MVA2002 IAPR Workshop on Machine Vision Applications, Dec. 11 - 13, 2002, Nara - ken New Public Hall, Nara, Japan

15—2

A Vision System that Recognizes Objects on General Streets

Osamu Hasegawa*
†Imaging Science and Engineering Lab.
Tokyo Institute of Technology

‡Neuro Science Research Institute
Advanced Industrial Science and Technology

Takeo Kanade†
The Robotics Institute
Carnegie Mellon University

Abstract

This paper describes a vision based monitoring system which (1) classifies targets (vehicles and humans) based on shape appearance, (2) estimates their colors, and (3) detects special targets, from images of color video cameras set up toward a street. The categories of targets were classified into {human, sedan, van, truck, mule (golf cart for workers), and others}, and their colors were classified into the groups of {red-orange-yellow, green, blue-lightblue, white-silver-gray, darkblue-darkgray-black, and darkred-darkorange}. On the detection of special targets, the test was carried out setting {FedEx van, UPS van, Police Car} as target and yielded desirable results.

The system tracks the target, independently conducts category classification and color estimation, extracts the result with the largest probability throughout the tracking sequence from each result, and provides the data as the final decision. For classification and special target detection, we cooperatively used a stochastic linear discrimination method (linear discriminant analysis : LDA) and nonlinear decision rule (K-Nearest Neighbor rule: K-NN).

1 Introduction

Recently many cameras for monitoring and security have come to be seen on streets[1-4]. In the present situation, typically monitoring is done by human operators because the effects of solar location or weather make it difficult to actualize an automatic processing system that analyzes the contents of outdoor images. However, the automation of processing is desirable because it is usually difficult to monitor multiple images at the same time or because human operators become careless with exhaustion or habituation.

In this paper, we describe a system that is capable of online classification of vehicles and humans, estimation of their color and detection of special targets, and detects special targets from street video images.

The information obtained from targets gradually changes due to the following effects because targets of this system move outdoors. (1) They hide themselves behind other objects, (2) they are hidden in the shadows of buildings, (3) they strongly reflect the sunlight, and so on. Therefore, it is not always possible to obtain both the substantial information useful for classification of the categories of vehicle and human (image with less shapes broken), and the information necessary for estimation of their colors (image with less effects of strong sunlight reflections or shadows) at the same time.

Therefore, tracking the target, this system independently conducts category classification and color estimation, abstracts the result with the largest probability throughout the tracking sequence from each result, and provides the data as the final decision. Special target detection is carried out based on both classification and color estimation results.

The categories of targets were classified into {human, sedan, van, truck, small cart for workers (mule), and others}, and their colors were classified into the groups of {red-orange-yellow, green, blue-lightblue, white-silver-gray, darkblue-darkgray-black, darkred-darkorange}. On the detection of special targets, the test was carried out setting FedEx van, UPS van, Police Car as target and yielded desirable results.

2 Previous research

Jolly et. al. [5] have conducted a study to classify vehicles into {sedan, pickup truck, hatchback, station wagon, van} from the shape of its outline segmented applying "deformable template" to a image that is shot to show the whole side of the vehicle on its approximate central. However, the vehicle in the image has to be shot with its side wholly visible.

Kanade et. al.[6] have conducted a study to segment a moving object from a common street image using the background subtraction and classify it into {one human, humans, vehicle} from such parameters as the complexity of its outline, the ratio of areas, and the zoom value when it was shot. In that study, the image of the background is adaptively renewed and the parameters are learned by a three layer neural net.

On the contrary, the proposed system dealt with six categories of targets including human, and also estimated their color, by using fixed camera. The images with overlap of targets was dealt with by an original means (the details mentioned later), and an additional function was mounted to detect the special target designated by the user.

*Address: †4259, Midori-ku, Nagatsuta, Yokohama, Kanagawa, 226-8503 Japan, ‡1-1-1, Umezono, Tsukuba, 305-8566, Japan, E-mail: oh@ieee.org

†Address: 5000 Forbes Avenue Pittsburgh, PA, U.S.A., E-mail: tk@cs.cmu.edu

3 Outline of system

This system consists of a fixed camera that is set on the roof of a two-storied building and a personal computer (CPU: PentiumIII 450MHz) that process the input image from the camera. We do not use special hardware for image processing. The camera was aimed toward the street on which there were vehicle and human traffic. Figure 1 shows the input image sample from the camera. The camera was a commercial CCD camera and the NTSC standard video signal was input to the system.

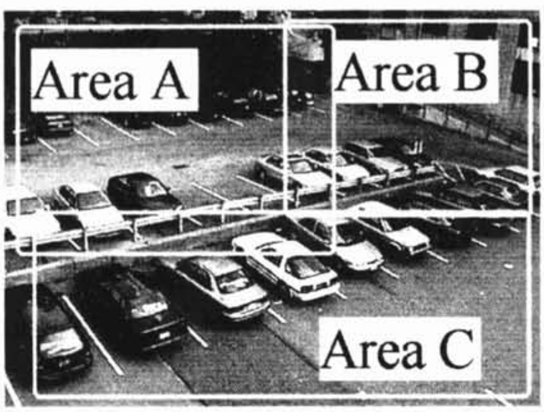


Figure 1: Input image and its segmentation.

The operating system (OS) of the computer was Linux and the program was described with C. The program consists of four modules, that are (a) a module to segment the target (vehicle or human) from the image, based on background subtraction and track it, (b) a module to classify the category of the target segmented, (c) a module to estimate the color of the target, (d) a module to decide synthetically the category from the above results of classification/estimation.

Module (a), which was studied and developed independently of this study[7], is able to renew the background image according to sunshine conditions or changes in weather, segment the target, and assign an ID to each target and track them. The details of the other three modules constructed for this study are as follows.

4 Classification of category based on shape appearance

4.1 Outline

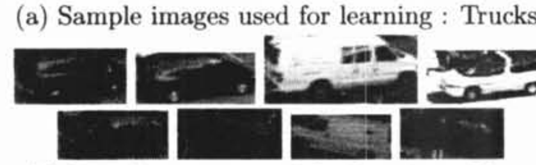
In this system, the classification targets were both vehicles and humans on common street. Thus it was difficult to make "models" of them in advance because there are many kinds of vehicles and humans sometimes walk in a group of more than two. Therefore, we adopted "appearance based approach" for category (shape) classification.

First we employed linear discriminant analysis (LDA) to learn variety of shape appearances and then classify them in the linear discriminant space. Scondly, we applied non-linear classification rule, K nearest neigbor (K-NN) rule, to them to obtain better classification results.

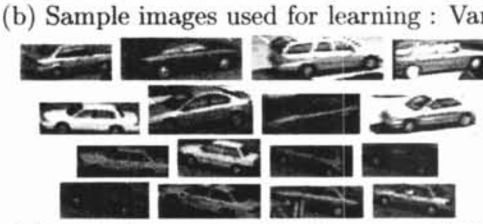
4.2 Collection of image samples for learning

Figure 1 shows the input sample image into the system. Its size is 640×480 pixels. In the image, the street appears from the left edge of the frame, passes through the parking lot, and then curves to the right and leads toward the bottom at the right edge. Vehicles and humans mainly pass on the street, but they sometimes appear on the bottom half of the frame (Area C).

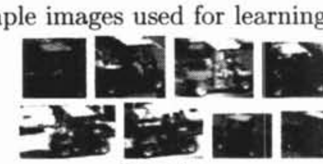
(a) Sample images used for learning : Trucks



(b) Sample images used for learning : Vans



(c) Sample images used for learning : Sedans



(d) Mules

Figure 2: Sample images used for learning (vehicles only)

Some 2000 images by background subtraction were collected as samples for learning. The processing speed from the extraction of a target image from the input image, tracking, and saving is three to four times per second when there was a single target to be extracted in the image. Sample images for learning were collected that varied in "appearance" as widely as possible. For example, even when targets of the same category varied in shape or in design, or even an identical target varied in "appearance" due to the direction of motion, they were included separately among the samples.

Shadows of vehicles or humans were also segmented as parts of the image when it was sunny. The samples were collected for all categories, whenever the shadows varied due to the location of sun.

The "appearance" (direction, size and so on) of the segmented target varies according to the position at which they are segmented. Therefore, the input image was divided into three areas, Area A, B, C, as shown in Figure 1.

The sample images for learning were collected from the online input images or the videotape image of 16 hours in total from 9:00 a.m. to sunset when it was sunny or cloudy.

Figure 2 shows the sample images segmented from Area A and B in Figure 1[1]. These show that vehicles to be classified into an identical category have various "appearances" (shapes).

The sample images were identified and the category label {human, sedan, van, truck, mule (small golf cart for workers), and others} was attached to them by the operator [2].

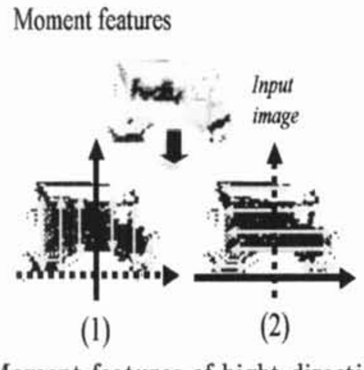


Figure 3: Moment features of hight-direction (1) and width-direction (2).

4.3 Learning of shape

Some 2000 sample images are doubled in height and width, and smoothed. 11-dementional feature vectors are composed from the data {width, height, width-direction and height-direction 1st, 2nd, 3rd moments of the image (see Fig.3), area of the image, centroid(x,y)} for each image.

The above vector data are sorted into classes (categories) corresponding to each area. These take LDA, and linear discriminant spaces for category classification are composed for each area.

Suppose B represents variance among all classes and W represents variance within a class, they are given by

$$W = \sum_{k=1}^{K}\sum_{i=1}^{n_k}(\boldsymbol{x}_{i,k} - \bar{\boldsymbol{x}}_k)(\boldsymbol{x}_{i,k} - \bar{\boldsymbol{x}}_k)^T \qquad (1)$$

$$B = \sum_{k=1}^{K} n_k(\bar{\boldsymbol{x}}_k - \bar{\boldsymbol{x}})(\bar{\boldsymbol{x}}_k - \bar{\boldsymbol{x}})^T \qquad (2)$$

K: number of classes.
n_k: number of objects in class k.
$\boldsymbol{x}_{i,k}$: vector of object i belonging to class k.
$\boldsymbol{x}_k$: centroid vector of class k.
T: transpose.

The following eigen value λ and eigen vector b_i are determined below.

$$(B - \lambda_i W) \times b_i = 0, \quad (\lambda_1 > \lambda_2 > ... > \lambda_N) \qquad (3)$$

[1]The Figure shows only part of the vehicles. There are also sample images of humans. The Figure shows the texture images within a box enclosing the targets in order to display it clearly. What were actually used for learning are the shading images of background subtraction.

[2]Then, if possible, a label of "direction" according to the course direction of the target (right or left) was attached to the sample images as attribute information.

where N represents a dimensional number of eigen vectors for learning, and these are the contribution ratio of the corresponding eigen vector b_i. In this study, we assumed that the dimension of discriminant space had reached its upper limit (which is assumedly M dimension) when the cumulative contribution ratio of eigen vector reaches 99.9%.

$$\sum_{i=1}^{M} \lambda_i \ \geq \ 0.999 \sum_{i=1}^{N} \lambda_i \ . \qquad (4)$$

When L represents the final dimensional number of discriminant space, there is the following relationship between M and class number K.

$$L \leq min(K-1, M). \qquad (5)$$

Consequently, the formula to translate input N-dimension image feature vector into L-dimension discriminant space is

$$\begin{array}{ccccc} L \times 1 & & L \times N & & N \times 1 \\ \boldsymbol{y} & = & [b_1\, b_2\, \ldots\, b_n]^T & & \boldsymbol{x} \end{array} \qquad (6)$$

4.4 Classification of shape

All the following classification processes are automatically processed by the system.

1. Suppose an image is extracted from a target of identification number "id" at the time t by background subtraction[3]. The image is doubled in height and width, and smoothed.

2. A 11-dimensional feature vector of the image ($\boldsymbol{x}_s^{(id,t)}$) is composed ($s$ denotes "shape").

3. $\boldsymbol{x}_s^{(id,t)}$ is projected into the discriminant space.

4. K nearest vectors, $\boldsymbol{V}_{s(k,c)}$ (c denotes category of the vector $\boldsymbol{V}_{s(k)}$), are extracted from pre-learned 2000 sample vectors.

5. Calculate "distance" between $\boldsymbol{x}_s^{(id,t)}$ and each category as follows

$$dist(\boldsymbol{x}_s^{(id,t)}, S_i) \overset{def}{\Longleftrightarrow} \sum_{if\ c=i} \frac{1}{\left\| \boldsymbol{x}_s^{(id,t)} - \boldsymbol{V}_{s(k,c)} \right\|} \qquad (7)$$

($dist(\boldsymbol{x}_s^{(id,t)}, S_i)$: distance between $\boldsymbol{x}_s^{(id,t)}$ and category S_i)

6. Output matched category and save "$Wieght_s$".

$$S_{max} = \max_i(dist(\boldsymbol{x}_s^{(id,t)}, S_i)) \Rightarrow \boldsymbol{x}_s^{(id,t)} \in S_{max} \qquad (8)$$

$$Weight_s(\boldsymbol{x}_s^{(id,t)}) = S_{max} \qquad (9)$$

[3]The id is assigned by the target extraction and tracking module[7] (see **3**).

5 Classification of target color

5.1 Outline

Because human eyes have stability for color, they are able to identify a color of identical targets under various conditions, for example, under direct sunlight, in evening light, or in shade. In this study, although the targets of color classification (estimation) move in any direction at various times under various weather, human eyes are able to identify them stably [4].

Therefore, in this study, the operators attached the color labels based on "their own sense" to the sample colors shot under various conditions, and the system was made to learn them.

However, when an identical target is actually shot under various lighting conditions, in general the color samples spread broadly in the color space as (R,G,B). Namely, the variance of sample data within a class is too large. Therefore, we devided samples of each class into sub-classes to obtain less distorted discriminant space by using LDA. This is one of the key devices of this study.

5.2 Learning of color

The sample color (R,G,B) images of vehicles or humans, shot at various times while sunny (under the sunlight, in the shade) or cloudy, are collected as variety as possible. About 1500 sample colors were segmented from the images of sunny days, and about 1000 samples from the images of cloudy days.

The operators attach the appropriate label {red-orange-yellow, green, blue-light blue, white-silver-gray, dark blue-dark green-black, dark red-dark orange} to these sample colors based on their sense. These labels are used in the decision process of the special targets later on.

From each labeled image, 25 sample data are extracted. These data in the (R,G,B) color space are translated into the (I1,I2,I3) color space[8] and then averaged.

$I1 = \frac{R+G+B}{3.0}$, $I2 = \frac{R-B}{2.0}$, $I3 = \frac{2.0\times G-R-B}{4.0}$

Feature vectors of each sample represents the minimum classes in the LDA process and two discriminant spaces for color classification are composed for sunny and cloudy days.

5.3 Classification of color

The following processes are automatically processed by the system. These are similar to those of the classification of shape.

1. Suppose an image is extracted from a target with with identification number "*id*" at the time "*t*" by background subtraction, and color values of the image is extracted[5].
2. The extracted values in the (R,G,B) color space are translated into the (I1,I2,I3) color space. From the translated values, a 3-dimension feature vector $\boldsymbol{x}_c^{(id,t)}$ is composed.
3. $\boldsymbol{x}_c^{(id,t)}$ is projected into the discriminant space corresponding to the weather during data collection.
4. K nearest vectors $\boldsymbol{V}_{\boldsymbol{c}(k,c)}$ (c denotes category of the vector $\boldsymbol{V}_{\boldsymbol{c}(k)}$) are extracted from pre-learned (1500 (sunny) or 1000 (cloudy)) sample vectors.
5. Calculate "distance" between $\boldsymbol{x}_c^{(id,t)}$ and each color group as follows;

$$dist(\boldsymbol{x}_c^{(id,t)}, C_i) \stackrel{def}{\Longleftrightarrow} \sum_{if\ c=i} \frac{1}{\left\| \boldsymbol{x}_c^{(id,t)} - \boldsymbol{V}_{\boldsymbol{c}(k,c)} \right\|} \tag{10}$$

($dist(\boldsymbol{x}_c^{(id,t)}, C_i)$: distance between $\boldsymbol{x}_c^{(id,t)}$ and color group C_i)

6. Output matched color group and save "$Wieght_c$".

$$C_{max} = \max_i(dist(\boldsymbol{x}_c^{(id,t)}, C_i)) \Rightarrow \boldsymbol{x}_c^{(id,t)} \in C_{max} \tag{11}$$

$$Weight_c(\boldsymbol{x}_c^{(id,t)}) = C_{max} \tag{12}$$

[4]For example, they always identify "the red vehicle" as "the red vehicle" both in sunlight and in shade.

[5]The *id* is assigned by the target extraction and tracking module[7] (see **3.2**).

6 Final decision

6.1 Decision process

When a new target appears in the input image, the target extraction and tracking module (see **3**) assigns an identifical number "id" to it and tracks its movement. The processes that have been mentioned above are conducted for every extracted and tracked image (see Figure1, **4.4**, **5.3**). These classified results are displayed on the input image as shown in Fig.3.

The module for the final decision calculate most matched shape category and color groups by equations (13)-(15) and output "$final\_output(D^{id})$".

$$D_s^{id} = \max_t(Weight_s(\boldsymbol{x}_s^{(id,t)})) \tag{13}$$

$$D_c^{id} = \max_t(Weight_c(\boldsymbol{x}_c^{(id,t)})) \tag{14}$$

$$final\_output(D^{id}) = \{D_s^{id}, D_c^{id}\} \tag{15}$$

6.2 Decision for special target

Among the top k sets of vectors for learning calculated during category classification mentioned in section 3, the weights of vectors with special target labels are counted for each special target. These counts are defined as points for each special target.

The maximum value of points for each special target is obtained among the tracking sequence. The top two point values are picked from the distribution of maximum points for each special target obtained in this way.

The color database of the special target with maximum points is compared to the final decision for the color of the previous target obtained. If the points distribution of each color in the final decision accords with the database, the target is recognized as the special target, and that is considered as the result of the decision.

If it disagrees with the database, the target with the second-highest point value is compared with the final decision for color of the target obtained in the same way. If it agrees with the database, the target with the second-highest point value is recognized as the special target. If it disagrees with the second database, it is determined to be a non-special target.

Labels with the maximum point value for special targets for the displayed image are displayed on the system monitor. In addition, font colors on the system monitor are changed according to the value of the points.

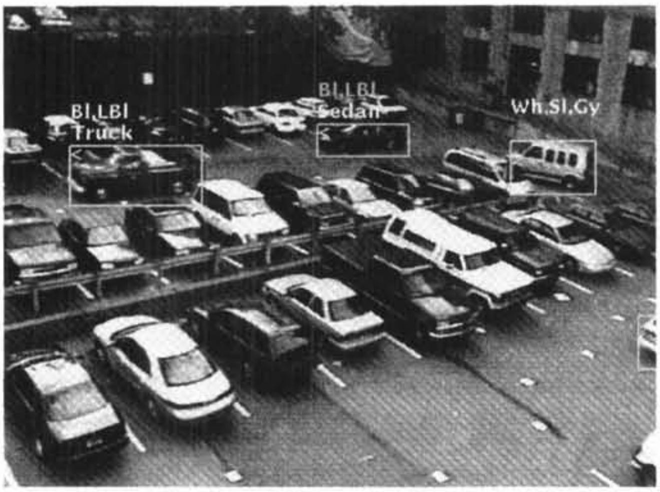


Figure 4: Sample of classified results.

7 Experiments and Results

7.1 Sample of processing display

Figure 4 shows a sample of input image and classified results. On the image, the $final\_output(D^{id})$ of each target are displayed, that were classified categories and colors.

The results of color and category classification displayed on the image are represented by the following codes.

```
<Color Codes>
  Rd.Or.Ye:  red-orange-yellow
  Gr      :  green
  Bl.LBl  :  blue-lightblue
  Wh.Sl.Gy:  white-silver-gray
  Db.Dg.B :  darkblue-darkgray-black,
  Dr.Do   :  darkred-darkorange

<Category Codes>
  Sedan : sedan,  Van    : van
  Track : truck,  Mule   : carts for workers
  Human : human,  Others: others
```

Font colors of the classified results shown in the display are automatically determined according to the value of $final\_output(D^{id})$ (reliability) as follows; "(high) red > green > white > not displayed (low)", in order to help the operator's understanding. As for these font colors, we substituted white for red and gray for green in Figure 4. The threshold of font color change was determined from the experimental results.

The processing speed at classification including target segmentation was two or three cycles per second where the input image contained only one target to process (PentiumIII 450MHz), the speed is dependent on the size of extracted images. A computer with a more powerful CPU is able to speed up the processing ability in principle, because the processing speed is merely dependent on the computer processing ability.

In Fig.4, the classification of category was made by the following methods. Discriminant space for Area A in Fig.1 was applied to the truck on the left and the sedan in the center of the display, and discriminant space for Area B was applied to the minivan on the right of the display. In the bottom right of the display, there was a vehicle with only the rear in sight, but it was cut off at the edge of the display, so it was not considered a target to process.

The truck on the left in Fig.4 was moving at a certain speed, so the image was blurred. Such blurs that were also found in sample images for learning were learned "as is."

In category classification, results for the truck and the sedan were shown using red fonts (white in Fig.4), because their whole shapes were clearly in sight. On the other hand, it is supposed that the result of the minivan on the right of the display was not displayed because the front most characteristic of that vehicle was hidden behind the parking car and the value of D_s^{id} became low.

In color estimation, it was considered that enough stable data were obtained for the truck and the minivan. Consequently, adequate results were shown using red fonts(white in Fig.3). On the other hand, the result for the sedan in the center was shown using green fonts(gray in Fig.4). The reason for this was that color information from the sedan became unstable because of the reflection of different colored vehicles in its body.

Table 1 : Experimental results.

	H	S	V	T	M	O	Totl	Err	%
H	67	0	0	0	0	7	74	7	91
S	0	33	2	0	0	0	35	2	94
V	0	1	24	0	0	0	25	1	96
T	0	2	1	12	0	0	15	3	80
M	0	0	0	0	15	1	16	1	94
O	0	2	0	0	0	13	15	2	87
								Avg	*90*

H:Human, S:Sedan, V:Van, T:Truck, M:Mule, O:Others

7.2 Experimental results

Table 1 shows the results of the experiments. The experiments were conducted with the following procedures. Considering both weather and solar location, images from the camera were input to the system online between 9:00 a.m. and 10:00 a.m., 1:00 p.m. and 2:00 p.m., 4:00 p.m. and 5:00 p.m. on sunny or cloudy days. Any input images were not included in the sample images for learning. The number of targets to process that appeared in the images amounted to 180.

In this experiment, the results for the final category and the color decision by the system were counted as a correct answer only when both the data corresponded with the results of human sight. As shown in the Table 1, the average of all correct answer rates was about 90%. Between sunny and cloudy days, the results on cloudy days were slightly better because the images were more stable on cloudy days.

For each area, the images segmented from Area A in Figure 1 were classified better than those from the other areas. The direction of the targets had little effect on the classification results.

The reason why the results for trucks were slightly lower than those for other kinds of vehicles is that although the poor number of samples for learning was collected because of the lower traffic of the trucks, there were various "appearances" (shapes) of trucks as shown in Figure 2. Sufficient samples for learning may improve the correct answer rate.

There were some other cases where the system had failed in category classification and color estimation as follows. {*Category classification:* The case that the targets overlapped each other, the case that the targets without samples for learning appeared (crane cars etc.), and so on. / *Color classification:* The body reflected strong sunlight, and that part was classified as being white, and so on.}

In this system, for the classification of overlapped target images, many images of overlapped targets were collected in advance and labeled as "Other" category, and then added as sample for learning. However, misclassifications often arose because of the various combinations of overlapping images. On the other hand, in case tracking for each target bore good results, some could be finally excluded as "temporal misclassification by overlapping," because the $final_output(D^{id})$s of each target become low in their tracking sequence.

Figure 5: Special target detection

7.3 Detection of special target

Table 2 shows the experimental results for the detection of special targets, and Figure 5 shows the display samples of detection results. We recorded the input pictures on the delivery time to videotape for a few days and conducted the experiment because both FedEX van and UPS van only appeared for delivery on morning and afternoon. The number of targets to process was 100 including special targets to detect.

In Table 2, the Detection Rate represents the number of times that the system detected the special target correctly (the number of correct detection / the number of special targets), and False Alarm represents the number that the system mis-detected a non-special target as a special target. The images for the test were not among the sample images for learning.

The False Alarm of Police Car was a relatively high value; on the other hand, the Detection Rate showed good results. The reason for this is that the tuning preferring the Detection Rate, which is important in the sight of security and monitoring, was selected.

One of the causes of the False Alarm connected with the UPS van detection was misdetection of the CMU deep wine-red van as the smaller type of UPS van, which had two types of vehicles as shown in Figure 3. The cause of the False Alarm in Police Car (police patrol car) detection was the misdetection of other large white or silver sedans because the police cars were also large white sedans as shown in Figure 3.

Table 2 : Special target detection.

	Detection rate (%)	False alarm (%)
FedEx Van	11/12 (91.0%)	3/88 (3.4%)
UPS Van	12/13 (92.3%)	7/87 (8.0%)
Police Car	16/17 (94.1%)	12/83 (14.5%)

Additionally in this test, the functions of category classification and color estimation were applied to the detection of the special target that had not been learned in advance. The taxies in the neighborhood of the test site were all yellow sedans. Thus, using the results of category classification and color estimation, we made the system output red-orange-yellow sedans as possible taxies. Consequently, the Detection Rate became 95.2%.

8 Conclusion

In this paper, we described a vision system for target classification, color estimation, and detection of special targets from street video images. The method of category (shape) classification in this study could also be applied to the images of noctovision (infrared camera) in principle.

With weather conditions other than sunny or cloudy days, especially rainy or snowy days, category and color classification of the targets are considered to be realized if the targets can be segmented stably, without the effects of raindrops or snowflakes drifted into the input images.

Acknowledgement: We would like to thank the members of CMU VSAM Project, R.Collins, A.Lipton, D.Duggins, H.Fujiyoshi and N.Enomoto, for their profitable comments.

References

[1] Koller, D., et al.: Towords Robust Automatic Traffic Scene Analysis in Real-time, *Proc. ICPR94*, pp.126-131, (1994)

[2] Beymer, D., et al.: A Real-time Computer Vision System for Measuring Traffic Parameters, *Proc. CVPR97*, pp.495-501, (1997)

[3] Mantri, S., Bullock, D. and Garret, J.Jr. : Vehicle Detection Using a Hardware-Implemented Nueral Net, *IEEE Expert*, pp.15-21, (Jan.-Feb., 1997)

[4] Ikeda, T., et al.: Traffic Measurement with a Roadside Vision System - Individual Tracking of Overlapped Vehicles, *Proc. ICPR96*, pp.859-864, (1996)

[5] Jolly, M-P D., et al.:Vehicle Segmentation and Classification Using Deformable Templates, *IEEE Trans. PAMI*, Vol.18, No.3, pp.293-308, (1996)

[6] Kanade, T., et al.: System for Video Surveillance and Monitoring, *Proc. DARPA IUW*, pp.3-24, (1998)

[7] Lipton, A., et al.: Moving target classification and tracking from real-time video, *Proc. WACV98*, pp.8-14, (1998)

[8] Ohta, Y. and Kanade, T. and Sakai, T. : Color Information for Region Segmentation, *Computer Graphics and Image Processing*, Vol.13, No.3, pp.222-241, (1980)