Object Type Classification Using Structure-based Feature Representation

Tomoyuki Nagahashi
Dept. of Computer Science
Chubu University
Aichi, Japan
kida@vison.cs.chubu.ac.jp

Hironobu Fujiyoshi
Dept. of Computer Science
Chubu University
Aichi, Japan
hf@cs.chubu.ac.jp

Takeo Kanade
The Robotics Institute
Carnegie Mellon University
Pittsburgh, PA, USA
tk@cs.cmu.edu

Abstract

Current feature-based object type classification methods information of texture and shape based information derived from image patches. Generally, input features, such as the aspect ratio, are derived from rough characteristics of the entire object. However, we derive input features from a parts-based representation of the object. We propose a method to distinguish object types using structure-based features described by a Gaussian mixture model. This approach uses Gaussian fitting onto 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 the 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 edges contain information on 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 one of the following: single human, human group, bike, or vehicle. We demonstrate that we can obtain higher classification performance when using both conventional and structure-based features together than when using either alone.

1. Introduction

Feature-based methods are commonly used for object recognition and type classification in visual surveillance [2]. 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 are invariant to the scaling and viewpoint of detected objects. Lipton et al. [1] have proposed a binary classification method that uses two feature vectors, i.e. dispersedness and area, to distinguish an image blob detected by adaptive background subtraction. The automated video surveillance system, called VSAM [2, 3], uses classification based on an artificial neural network(ANN) 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 edges and use 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 rough characteristics of an entire object. However, we derive input features from parts-based representation of an object.

In this paper, we propose a method to distinguish object types using structure-based features described by a Gaussian mixture model. Our approach uses Gaussian fitting of an object's 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, where the nodes contain the vector quantization histograms of SIFT(Scale Invariant Feature Transform) obtained from the corresponding segmented regions, and the edges contain information of 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: a single human, human group, bike, or vehicle. We demonstrate that we can obtain higher classification performance when using both conventional and structure-based features together than when using either set of features alone.

2. Structure-based feature representation

Our approach uses the Gaussian fitting onto 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, where the nodes contain the vector quantization histograms of SIFT obtained from the corresponding segmented regions, and the edges contain information about distance 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 for the object's internal patterns. Moreover, this approach can eliminate influences causes by individual pattern differences. Thus, we apply the GMM to segment a detected object into several regions. Let $\boldsymbol{x}_i = \{u_i, v_i, I_i\}^T$ denote coordinate (u, v) and intensity I in the image and $\boldsymbol{\Phi} = \{\alpha_j, \boldsymbol{\phi}_j = (\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)\}_{j=1}^{c}$ denote the GMM parameter. To fit the GMM, we use the deterministec

annealing EM (DAEM) algorithm [9] to estimate the parameters $\mathbf{\Phi}_{ML}$ with the following equation:

$$\mathbf{\Phi}_{ML} = \arg\max_{\mathbf{\Phi}} \sum_{j=1}^{c} (\alpha_j \cdot p_j(\boldsymbol{x}|\boldsymbol{\mu}_j, \mathbf{\Sigma}_j))^{\beta}$$

$$p(\boldsymbol{x}|\boldsymbol{\mu}_j, \mathbf{\Sigma}_j) = \frac{1}{\sqrt{(2\pi)^3|\mathbf{\Sigma}_j|}} \cdot \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_j)^T \mathbf{\Sigma}_j^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_j)\right\}, \quad (1)$$

where $\boldsymbol{\mu}_j$ is the average, $\mathbf{\Sigma}_j$ is the covariance matrix, $\boldsymbol{\phi}_j = \{\boldsymbol{\mu}_j, \mathbf{\Sigma}_j\}$ 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 expressed as $\mathbf{\Phi}_{ML}$ to projected onto the (u, v) plane. We see that each Gaussian distribution corresponds to the internal pattern of an object.

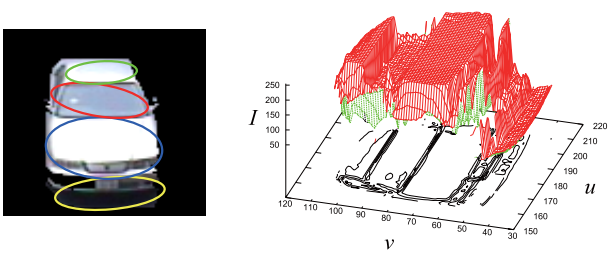

Figure 1: Example of GMM fitting for detected pixels.

Region Segmentation by Mixture of Gaussian Distribution We propose a method of region segmentation using Gaussian distribution parameter $\boldsymbol{\phi}$. A detected pixel $\boldsymbol{x}$ can be distinguished into the sub-region C_i using the following equation:

$$C_i = \arg\max_i p_i(\boldsymbol{x}|\boldsymbol{\phi}_i). \quad (2)$$

Figure 2 shows examples of GMM-based segmentation. We see that each Gaussian corresponds to the physical part of an object. Figure 3 compares the proposed and conventional methods (Mean-Shift clustering [10]) for region segmentation. We see 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 the proposed method clusters the region in the $\{u_i, v_i, I_i\}^T$ space.

2.2. Features Extraction

At each pixel, SIFT features are extracted. Then, 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.

SIFT Descriptor The SIFT descriptors are computed for normalized image patches with the code provided by Lowe [11]. A gradient orientation $\theta(x, y)$ and magnitude $m(x, y)$ of image $L(x, y)$ is computed as:

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

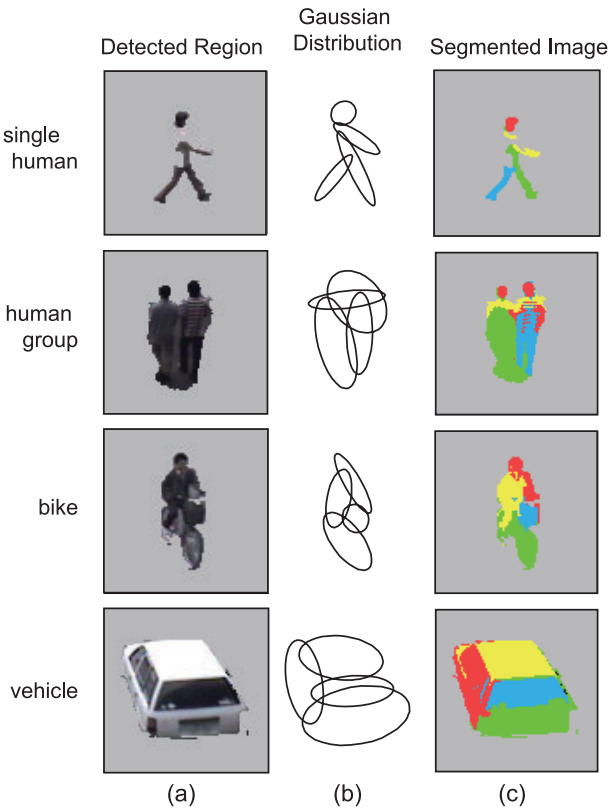

Figure 2: Examples of GMM-based segmentation.

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

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:

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

$$w(x,y) = G(x,y,\sigma) \cdot m(x,y), \quad (6)$$

where $G(x,y,\sigma)$ is the Gaussian distribution, and θ is 36 bins covering the 360° range of orientations. SIFT features are local histograms of edge directions computed over different parts of the region of interest. Using 8 orientation directions and a 4x4-grid gives the best results, leading to a descriptor of size 128.

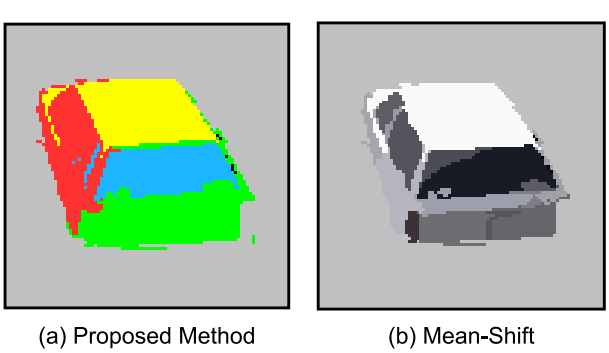

Figure 3: Difference Mean-Shift Segmentation.

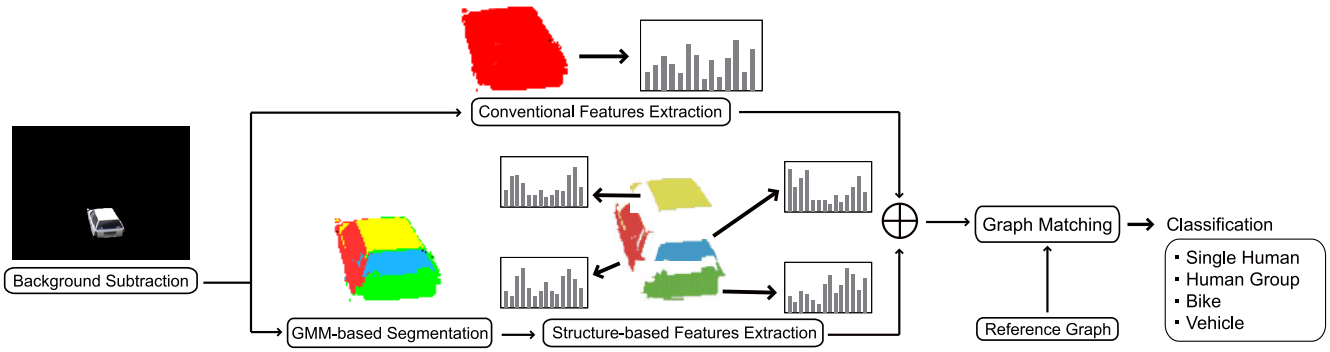

Figure 5: Outline.

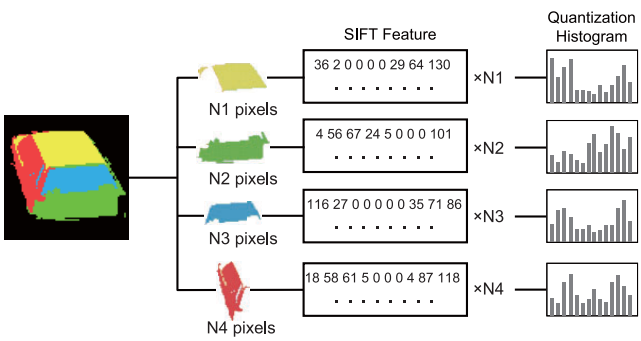

Figure 4: Feature Extraction.

Vector Quantization Histogram A codebook is created with the LBG algorithm using a SIFT descriptor. Vector quantization is performed to encode the SIFT descriptors using the codebook, which is trained in advance using 800 samples. We computed a vector quantization histogram for each segmented region, as shown in Figure 4. This histogram is normalized by the number of pixels belonging to the same region.

3. Object Type Classification of Graph Matching

3.1. Graph representation for structure-based features

We model the object as a graph, where the nodes contain the vector quantization histogram based on SIFT features obtained from the corresponding segmented regions, and the edges 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 as one of the following: a single human, human group, bike, or vehicle.

3.2. Graph Matching

The nodes contain the vector quantization histograms for each segmented region, and the edges contain the Euclidean distance between two connected regions. Let $\boldsymbol{N} = \{\boldsymbol{n}_1, \ldots, \boldsymbol{n}_4\}^T$ denote a set of the nodes, and $\boldsymbol{E} = \{\boldsymbol{e}_{12}, \ldots, \boldsymbol{e}_{34}\}^T$ denote a set of the edges. The distance between reference graph $\boldsymbol{T} = \{\boldsymbol{n}_1^t, \ldots, \boldsymbol{n}_4^t\}^T$ and input graph $\boldsymbol{X} = \{\boldsymbol{n}_1^x, \ldots, \boldsymbol{n}_4^x\}^T$ is given by

$$cost(\boldsymbol{T}, \boldsymbol{X}) = \sum_{i=1}^{4} ||\boldsymbol{n}_i^t - \boldsymbol{n}_i^x|| + \sum_{j=1}^{6} ||\boldsymbol{e}_j^t - \boldsymbol{e}_j^x||. \quad (7)$$

Since correspondence of the nodes between $\boldsymbol{T}$ and $\boldsymbol{X}$ is unknown, the cost of all combinations of $\boldsymbol{T}$'s nodes and $\boldsymbol{X}$'s node are calculated. Then, the minimum cost is selected from all combinations of $\boldsymbol{T}$ and $\boldsymbol{X}$ as

$$Cost(\boldsymbol{T}, \boldsymbol{X}) = \min_i \{cost(\boldsymbol{T}, \boldsymbol{X}_i)\}. \quad (8)$$

A final matching score is calculated by the following equation:

$$Cost = \alpha \cdot Cost_l + (1 - \alpha) \cdot Cost_g, \quad (0 \leq \alpha \leq 1) \quad (9)$$

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

4. Experimental Results

4.1. 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 of 23 hours. A total of 800 images was used for training. A human operator collected sample images and assigned class labels to them. Another 800 images were used for the discriminating experiments described below. Figure 6 shows examples of video image used in this experimentation.

4.2. 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. The classification accuracy for four classes found to be about 88.2%. We

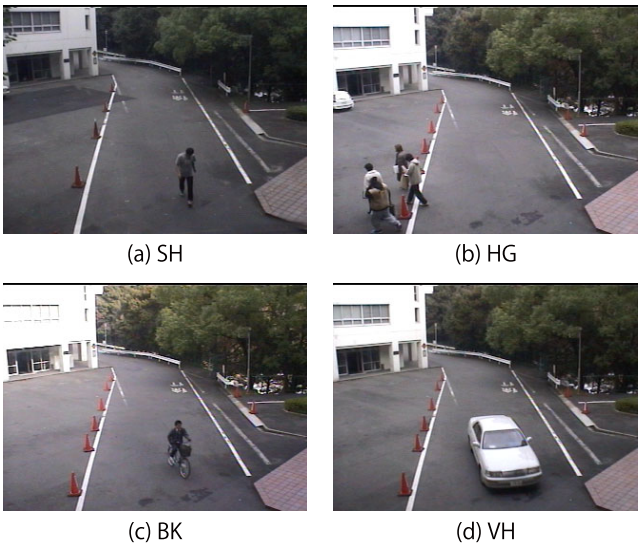

Figure 6: Example of video image.

Figure 7: Example of object classification.

can obtain higher classification performance when using both conventional and structure-based features together than when using either set of feature 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 7 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: Classification Result ($\alpha = 0.1$)

		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

5. Conclusion

We presented an approach to object type classification using structure-based feature representation. We proposed GMM-based segmentation and object classification by graph matching using SIFT. The effectiveness of the integrating conventional and structure-based features was confirmed by experimentation.

References

[1] A. Lipton, H. Fujiyoshi, and R.S. Patil, "Moving target detection and classification from real-time video.", Proc. of the 1998 Workshop on Applications of Computer Vision, 1998.

[2] R. Collins, A. Lipton, H. Fujiyoshi, and T. Kanade, "Algorithms for cooperative multisensor surveillance", Proc. of the IEEE, Vol. 89, No. 10, October, pp. 1456 - 1477, 2001.

[3] O. Hasegawa and T. Kanade, "Type Classification, Color Estimation, and Specific Target Detection of Moving Targets on Public Streets", Machine Vision & Applications, Springer, Vol.16, No.2, pp.116-121, 2005.

[4] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection", In Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, San Diego, CA, pp.886-893, 2005

[5] P. Viola, M. J. Jones, D. Snow. "Detecting Pedestrians Using Patterns of Motion and Appearance" Proc. of the Ninth IEEE International Conference on Computer Vision (ICCV'03) - Volume 2, 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", IEICE Technical Report PRMU2005-7, pp37-42, 2005.

[8] C. Stauffer and W.E.L Grimson, "Adaptive background mixture models for real-time tracking", Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, pp.246-252, 1999.

[9] N. Ueda, and R. Nakano, "Deterministic Annealing EM Algorithm", IEICE Journal(D-II), Vol. J80-D-II, No. 1, pp. 267-276, 1997.

[10] D. Comaniciu, and P. Meer, "Mean Shift Analysis and Applications", Proc. IEEE Seventh Int'l Conf. Computer Vision, vol. 2, pp. 1197-1203, 1999.

[11] D. G. Lowe, "Distinctive image features from scale-invariant keypoints", International Journal of Computer Vision, 60(2), pp. 91-110, 2004.