 - \alpha) \cdot Cost_g, \quad (0 \le \alpha \le 1) \tag{13}$$

where $Cost_l$ is the cost calculated by structure-based feature representation, and $Cost_g$ is the cost calculated by the conventional approach. The cost calculated by the conventional approach is $c = 1$. By calculating the matching cost between the input and reference graphs, we can classify an object using a k-NN-based classifier.

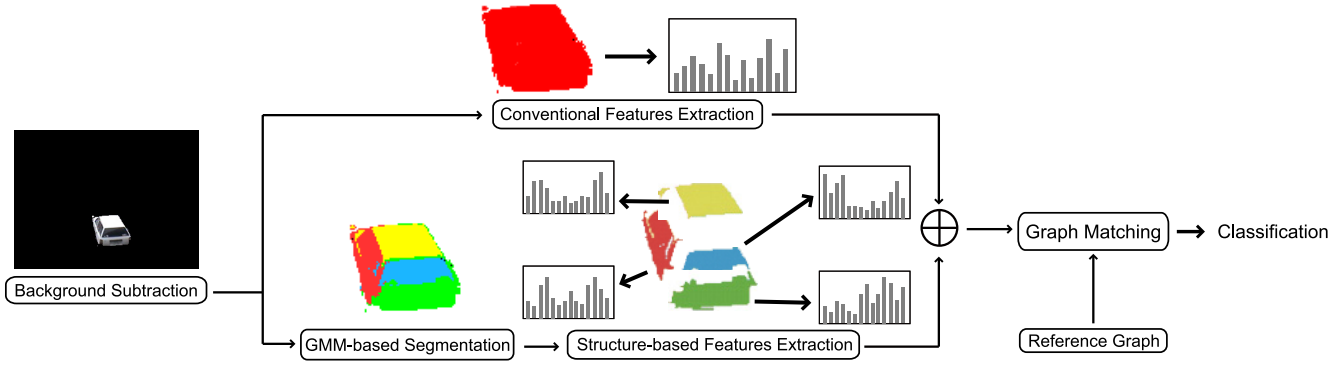


Fig. 5 Outline

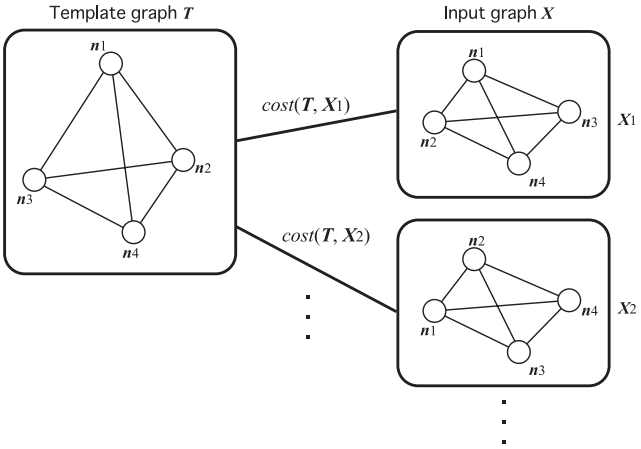


Fig. 6 Correspondence of nodes between $\boldsymbol{T}$ and $\boldsymbol{X}$

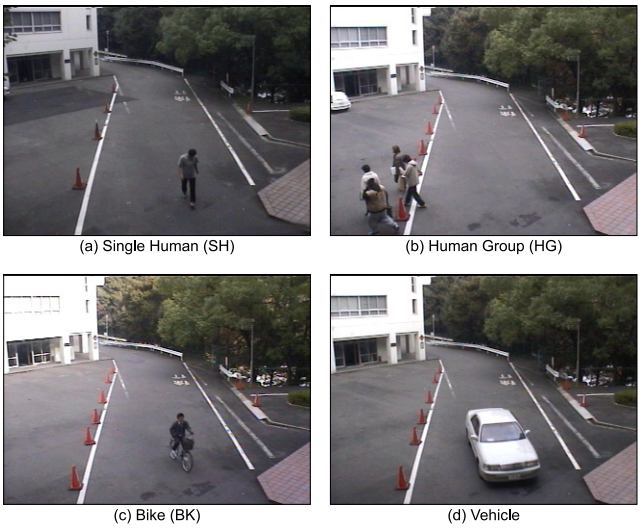


Fig. 7 Example of video images

4. Experimental Results

We have present the results of from two experiments. For the first experiment, we describe the results of using a four-class dataset an outdoor surveillance site. For the second experiment, we describe the results of using generic object recognition and compare them with a conventional approach with bags-of-keypoints [12][13].

4.1 Experiment on Surveillance Data

4.1.1 Dataset

We collected 200 images for our learning sample for each category (SH:single human, HG:human group, BK:bike, VH:vehicle) from a video database for 23 hours. A total of 800 images was used for training. A human operator collected sample images and assigned them class labels. Another 800 images were used for the discriminating experiments described below. Figure 7 shows examples of video images used in this experiment.

4.1.2 Results

We tested structure-based classification with about 200 sample images for each class, which were not contained in the training sets. Table 1 shows the classification results when α changed. In the conventional feature ($\alpha = 0$) , HG has a lot of variations because a person's position changed. However, in the structure-based feature ($\alpha = 1$), the classification rate is high because each region represents the person. In addition, the conventional feature is better than the structure-based feature in VH. The classification accuracy for four classes was to be about 88.2%. A higher classification performance can be obtain using both the conventional and structure-based features together compared with using either set of features alone.

Table 2 shows a confusion matrix of the classification results when $\alpha = 0.1$. Although the appearances of a single human and bike are very similar from some viewpoints, structure-based feature representation can distinguish them correctly using information obtained

from the bottom part of a sub-region (Figure 8). Figure 9 shows an example of correct data using conventional and structure-based features.

Table 1 Classification rate [%]

		α						
		0.0	0.1	0.3	0.5	0.7	0.9	1.0
class	SH	75.6	80.8	77.5	74.7	71.4	70.9	71.8
	HG	80.4	87.1	85.7	85.7	85.7	85.2	85.2
	BK	86.3	87.7	86.3	87.2	86.3	85.3	85.8
	VH	97.3	96.8	95.9	95.9	96.4	96.4	96.4
	Total	85.0	88.2	86.4	85.9	85.0	84.5	84.9

Table 2 Confusion matrix ($\alpha = 0.1$)

		out					
		SH	HG	BK	VH	correct	rate[%]
in	SH	172	24	16	1	172	80.8
	HG	9	182	16	2	182	87.1
	BK	15	10	185	1	185	87.7
	VH	7	0	0	212	212	96.8
	Total					751	88.2

4.2 Experiment on Generic Object Recognition

In the second experiment, we performed generic object recognition in the Caltech 256 database †. Object boundaries in the Caltech 256 database, which are used in this experiment, are extracted in advance by a human operator. Images are shown in Figure 10. Five images in each category were used for a reference pattern. The evaluation data consists of positive class 103 images and negative class 2524 images.

In this experiment, we compared the following three methods: bag-of-keypoints [12](bok1), [13](bok2), and the proposed method.

bok1 [12] Bag-of-keypoints, which is often used for object categorization, is based on vector quantization of SIFT descriptors of image patches. This method is robust to background clutter and produces good caterorization accuracy even without exploiting geometric information.

bok2 [13] This method subdivides an image in to a grid and computes histograms of an image features over the resulting sub-regions.

proposed method Our method is GMM-based segmentation, which computes histograms of an image features over the resulting sub-regions.

†http://www.vision.caltech.edu/Image_Datasets/Caltech256/

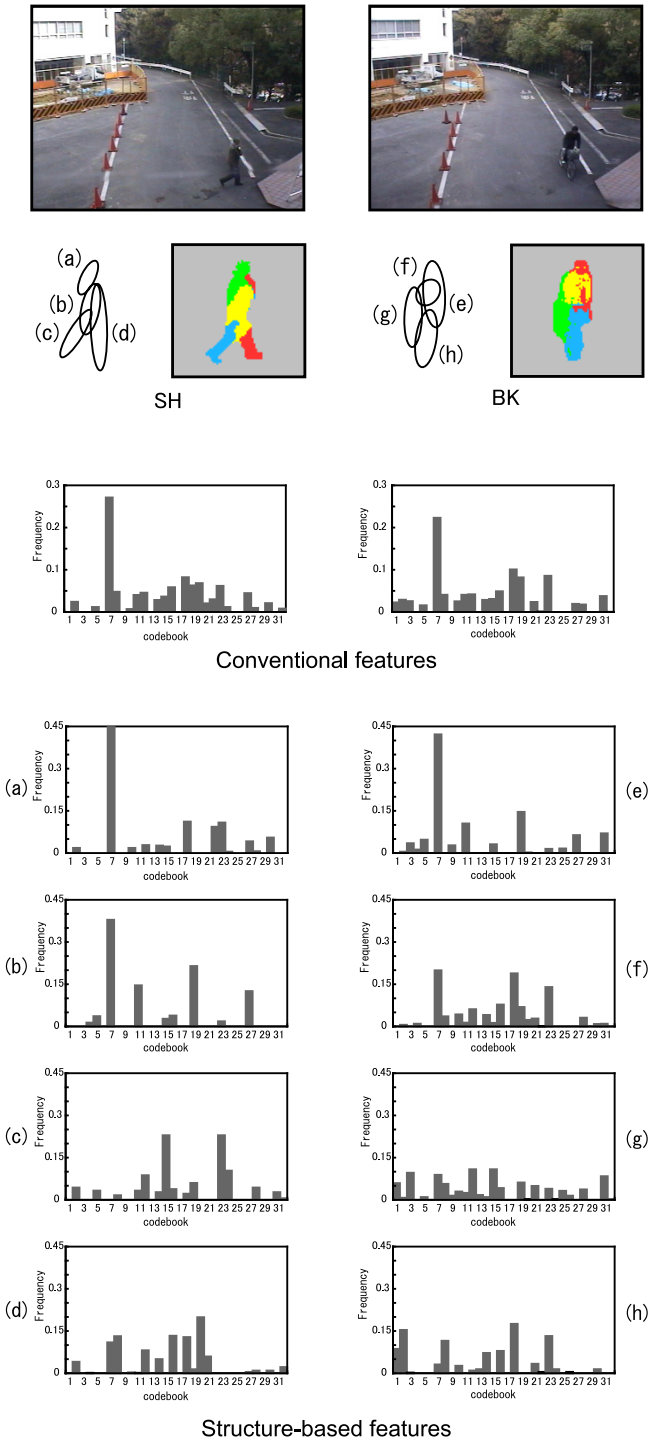


Fig. 8 Structure-based feature for mis-classification pattern (single human and bike). In the conventional feature, the SH and BK features are similar. However, the structure-based feature representation can distinguish them correctly using information obtained from the bottom part of a sub-region (c)(d) and (g)(h).

4.2.1 Results

Figure 11 shows classification results obtained using

Fig. 9 Example of correct sample with proposed method

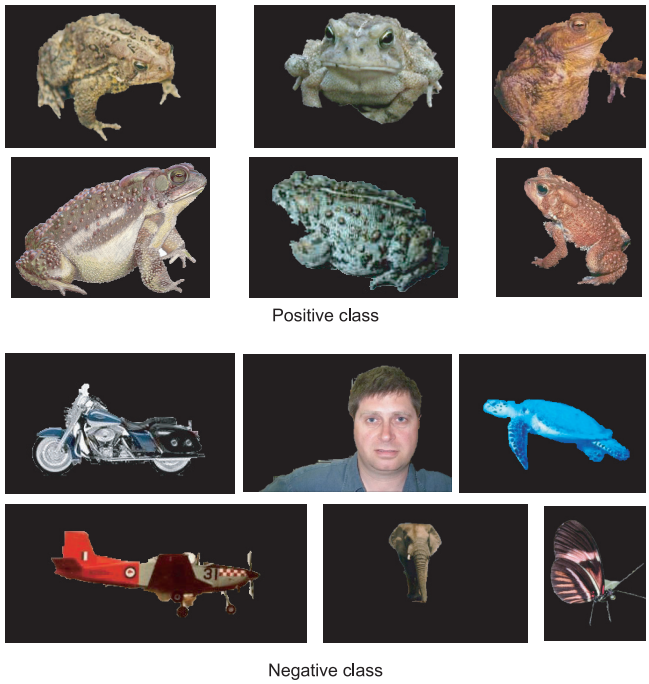


Fig. 10 Example of Caltech Images

bok1[12], bok2[13] and the proposed method. As shown in Figure 11, the proposed method can achieve 17.6% better classification than that with bok1. However, we can see that bok2 can achieve 5.6% better classification than that with the proposed method. This is because bok2 exploits geometric information. Correspondence of the sub-regions is clear in bok2. However, the correspondence of the sub-regions is unknown in the proposed method. Therefore, the proposed method can make a mistake in the correspondence of the sub-regions. When input images are rotated 45 degrees, the classification rate of the proposed method does not change. However, the classification rate of bok2 decreases. This is because GMM-based segmentation is robust to rotation changes. Figure 12 shows an exam-

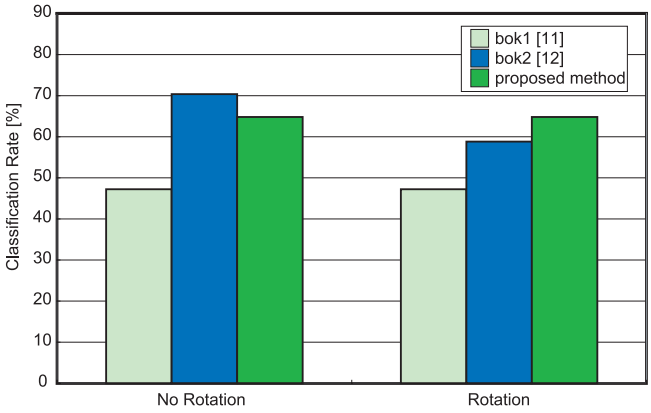


Fig. 11 Classification results.

ple of a GMM-based segmentation of a rotated image. We can see that the segmentation result of the rotated image is the same as that of the original image. With bok2, geometric information changes when the image is rotated. Therefore, bok2 cannot be correctly match.

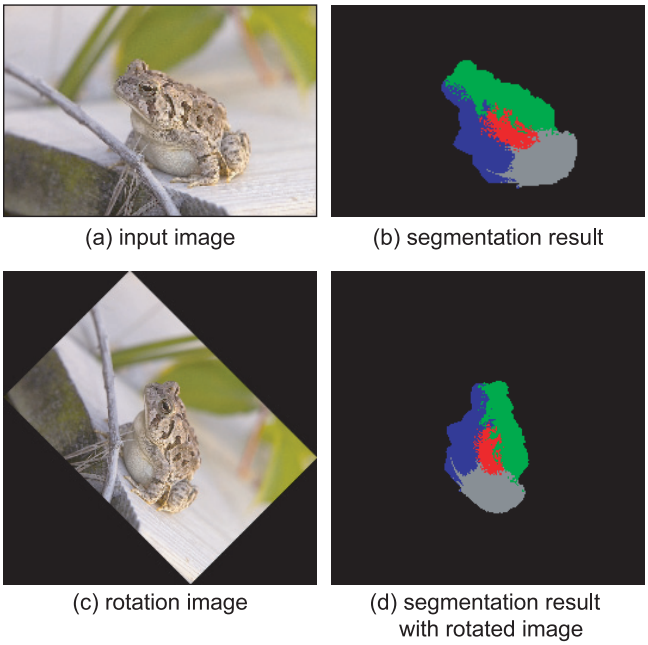


Fig. 12 Segmentation result by rotation image.

5. Conclusion

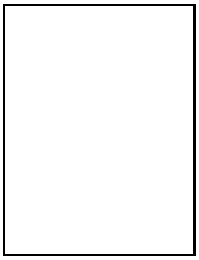
We developed an method for object type classification using structure-based feature representation. We proposed GMM-based segmentation and object classification by graph matching using SIFT. The effectiveness of integrating conventional and structure-based features was confirmed through experimentation. A

higher classification performance can be obtained using both the conventional and structure-based features together with using either set of features alone. We also confirmed that our method can extract appearance and geometric information that is robust to rotation changes.

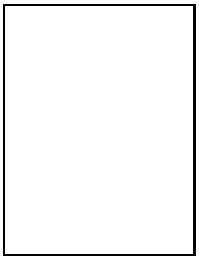
References

[1] R. Collins, A. Lipton, H. Fujiyoshi, and T. Kanade, "Algorithms for cooperative multisensor surveillance,", pp.1456 – 1477, October 2001.
[2] A. Lipton, H. Fujiyoshi, and R. Patil, "Moving target classification and tracking from real-time video," Proc. of the 1998 DARPA Image Understanding Workshop (IUW'98), November 1998.
[3] O. Hasegawa and T. Kanade, "Type classification, color estimation, and specific target detection of moving targets on public streets," vol.16, no.2, pp.116–121, February 2005.
[4] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," International Conference on Computer Vision & Pattern Recognition, ed. C. Schmid, S. Soatto, and C. Tomasi, INRIA Rhône-Alpes, ZIRST-655, av. de l'Europe, Montbonnot-38334, pp.886–893, June 2005.
[5] P. Viola, M.J. Jones, and D. Snow, "Detecting pedestrians using patterns of motion and appearance," ICCV, vol.02, p.734, 2003.
[6] M. Seki, K. Sumi, H. Taniguchi, and M. Hashimoto, "Gaussian mixture model for object recognition," MIRU2004, vol.1, pp.344–349, 2004.
[7] N. Hirata, M. Seki, H. Okuda, and M. Hashimoto, "Vehicle detection using gaussian mixture model from IR image," Technical report of IEICE. PRMU, vol.105, no.62, pp.37–42, 2005.
[8] N. Ueda and R. Nakano, "Deterministic annealing em algorithm," Neural Netw., vol.11, no.2, pp.271–282, 1998.
[9] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.24, no.5, pp.603–619, 2002.
[10] D.G. Lowe, "Distinctive image features from scale-invariant keypoints," Int. J. Comput. Vision, vol.60, no.2, pp.91–110, 2004.
[11] D. Linde, A. Buzo, and R.M. Gray, "An algorithm for vector quantizer design," IEEE Trans. Commun., vol.28, no.1, pp.84–95, 1980.
[12] C. Dance, J. Willamowski, L. Fan, C. Bray, and G. Csurka, "Visual categorization with bags of keypoints," ECCV International Workshop on Statistical Learning in Computer Vision, 2004.
[13] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," CVPR '06: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Washington, DC, USA, pp.2169–2178, IEEE Computer Society, 2006.

Tomoyuki Nagahashi received a bachelor's degree in computer science from Chubu University, Japan, in 2006. Currently he is a Masters student at the Department of Computer Science, Chubu University.



Hironobu Fujiyoshi received the Ph.D. degree in electrical engineering from Chubu University, Japan, in 1997. For his thesis, he developed a fingerprint verification method using spectrum analysis, which has been incorporated into a manufactured device sold by a Japanese security company. He is a Member of Faculty at the Department of Computer Science, Chubu University. From 1997 to 2000 he was a post-doctoral fellow at the Robotics Institute of Carnegie Mellon University, Pittsburgh, PA, USA, working on the DARPA Video Surveillance and Monitoring (VSAM) effort and the humanoid vision project for the Honda Humanoid Robot. He performs research in the areas of real-time object detection, tracking, and recognition from video.



Takeo Kanade received the B. E. degree in electrical engineering in 1968, the M. E. degree in 1970, and the Ph. D. degree in 1973 from Kyoto University, Japan. Currently, he is Helen Whitaker Professor of Computer Science at Carnegie Mellon University, Pittsburgh, PA, USA. Dr. Kanade was a recipient of the Robotics Industry Association, Joseph F. Engelberger Award 1995, the Japan Robotics Association, JARA Award in 1997, the Yokogawa Prize at the International Conference on Multi Sensor Fusion and Integration for Intelligent Systems in 1997, the Hip Society, Otto AuFranc Award in 1998, the Hoso Bunka Kikin Foundation Award in 1994, and the Marr Prize at The Third International Conference on Computer Vision in December 1990. He was also selected as the author of one of the most influential papers that appeared in the Artificial Intelligence journal in the last ten years in 1992. He is Founding Chief Editor of the *International Journal of Computer Vision*. He is a Member of the National Academy of Engineering and a Fellow of the American Association for Artificial Intelligence (AAAI).