Moving Objects Segmentation at a Traffic Junction from Vehicular Vision

Joo Kooi Tan[1], **Seiji Ishikawa**[2], **Shinichiro Sonoda**[3],
Makoto Miyoshi[4], and **Takashi Morie**[5], Non-members

ABSTRACT

Automatic extraction/segmentation and the recognition of moving objects on a road environment is often problematic. This is especially the case when cameras are mounted on a moving vehicle (for vehicular vision), yet this remains a critical task in vision based safety transportation. The essential problem is twofold: extracting the foreground from the moving background, and separating and recognizing pedestrians from other moving objects such as cars that appear in the foreground.

The challenge of our proposed technique is to use a single mobile camera for separating the foreground from the background, and to recognize pedestrians and other objects from vehicular vision in order to achieve a low cost and intelligent driver assistance system.

In this paper, the normal distribution is employed for modelling pixel gray values. The proposed technique separates the foreground from the background by comparing the pixel gray values of an input image with the normal distribution model of the pixel. The model is renewed after the separation to give a new background model for the next image. The renewal strategy changes depending on if the concerned pixel is in the background or on the foreground. Performance of the present technique was examined by real world vehicle videos captured at a junction when a car turns left or right and satisfactory results were obtained.

Keywords: Background Detection, Foreground Extraction, Pedestrians, Normal Distribution, Mobile Camera, vehicular Vision

1. INTRODUCTION

Automobile technology has come to the stage of realizing safe driving by employing various sensors to prevent car accidents. These systems are referred to as Advanced Driver Assistance Systems (ADAS) [1] or Intelligent Transportation System (ITS) [2]. A vehicle can gather image information from a variety of sources. These include laser, radar, video (a single camera, stereo cameras, OMNI cameras), etc. They can be employed for assisting a driver in recognizing traffic signs, pedestrians, road lanes or parking lanes in order to make appropriate decisions.

Over the last decade, many researchers have addressed on-board pedestrian and driving car detection to anticipate accidents in order to avoid them. Fardi et al. [3] combine laser scanner with a Thermal Infrared Radiometer (TIR) to extract shape of an object by using Kalman filters in a parallel way. Nanda et al. [4] use infrared images to obtain correlation with probabilistic human templates. Detecting an object shape including human detection has as well been enthusiastically studied in the computer vision field [5]. Many researchers use stereo cameras to obtain a depth map of a frontal view from which object regions are detected [6]. Although they achieved reasonable performance with a video captured from a car driving in a normal speed, they have thus far detected only human locations and not detailed shapes. To extract the exact shape of a human for purposes of recognition, the disparity between the images should be precisely calculated in order to alert the driver so that he/she may pay attention to only a high risk situations such as pedestrians who is “walking in front of the car”, “absorbed in the mail/talk of his/her cellular phone”, “ has unexpectedly fallen-down”, etc.

The use of a single camera helps to avoid disparity calculation. Some of the techniques use feature-based techniques for detecting pedestrians; e.g., edge template of Gavrila et al. [7,8,9], HOG (Histogram of Orientated Gradient) descriptor [10], the motion enhanced Haar feature [11], human templates [12], sparse Gabor filters and support vector machine (SVM) [13]. Unlike these direct methods, Zhang et al. [14] proposed a method of using optical flow to estimate a FOE (focus of expansion), and make a FOE residual map extraction. The researchers then created a FOE residual map segmentation for extracting a human under the condition that vehicle goes straight. But the technique is not very accurate when a car moves slowly, because small changes of flows make it difficult to estimate and obtain the FOE. Thus, in their paper, the speed of the vehicle is between 30- 40miles/hour (approx 48-64km/h).

Manuscript received on March 18, 2011 ; revised on September 13, 2011.

[1,2,3,4] The authors are with Department of Mechanical and Control Engineering, Kyushu Institute of Technology, Japan, Email: etheltan@cntl.kyutech.ac.jp, ishikawa@cntl.kyutech.ac.jp and

[5] The author is with School of Brain Science and Engineering, Kyushu Institute of Technology, Japan, Email:morie@brain.kyutech.ac.jp

On the other hand, some recent approaches use landmarks such as a pedestrian crossing (a zebra crossing) [15] and a road lane [16] to estimate obstacles on a road. Both detectors are a feature point based method which assume that the location of the pedestrian crossing or a zebra crossing is known in advance. The former paper uses the MSER (Maximally Stable Extremal Region) detector as a feature detector to detect border points of stable regions and yields a few hundred MSER regions. It then construct all possible triplets from the regions to obtain zebra crossing stripes. However, the latter method requires stereo image points to extract white lanes points on a straight and a single road. However, it is difficult to extract good features from a roadway, since white lanes or zebra crossings are usually not clear and discontinuous.

Background detection is a popular technique for extracting foreground objects based on a single camera. The background is sequentially detected from a video image sequence that may contain moving foreground objects. If a background image is obtained, human shape can be directly extracted by comparing the background image and the present image. There are reported techniques for sequential background image detection [17-20]. All of these methods, however, deal with the video images captured from a stationary camera. A background detection technique based on a pan-tilt camera is reported by Hayman & Eklundh [21]. But it made use of control signals from the pan-tilt camera to detect the background, which is not useful for a camera on a car. Jota et al. [22] proposed a method of extracting moving objects from a video provided by a moving camera by selecting the trajectories of background points tracked over a number of frames. But the method is difficult to be employed in car vision, since it is computationally heavy, particularly, in the part of a generated panoramic image of the background where a sequence of all images is required in the process.

Thus, extraction of a background image to obtain foreground objects and retrieve only valid moving objects by a moving camera is still an important and difficult problem. To the best of our knowledge, a practical sequential background detection method has not yet been proposed with respect to a video captured from a moving camera.

However, when considering an object extraction problem, some previous techniques [23-24] perform object segmentation that segment an object boundary into one dimensional curve. An Interactive Graph Cuts [25] technique has been proposed to segment an interested object. However the technique needs to prepare for a correct label and a graph of the interested object in advance. For segmenting pedestrians, some techniques use global shape models, e.g., separating shape models into hierarchical parts and template matching is done combined with background subtraction [26]. Obviously, the technique needs to define the interest parts (seeds) of an object, hence high computing cost.

The present paper consists of two parts. First, we propose a technique for separating the foreground, *i.e.*, moving objects, from the background in a video captured by a moving camera. Second, we propose a method of segmenting moving objects from the acquired foreground and judge if they are cars or pedestrians. The diagram of the proposed technique is shown in ***Figure 1***. In the first part, the gray value of each pixel in the background image is represented by the normal distribution. Once the background model is provided, moving objects are directly extracted by the pixelwise comparison of the gray values between the fed image and the background image. The next idea of the technique is to calculate camera motion between the previous frame and the incoming/input frame, and to modify the background image frame by frame. Due to the fact that a moving camera is applied in the technique, camera motion detection and 2-D projective transform between two successive images are performed by employing feature points tracking. Furthermore, frame compensation by bilinear interpolation is done to the camera images to get a background model for achieving more precise model estimation. Then we focus our attention on segmenting and clustering moving objects (e.g., pedestrians, cars, bicycles, etc.) which appear on the foreground, in order to give warnings to the driver to let him/her perform braking or slowing down the speed.

The technique is applied to the video captured near a zebra crossing at a traffic junction in order to extract and recognize pedestrians and other vehicles. Experimental results on foreground extraction are shown along with the comparison of accuracy and computation time with Temporal median, Running average, Mixture of Gaussian (MoG) methods. Furthermore, to show the accuracy in the segmentation of moving objects from the foreground, two vehicular videos captured in an urban area are used: (1) a vehicle moves slowly at a junction and turns left, and (2) it turns right at a junction.

2. OVERVIEW OF THE PROPOSED METHOD

A pixel on a video image frame changes its gray value by camera motion or by appearance of an object in the time lapse, as well as random image noise. In this dynamic situation, camera motion is detected from a video by finding pixel to pixel correspondence between successive frames. Then the background image is successively renewed, starting from the initial background image which is equal to the initial image frame of the video. The renewal is done by modifying the normal distribution defined on each pixel on an image frame and representing its gray value.

The initial background model of the first image

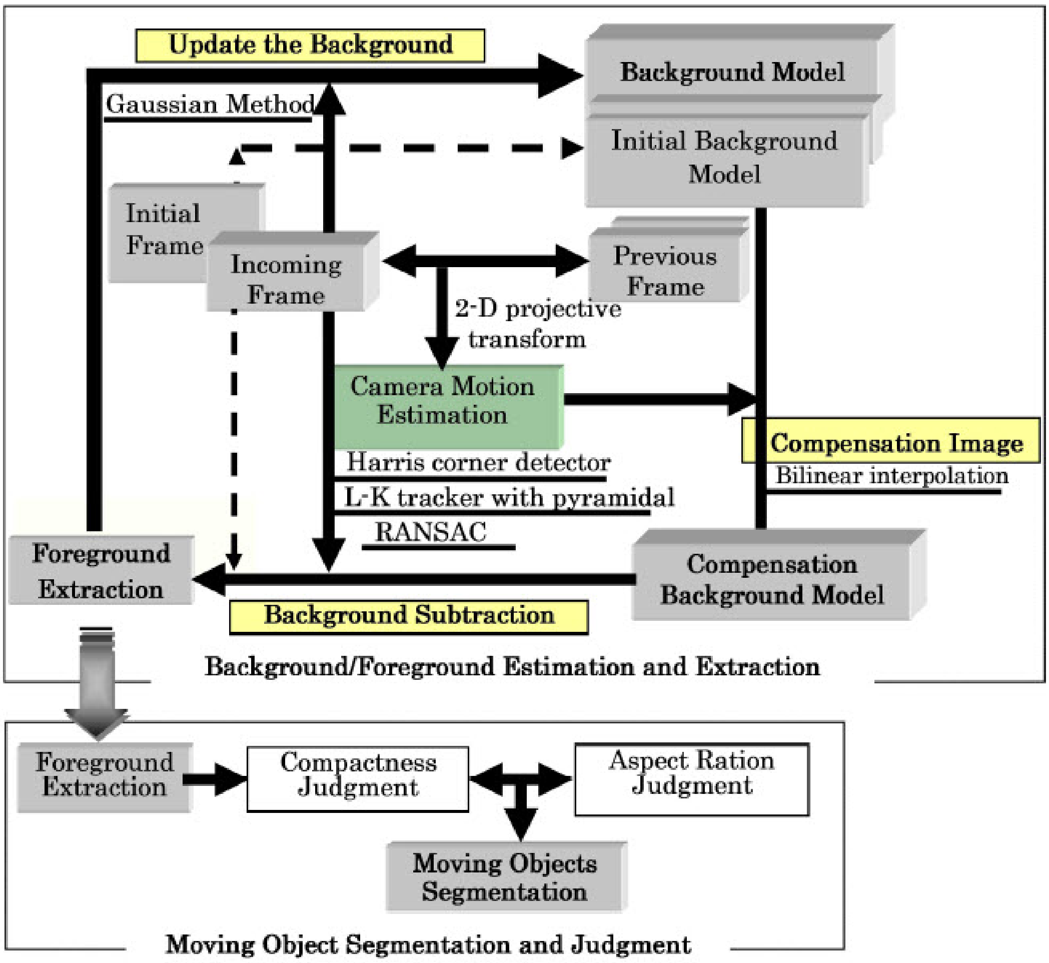


Fig.1: *Diagram of the Proposed Technique.*

frame is formed by an initial normal distribution

$N_0(\mu_0, \sigma_0)$ in which the initial standard deviation σ_0 is between 15 and 30 (the authors defined the value $\sigma_0 = 30$ in ***Figure 5***). This value is obtained experimentally with respect to our vehicle video database. The mean value $\mu_0(x.y)$ is assumed to be the initial gray value at each pixel. Here, if moving objects are on the initial image, then the moving objects become part of the background. Now, suppose that normal distributions are defined on the initial image frame. The normal distribution of the pixel $p(x, y)$ on the ith image frame $(i = 1, 2, \ldots, I)$ is calculated by bilinear interpolation of the mean and the variance from those normal distributions of the four pixels on the (i-1)th image nearest to the point which corresponds to the pixel $p(x, y)$ on the ith image. Referring to the normal distribution, the pixel $p(x, y)$ is judged if it belongs to the foreground or to the background. This is done with all the pixels on the ith image frame. Then we have a set of pixels on the foreground on the ith image frame and it is reported at time i. The normal distribution at every pixel on the ith image frame is renewed their parameters according to if it belongs to the foreground or the background. Examples of the normal distribution on pixels along with time are illustrated in ***Figure 2***. As shown in $Fig.2(b1)$ and $(b2)$, the input pixel (green line) matches with the normal distribution. This means that the input pixel p(160,100) in a red square on the original image frame 1 and frame 2 are judged as belonging to the background. However the input pixel p(50,60) on image frame 1 and frame 2 shown in $Fig.2(d1)$ and $(d2)$ do not match the background normal distribution. Thus the input pixel p(50,60) on frame 1 and frame 2 are judged as belonging to the foreground. Since the preceding car which is captured by the ego-car is moving, part of the car appears initially on the foreground.

The point on the (i-1)th image correspondent to the pixel p(x,y) on the ith image is calculated using the inverse projective transform $T_{i \to i-1}$ explained in the next section.

3. DETECTING CAMERA MOTION

In order to detect camera motion, feature points are chosen on image frame f_{i-1} by Harris corner detector [28] and their corresponding locations are searched on the image frame f_i employing Lucas-

Kanade tracker [29] with pyramidal search, taking larger displacement of the feature points at faster camera motion into account. From the set of point pairs $\{(p_{k,i-1}, p_{a(k),i})\}$, the 2-D projective transform $T_{i-1\to i}$ is defined by

$$\begin{pmatrix} x \\ y \\ 1 \end{pmatrix}_i = m \begin{pmatrix} h_0 & h_1 & h_2 \\ h_3 & h_4 & h_5 \\ h_6 & h_7 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ 1 \end{pmatrix}_{i-1} \quad (1)$$

between the images f_{i-1} and f_i. Here point $p_{k,i-1}(x_k, y_k)$ and $p_{a(k),i}(x_{a(k)}, y_{a(k)})$ corresponds with each other: Function $a(k)$ makes the correspondence between the two matched points. Real value m is a value which makes the equation hold correctly.

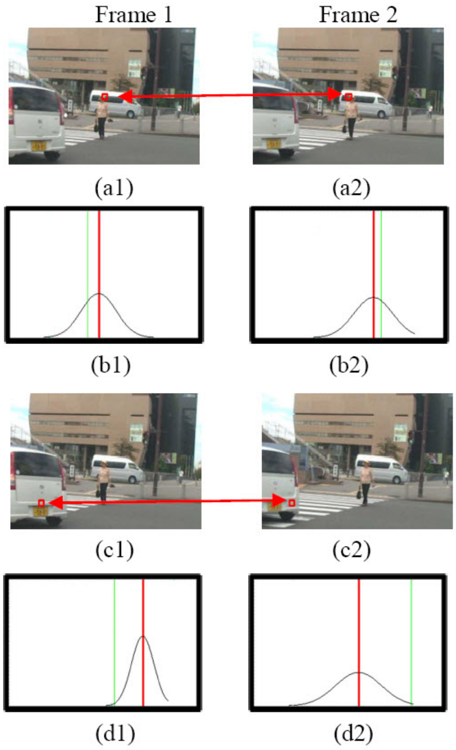


Fig.2: *The Normal Distribution on Pixels of the Frames. (a)and(c) The Original Frame (The Red Square is the Position of the Pixel and Red Arrow Stands for Pixel Correspondence (i.e. i-th images)), (b)and(d) The Normal Distribution on Pixel p(160,100) and Pixel p(50,60) where (b1)(d1) are on the Image Frame i=1 and (b2)(d2) are on the Image Frame i=2.(The red line stands for mean value and the green line shows the pixel value, respectively.)*

The camera motion is assumed to be given by this transform between the image frame f_{i-1} and f_i. Although Eq.(1) holds between two planes in a 3-D space, it is employed for camera motion detection in the proposed technique in order to reduce computational cost. Uemura et al.[32], proposed a dominant planes estimation technique from an image sequence to solve this problem. They obtained good accuracy of camera motion compensation, but it had a high processing cost and did not work effectively as it was affected by low texture scene such as buildings or walls, ground with similar texture, and color conditions.

On the other hand, in the calculation of the 2-D projective transform, foreground objects can be included in the initial image frame f_0. Their motion vectors act as outliers there. But they are effectively excluded by use of RANSAC [27], if the number of motion vectors on the foreground objects is much less than those in the background.

Here we show the effect of camera motion compensation using the 2-D projective transform by the employment of six video scenes (1,231 image frames). After the compensation, subtraction between two successive images is performed and binarization is done to the subtracted images for the evaluation. The removal rate is defined by

$$\text{Removalrate}[\%] = \left(1.0 - \frac{Comp\_af}{Comp\_bf}\right) \times 100 \quad (2)$$

Here, $Comp\_bf$ means the total pixels appeared on the background, and $Comp\_af$ means the total pixels appeared on the background after the camera compensation.

The result is given in ***Table 1***, in which the compensation employing parallel translation is also shown for reference.

The original six video scenes are shown in ***Fig.3***, in which the scenes are acquired in different driving directions such as moving straight (MS), turning right (TR), turning left (TL) and turning round a small curve (TC) at various places and different weather conditions. Table 1 shows that 2-D projective transform compensates the camera motion enough to be adopted in the proposed technique. It also shows that the parallel translation gives a poorer result than the 2-D projective transform.

4. JUDGMENT OF THE FOREGROUND

Let the gray value and the normal distribution at the pixel $p_{i(x,y)}$ be denoted by $f_i(x,y)$ and $N_i(\mu,\sigma)$, respectively. If, for a certain threshold T_σ, the following relation

$$\frac{|f_i(x,y) - \mu|}{\sigma} \le T_\sigma \quad (3)$$

holds, the pixel $p_i(x,y)$ is recognized as a pixel in the background: Otherwise it is recognized as a pixel on the foreground. This judgment is performed with all the pixels in the image. Hence a set of foreground pixels, and therefore foreground objects, in the image f_i are obtained. This is the output at time i.

5. RENEWAL OF THE NORMAL DISTRIBUTION

Based on the video taken from a fixed camera, various techniques [17-19] have been proposed to make a background model. However, if one wants to apply the technique to detecting the background from a video taken by a moving camera, camera motion compensation should be considered in the algorithm. At the same time, it should be done in high speed with limited noise. Therefore, we employed a background modelling based on a single normal distribution and a renewal strategy explained in the following, because it is effective and requires less computation cost than the MoG(Mixture of Gaussian)[17] method. In order to verify the proposed technique, comparison with other methods including the MoG model is done experimentally, which is shown in Section 8.

Let the normal distribution of pixel $p_i(x,y)$ be denoted by $N_i(\mu_i(x,y), \sigma(x,y))$. It is renewed according to whether the pixel $p_i(x,y)$ exists in the background or on the foreground. The average value $\mu_i(x,y)$ and the variance $\sigma_i^2(x,y)$ are renewed, respectively, by

$$\mu_i^{\#}(x,y) = \begin{cases} \alpha f_i(x,y) + (1-\alpha)\mu_i(x,y) \\ \qquad \text{if. } p_i(x,y); background \\ (1-\beta) f_i(x,y) + \beta\mu_i(x,y) \\ \qquad \text{if. } p_i(x,y); foreground \end{cases} \tag{4}$$

$$\sigma_i^{2\#}(x,y) = \begin{cases} \alpha(f_i(x,y)-\mu_i(x+y))^2+(1-\alpha)\sigma_i^2(x,y) \\ \qquad \text{if. } p_i(x,y); background \\ (1-\beta)(f_i(x,y)-\mu_i(x+y))^2+\beta\sigma_i^2(x,y) \\ \qquad \text{if. } p_i(x,y); foreground \end{cases} \tag{5}$$

Here α and β are defined, respectively, as follows;

$$\alpha \equiv \alpha_i(x,y) = \frac{c}{\sqrt{2\pi\sigma}} \exp\left\{-\frac{1}{2}\frac{(f_i(x,y)-\mu_{i-1}(x,y))^2}{\sigma_{i-1}(x,y)^2}\right\} \tag{6}$$

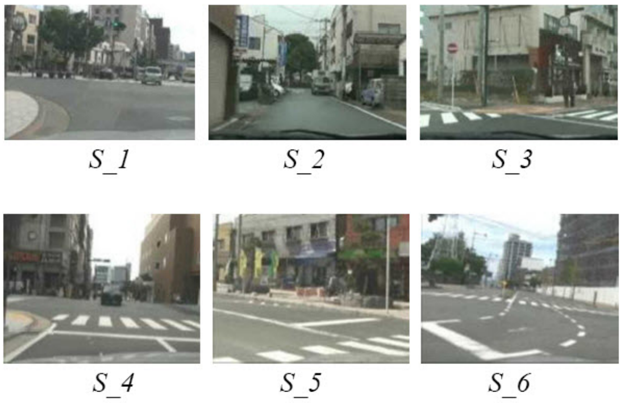
$$\beta \equiv \beta_i(x,y) = \frac{1}{1+kC_i^2(x,y)} \tag{7}$$

Parameter α is called a variable learning rate. Constant c contributes to sensitivity of the background model change to the input pixel value and is determined experimentally. On the other hand, parameter $\beta(0 < \beta \leq 1)$ defined by Eq.(7) is another variable learning rate. $C_i(x,y)$ is the number of successive frames where the pixel $p(x,y)$ has been judged as the foreground (k is a constant). Once the pixel $p(x,y)$ is judged as the background, $C_i(x,y)$ is reset to 0. The reason of employing $C_i^2(x,y)$ instead of $C_i(x,y)$ in Eq.(7) is to accelerate the effect of a long foreground period.

In effect, by considering the two variable learning rate α and β, the former for the background and the later for the foreground, the proposed normal distribution model tries to adapt itself to small change or disturbance in the background, e.g., swaying trees, shadows in a scene, etc., whereas it adapts to sudden change in the scene such as changes of lights or weather and longer change caused by an object on the foreground. Therefore the one normal distribution adopted in the proposed technique adapts to the input image robustly than employing MoG[17]. The results of the comparison are shown in the experiment section.

Table 1: *Average Removal Rate with respect to the Camera Motion Compensation method. The averaging is done with every video scene.*

	Average Removal Rate[%]		
Video	*Num.of*	*Parallel*	*2-D projective*
Scene	*frames*	*Translation*	*Transform*
1	329	76.12	82.60
2	160	17.98	76.18
3	201	93.76	96.15
4	165	83.56	91.32
5	196	57.87	75.68
6	180	68.94	80.41

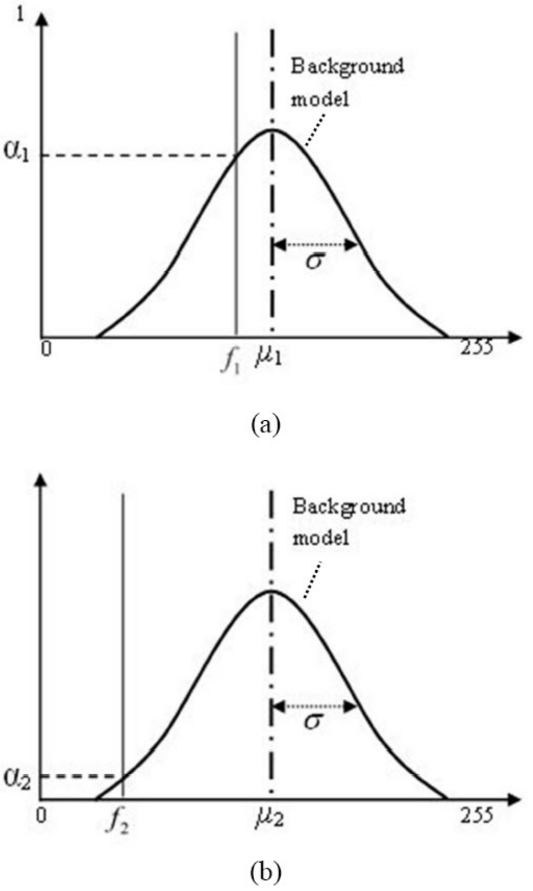

S_1 S_2 S_3

S_4 S_5 S_6

Fig.3: *Scenes Compared for Evaluating the Camera Motion Compensation.*

$(S\_1: MS \to TC \to TL \to MS, S\_2: MS \to MS, S\_3: MS \to TR, S\_4: MS \to TL, S\_5: MS \to TL, S\_6: MS \to TR)$

Figure 4 illustrates an example of the normal distribution background model. In $Fig.4(a)$, the input pixel f_i matches the background model as f_i is within the standard deviation from the average value. Then the learning rate 1 α_1 becomes large with respect to the distribution. In $Fig.4(b)$, on the other hand, the input pixel is outside of the standard deviation with the distribution. It is then regarded as not matching the background and the learning rate α_2 becomes small. Thus the background pixel is judged as a foreground pixel. The normal distribution model of the background and the foreground is renewed by Eqs. (4) and (5).



(a)

(b)

Fig.4: *Strategy of Match Level with the Background Corresponding to the Normal Distribution Model of an Observed Pixel. (a) Matched with the Distribution, (b) Unmatched with the Distribution)*

6. OVERALL ALGORITHM

The overall algorithm is given in the following. The number of processed image frames is I.

0: i:=1

1: Each pixel on the initial image frame f_0 is given a normal distribution model $N_0(\mu, \sigma)$ with its RGB values.

2: Harris corner detector is applied to the image f_{i-1} in order to detect feature points $p_k(k = 1, 2, \dots, K)$ on the image.

3: The corresponding locations of the feature points $p_k(k = 1, 2, \dots, K)$ are searched on the next image f_i by applying Lucas-Kanade tracker. Pyramidal image structure is employed in the search.

4: The 2-D projective transform $T_{i-1 \to i}$ between the images f_{i-1} and f_i is defined from the obtained corresponding feature point pairs. Here RANSAC is employed to exclude outliers.

5: The normal distribution $N_i(\mu, \sigma)$ is calculated with respect to every pixel on image f_i from the normal distributions in image f_{i-1}.

6: Employing $N_i(\mu, \sigma)$, every pixel on image f_i is examined if it is in the background or on the foreground. All the pixels judged as foreground pixels are reported at time i.

7: The normal distribution with the pixel in image f_i is redefined by Eqs.(4) and (5) to get a renewed normal distribution $N_i^{\#}(\mu, \sigma)$.

8: Let $N_i(\mu, \sigma) \equiv N_i^{\#}(\mu, \sigma)$.

9: Set $i := i + 1$.

10: If $i > I$, stop. Otherwise, go to 2.

7. SEGMENTATION OF MOVING OBJECTS

Our goal in the present system is to detect moving objects from the background and, in particular, to extract the figures of pedestrians. As for identifying a human detection from an image, the use of the HOG feature [10] is a well-known technique. Our system, however, employs a simple but effective rule for judging if a moving object on a road is a human, since the overall computation time needs to be conducted in real-time in a driver assistance system.

The employed rule evaluates two indices related to the shape of an extracted object. *Complexity* C of shape S, whose area and perimeter are denoted by A and L, respectively, is defined by

$$C = L^2/4\pi A, \tag{8}$$

where as *aspect ratio* R of shape S, whose width and height are denoted by W and H, respectively, is defined by

$$R = H/W. \tag{9}$$

Shape S is judged as a pedestrian, if the complexity and the aspect ratio are larger than certain thresholds; i.e., the condition for shape S to be a pedestrian is

$$(C \geq Th1) \cap (R \geq Th2). \tag{10}$$

Here thresholds $Th1$ and $Th2$ are determined experimentally. Consequently, applying Eq(10) can simply discriminate a pedestrian from other moving objects such as cars. Here, we consider that the shape of a pedestrian is more complex than a car. On the other hand, a car is more similar to a circle and C is rather closer to 1. Similarly, its aspect ratio R is considered to be small, because generally a car is wider than a pedestrian. However, a more effective idea needs to be developed.

One might think of a shape of a pedestrian who had fallen down on the ground as a possible case

$$(C \geq Th1) \cap (R < Th2) \tag{11}$$

may happen. Since each image frame in a video is processed, we may have a sequence such as (10)→(10) → ... →(10)→(11). Having obtained this sequence, we can judge that a pedestrian has fallen down on the ground. For a person lying long time on the zebracrossing, for example, a sequence (11)→(11)→ ... →(11) would be obtained. A technique for 3-D static obstacles detection in front of an ego-car [31] can be applied in such a case. In this paper, we exclude the case, however.

8. EXPERIMENTAL RESULTS

The proposed technique was applied to the video image data captured from a car and it was compared to the methods employing temporal median and running average.

In the first experiment, the technique is applied to a video of 320×220 pixel images in which a car turns left at a junction with a speed of around 9-16miles/hour (15- 25km/h) under the following situation; goes slowly → stops → turns to the left → moves forward. The video includes total of 98 frame images. In the video, there is another car in front of the ego-car and it soon disappears at the initial part of the video: A pedestrian walks from the right-side to the left-side on the zebra crossing and the ego car stops in front of the zebra crossing. The result of the comparison of the proposed technique with other techniques on the foreground (a pedestrian) extraction is shown in ***Fig. 5***. To allow fair comparison, we make other methods work on the same set of original images, in which six image streams are shown from the left to the right: (a) is the original video image, (b) is the result obtained from the temporal median (TM) method, (c) is the result obtained from the running average (RA) method, (d) is the result obtained from the proposed technique employing the normal distribution model (NM), and finally in (e) we apply an expansion-contraction operator to the obtained foreground images in (d) to remove isolated pixels. It is obvious that the proposed technique exceeds the other two methods from the point of the amount of the noise remained in the background.

Superiority of the present technique over temporal median and running average methods is also shown by the ROC curves depicted in ***Fig. 6***. In the figure, the abscissa is a false positive rate (FPR) [%] and the ordinate is a true positive rate (TPR) [%]. It is shown that the ROC curve of the proposed technique, illustrated by a green line, is nearer to the top left corner than other methods (temporal median: red, running average: blue) when FPR<3.5%.

The effect of the renewal process described in Section 5 compared with MoG[17] is shown in ***Fig. 7***, where (a) Original images, (b) Proposed technique; a background and a foreground normal distribution model, (c) MoG technique using a single Gaussian distribution, (d) MoG technique using three Gaussian distributions. The video scene has a scenario that a car goes slowly near to the junction (Frame_5) and then stops for a while when a lady crosses the zebra-crossing (Frame_15 to Frame_75) and moves again after the lady crossed the road (Frame_75 to Frame_85, till Frame_98). The original video in Fig. 5 as abovementioned contains a frontal car in some of the initial image frames as shown in frame_5 in *Fig. 7(a)*. Therefore the initial background contains it as shown in *Fig. 7(b)*. Although the car goes away and disappears in frame_15, the background still contains some of its part. Finally it disappears from the background in frame_25. The time for the frontal car disappeared from the background was 0.43 s (=13/30) from the time when it appeared in the view. After all, the background update of the proposed technique operates well and robustly even under small (e.g., ego-car moves slowly) changes or sudden (e.g., frontal car or moving objects disappeared) changes. However, the compared technique MoG[17] in (c) and (d) shows that renewal of the background model does not run well even when the ego-car stops at the zebra-crossing(Frame_25 to Frame_75). This is because update speed/sensitivity of the background model are not suited to complex background change and it accompanies a large delay in the complex background modeling. Thus, it is visibly shown in (d) that, when three Gaussian distributions are introduced, the foreground pixels do not have enough time to assimilate to background pixels. Moreover, the MoG technique considers only a background learning rate, but the proposed technique contains two significant parameters; *i.e.*, the background and the foreground learning rate, in order to consider a sudden small or large change of the background as well as the foreground. We furthermore evaluated the computation time of the proposed method and the MoG method in ***Table.2***. Although the computation time of the MoG (with a single Gaussian distribution) technique is approximately 10ms/frame less than our technique, it does not give acceptable results.

The result of objects segmentation with the first experiment is shown in ***Fig. 8***, in which a red rectangle indicates a segmented pedestrian, and a green line shows its trajectory. Note that a human tracking technique is not discussed in this paper. The proposed technique successfully segmented an object from the background on frame $f = 23$ to $f = 98$. However we also obtained some false results as shown in ***Fig. 9***. The lady in the video is carrying and shaking an umbrella. Especially, when it is shaken backward, the umbrella and the human are segmented into two moving objects. Adversely, an occlusion of the lady happens when the umbrella is shaken forward or near her chest. The both cases may affect the precision and the recall defined in Eq.(12) and Eq.(13), respectively, negatively.

In the second experiment, the proposed technique is applied to the video image in which a car turns right at a junction. The video is taken from the car which is stopping in the intersection until all the cars in the facing lane have passed and every pedestrian has finished walking on the right-hand side zebra crossing. The difficulty in the right-turn case is that one needs to segment not only the pedestrians on the zebra crossing but also the facing cars having various sizes caused by perspective view.

The result of the moving object extraction by the proposed technique is given in ***Fig. 10***. *Figure 10(a)* shows some frames in the original video, *Fig.10(b)* gives the extracted foreground objects and *Fig.10(c)* illustrates the result of discriminating pedestrians (blue rectangles) from other moving objects (red rectangles).

In order to evaluate the result of extraction of the foreground objects, they are compared to the ground truth images which are extracted manually from the input images frame by frame. An example of the ground truth images is shown in ***Fig.11(a)***. This image is compared to the extracted foreground objects, shown in *Fig.11(b)*, frame by frame. Comparing *Fig.11(a)* and *Fig.11(b)*, we obtain the image as shown in *Fig.11(c)*. In *Fig.11(c)*, pixels drawn in red represent common pixels between *Fig.11(a)* and *Fig.11(b)*, and they are true positive (TP) pixels. The pixels in green are those included in *Fig.11(a)* but not in *Fig.11(b)*: They are false negative (FN) pixels. Conversely, blue colour pixels are those contained in *Fig.11(b)* but not in Fig.11(a) and they are false positive (FP) pixels.

For evaluating the accuracy of the detected objects, consideration on the number of detected objects, and on the shape of the detected objects are generally introduced. Ideally, our system is to recognize and judge pedestrians that have high risk to a vehicle driver from the results of the segmentation, thus giving a barometer of danger to the driver. For this reason we take the shape based detection into account. To evaluate the accuracy of the result of the detection, some definitions are given as follows;

$$precision = \frac{TP}{TP + FP} \tag{12}$$

$$recall = \frac{TP}{GT} \tag{13}$$

$$FPR = \frac{FP}{TP + FP} \tag{14}$$

$$undetected = \frac{N_{un}}{N_{GT} + N_{un}} \tag{15}$$

Here TP and FP stand for respective areas (the number of pixels); GT stands for the area in the ground truth image. The *precision* given by Eq.(12) means a measure of exactness, whereas the *recall* of Eq.(13) is a measure of completeness of each segmented object. If the value of Eq.(13) is greater than 0.7, the object is regarded as detected successfully; otherwise undetected. FPR is defined by Eq.(14) and it is a measure of incorrectness. The *undetected* defined by Eq.(15) is a measure of the number of undetected objects occluded by vehicle or other objects or the recall value was less than 0.7. Note that, N_{un}, and N_{GT} are the number of undetected objects, and the number of the objects in the ground truth image, respectively.

With respect to the video shown in *Fig.10*, 112 successive frames were evaluated and the extraction achieved 82.8[%] with *precision* and 68.2[%] with *recall*, whereas *FPR* was 17.3[%], and *undetected* was 15[%]. ***Figure 12*** illustrates the relation between the FPR (the abscissa) and the recall (the ordinate) with respect to the accuracy of the extraction of pedestrians (a blue line) and other moving objects (a red line) in the video of *Fig.10*.

The used PC contains Intel Core2 2.4 GHz CPU with 4GB memory. The total computation time for processing a single image frame is 39.6 ms/frame in average. It depends on the number of objects, the complexity of the urban environment and the background.

9. DISCUSSION

This paper proposed a technique for detecting the background sequentially from a video taken by a moving camera (vehicular vision) and for recognizing the moving objects that appears on the foreground. The technique employs a single normal distribution model as a background model at each pixel on an image. In the performed experiment, a pedestrian was successfully extracted from the video taken near a zebra crossing. Almost real time processing, i.e., 39.6 ms/frame in average was achieved. Temporal median filtering, running averaging and mixture of Gaussian modelling were also considered in the background detection, but they didnt show better performance than the proposed single normal distribution model.

It is also emphasized that the proposed technique clusters pedestrians and other transports robustly even in turning-right scenes in which moving objects exist more than going-straight or turning-left scenes.

The advantage of human extraction based on the background detection over existent stereo-based techniques is that it can directly extract human shape, which can be employed for motion or action recognition employing a motion database [30], and human body direction detection [33].

The drawback of the background detection technique is that motionless objects are included in the background. Change in the appearance when a camera moves may be employed for its detection [31]. This remains for further study, however.

Figure 12 shows that the *recall* of a pedestrian is approximately 0.82 when the FPR is 0.28, whereas the *recall* of other moving objects (cars, trucks, bicycles, etc.) gains 0.96 when the FPR is less than 0.08. The reason for the lower *recall* values of a pedestrian compared to other moving objects may come mainly from the variation of its shape. The criterion given by Eq.(10) may not hold for a pedestrian carrying a large baggage or riding on a bicycle. The criterion may also be weak for a pedestrian occluded by other objects. Hence the improvement with the present technique needs to be done in particular for extracting exact shape of a pedestrian such as flying out into the road lane, absorbed in reading, writing or talking using his/her cellular phone (or smart phone), fallen-down, etc. We also intend to investigate on the selection of the threshold $Th1$, $Th2$ and the aspect ratio, or even to introduce a new decision method to improve the segmentation of each object. In addition, for achieving practicality of the proposed technique, the overall computation time of the technique/system is expected to be less than 30ms/frame.

10. CONCLUSIONS

In order to assist safe driving, we have proposed a technique for detecting background images sequentially from a video provided by a camera installed in a vehicle and extracting moving objects including pedestrians by segmenting them from the background. The technique was examined and compared to other techniques by real video images captured at a traffic junction of an urban area when a vehicle turns left or right and satisfactory results were obtained. The shape of a pedestrian in various situations needs to be extracted exactly in order to achieve higher *recall* values.

11. ACKNOWLEDGEMENTS

This work was supported by JSPS KAKENHI (22510177) and a grant of Knowledge Cluster Initiative implemented by MEXT, which are greatly acknowledged.

References

[1] R. Bishop, Intelligent vehicle technology and trends, Artech House, Inc., 2005.

[2] S. Ezell, Intelligent Transportation Systems, The Information Technology & Innovation Foundation, 2010

[3] B. Fardi, U. Schuenert and G. Wanielik, “Shape and Motion-based Pedestrian Detection in Infrared Images :A Multi Sensors Approach,” *Proceedings of IEEE Intelligent Vehicles Symposium*, pp.18-23,2005.

[4] H. Nanda and L. Davis, “Probabilistic Template Based Pedestrian Detection in Infrared Videos,” *Proceedings of IEEE Intelligent Vehicles Symposium*, p.15-20, 2002.

[5] D. M. Gavrila and S. Munder, “Multi-cue Pedestrian Detection and Tracking from a Moving Vehicle,” *International Journal of Computer Vision*, Vol.73, No.1, pp.41-59, 2007.

[6] L. Zhao and C. Thorpe, “Stereo- and Neural Network- Based Pedestrian Detection,” *IEEE Transactions on Intelligent Transportation Systems*, Vol.1, No.3, pp. 148-154, 2000.

[7] D. M. Gavrila and V. Philomin, “Real-time Object Detection for Smart Vehicles,” *Proceedings of IEEE International Conference on Computer Vision*, pp.87- 93, 1999.

[8] D. M. Gavrila, “Detection from a Moving Vehicle,” *Proceeding of the Europ. Conference on Computer Vision*, No.2, pp.37-49, 2000.

[9] D. M. Gavrila and S. Munder, “Multi-cue Pedestrian Detection and Tracking from a Moving Vehicle,” *International Journal of Computer Vision*, Vol.73, No.1, pp.41-59, 2007.

[10] N.Dalal and B.Triggs, “Histograms of Oriented Gradients for Human Detection,” *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Vol.1, pp.886-893, 2005.

[11] P. Viola, M. Jones and D. Snow, “Detecting Pedestrians Using Patterns of Motion and Appearance,” *International Journal of Computer Vision*, Vol.63, No.2, pp.153-161, 2005.

[12] A.Broggi, M.Bertozzi, A.Fascioli and M. Sechi, ”Shape-based pedestrian detection”, *Proceedings of the IEEE Intelligent Vehicles Symposium*, pp. 215-220, 2000.

[13] H. Cheng, N.Zheng and J. Qin, “Pedestrian Detection Using Sparse Gabor Filter and Support Vector Machine,” *Proceedings of IEEE Intelligent Vehicles Symposium*, pp.583-587, 2005

[14] Zhang, Y. et al., “Robust Moving Object Detection at Distance in the Visible Spectrum and Beyond Using a Moving Camera,” *Proceedings of CVPR Workshop*, 2006.

[15] M. Jogan, J. Sorli, I. Vidmar, and A. Leonardis, “Detection of Pedestrian Crossings Using Triplets,” *Proceedings of Vision Technologies & Intelligent Maps for Mobile Attentive Interfaces in Urban Scenarios*, 2006.

[16] Bertozzi, M. and A. Broggi, “GOLD: A Parallel Real-time Stereo Vision System for Generic Obstacle and Lane Detection,” *IEEE Transactions on Image Processing*, Vol.7, No.1, pp.62-81, 1998.

[17] C. Stauffer and W.E.L. Grimson, “Adaptive Background Mixture Models for Real-time Tracking,” *Proceedings of Conference on Computer Vision and Pattern Recognition*, Vol.2, pp.246-252, 1999.

[18] A. Elgammal, R. Duraiswami, D. Harwood and L. Davis, “Background and Foreground Modeling Using Non-parametric Kernel Density Estimation for Visual Surveillance,” *Proceedings of the IEEE-PIEEE*, Vol.90, No.7, pp.1151-1163, 2002.

[19] K. Kim, T.H. Chalidabhongse, D. Harwood and L. Davis, "Background Modeling and Subtraction by Codebook Construction," *Proceedings of International Conference on Image Processing*, Vol.5, pp.3061-3064, 2004.

[20] R. Cucchiara, M. Piccardi and A. Prati, "Detecting Moving Objects, Ghost, and Shadows in Video Streams," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.25, No.10, pp.1337-1342, 2003.

[21] E. Hayman and J.-O. Eklundh, "Statistical Background Subtraction for a Mobile Observer," *Proceeding of the 9th International Conference on Computer Vision*, Vol.1, pp.67- 74, 2003.

[22] K. Jota, T. Tsubouchi, Y. Sugaya and K. Kanatani, "Extracting Moving Objects from a Moving Camera Video Sequence," *Workshop of Information Processings Society of Japan*, 2004-CVIM-143-6, pp.41-48, 2004.

[23] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes:Active Contour Models," *International Journal of Computer Vision*, Vol.1, No.4, pp.321-331,1988.

[24] L. D. Cohen, "On Active Contour Models and Ballons," *Computer Vision Graphics, and Image Processing: Image Understading*, Vol.53, No.2, pp.211- 218, 1991.

[25] Y. Y. Boykov, M.P. Jolly, "Interactive Graph Cuts for Optimal Boundary and Region Segmentation of Objects in N-D Images," *Proceedings of International Conference on Computer Vision*, pp.105-112, 2001.

[26] Z. Lin, L. S. Davis, D. Doermann and D. Dementhon, "Hierarchical Part-template Matching for Human Detection and Segmentation," *IEEE Proceeding of 11th International Conference on Computer Vision*, pp.1-8, 2007.

[27] M.A. Fischler and R.C. Bolles, "Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography," *Communication of ACM*, Vol.6, No.24, pp.381-395, 1981.

[28] C. Harris, M. Stephens, "A Combined Edge and Corner Detector," *Proceeding of the 4th Alvey Vision Conference*, pp.147-151, 1988.

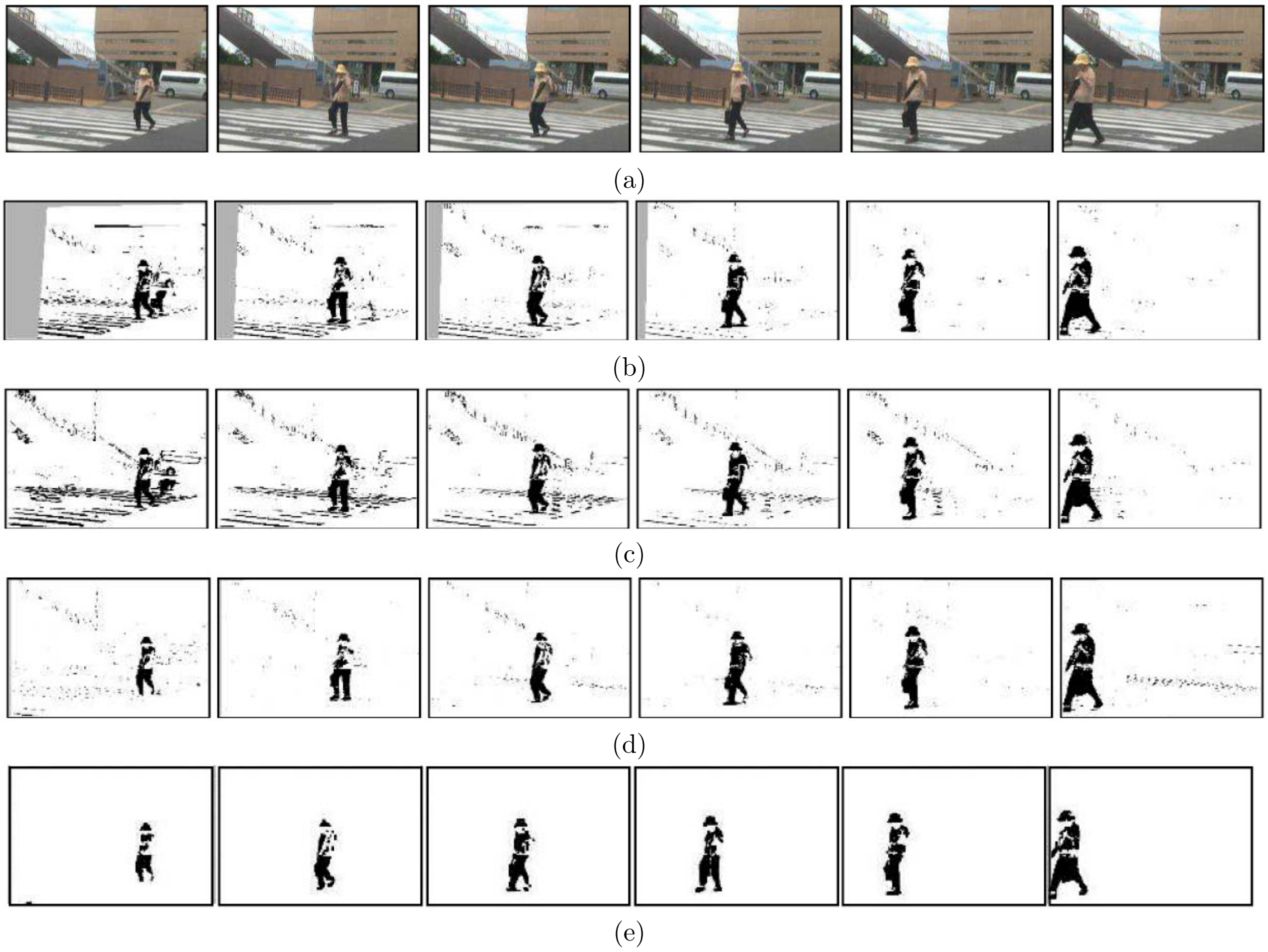
[29] B.D. Lucas and T. Kanade, "An Iterative Image Registration Technique with an Application to Stereo Vision," *Proceeding of the 7th International Joint Conference on Artificial Intelligence*, pp.674-679, 1981.

[30] J.K. Tan, K. Kouno, S. Ishikawa, H. Kim and T. Shinomiya, "High Speed Human Motion Recognition Employing a Motion Database," *Imaging and Visual Computing*, Vol.36, No.5, pp.738-746, 2007. (in Japanese).

[31] S. Qian, J.K. Tan, S. Ishikawa, T. Morie, "Obstacle Detection Using a Moving Camera," *Proceedings of International Symposium on Artificial Life & Robotics*, 2011, pp. 767-770, 2011.

[32] H. Uemura, K. Mikolajczyk, J.K. Tan, S. Ishikawa, "Multiple Feature Points Tracking for Camera Motion Compensation," *Journal of Biomedical Fuzzy Systems Association*, Vol.11,No.1, pp.1-9, 2009.

[33] Yuuki Nakashima, J.K. Tan, S. Ishikawa, T. Morie, "Detecting a Human Body Direction Using Multiple- HOG," *Proceedings of the First International Symposium on Future Active Safety Technology toward zero-traffic-accident*, CD-R TS3-8-2- 5(6pages), 2011.



(a)

(b)

(c)

(d)

(e)

Fig.5: *Comparison of the Proposed Background Estimation Technique with Other Techniques. Experimental Results on Turning-left Video which Consists of 98 Frames of 320×220pixels. (a) Part of the Original Frames, (b) The Result Obtained from the Temporal Median (TM) Method, (c) The Result Obtained from the Running Average (RA) Method, (d) The Result Obtained from the Proposed Technique Employing the Normal Distribution Model (NM).(e) The Result after Applying Expansion-Contraction Post-processing to (d)*

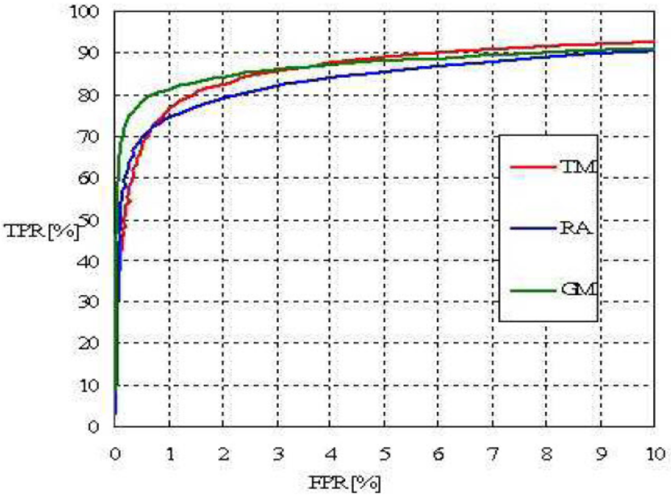


Fig.6: *The ROC Curve with the Result in Fig. 5.*

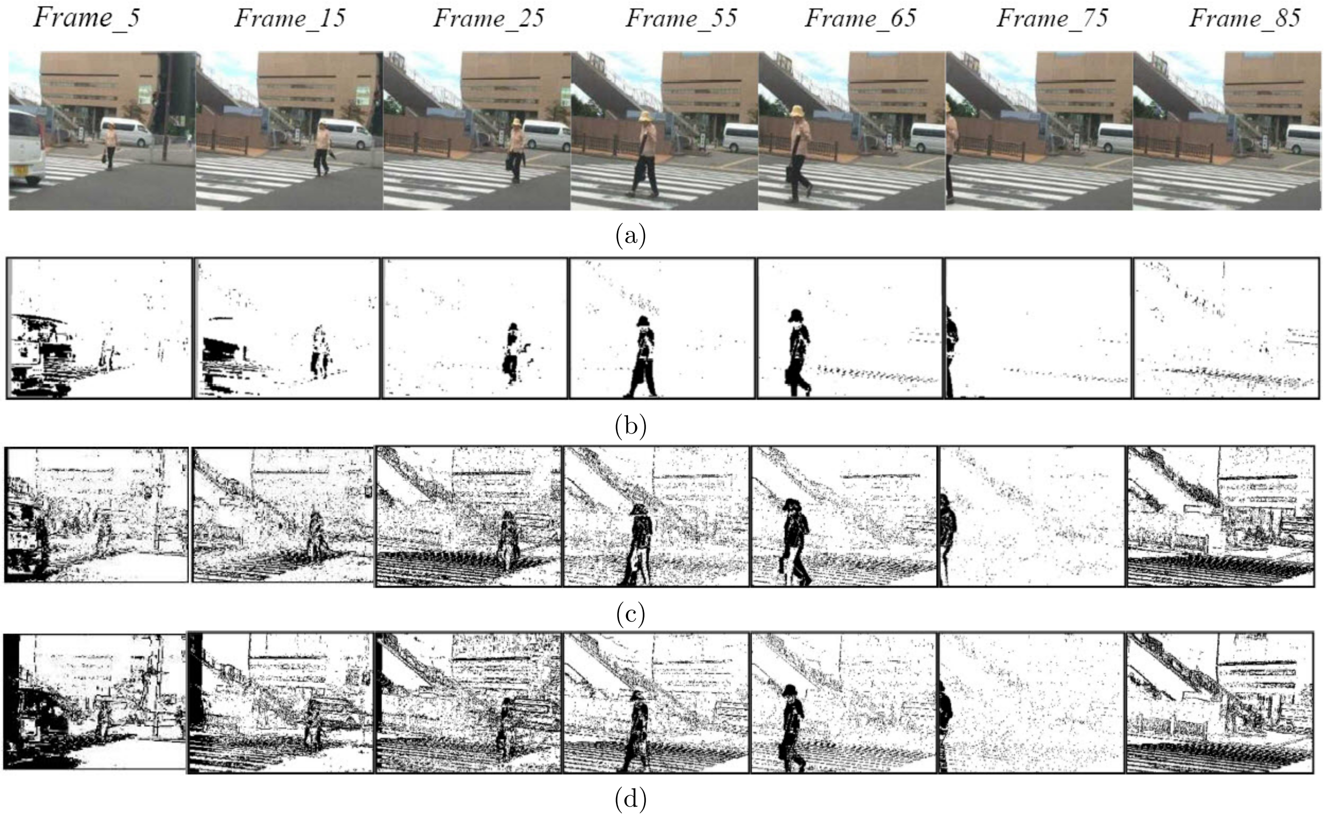


Fig.7: *Comparison of the Effectiveness of Renewing the Background between the Proposed Technique(NM) and Mixture of Gaussian Models Technique(MoG).(a) The Original Video Images,(b) Extracted Foreground by the Proposed Method , (c) Extracted Foreground by the MoG Technique using a Single-Gaussian Distribution, (d) Extracted Foreground by the MoG Technique using Three-Gaussian Distributions..*

Table 2: *Computation Time of the Proposed Technique along with Other Techniques (CPU: Intel Core2 2.40GHz, Total Number of Image Frames: 98)*

Computation Time [msec / frame]					
Method	*Proposed*	*TM*	*RA*	*MoG (Single Gaussian Distribution)*	*MoG (Three Gaussian Distributions)*
Total Processing Time	51.45	139.10	30.46	41.12	58.53

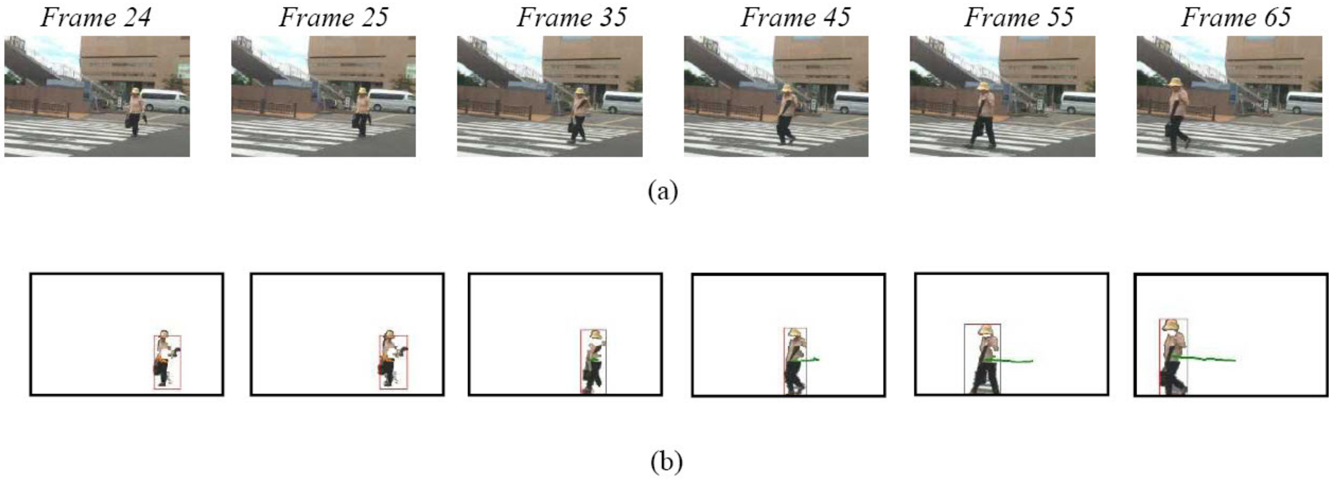


Fig.8: *Result of the Objects Segmentation on Fig.5. (a) Original Video Scene, (b) Result of the Segmentation (Red Rectangle: Pedestrian, Green Line: Tracked Trajectory).*

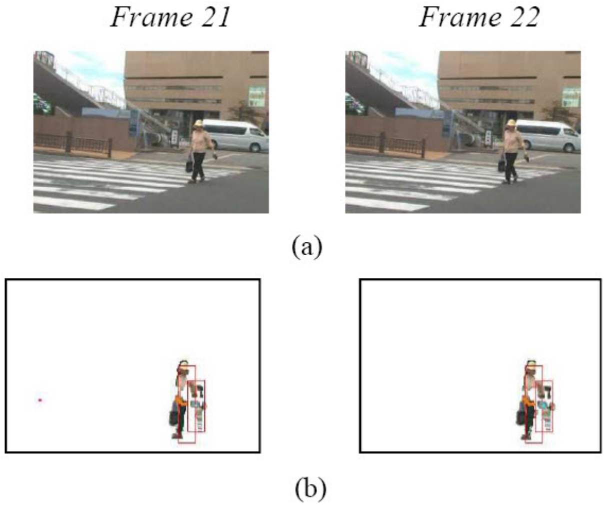
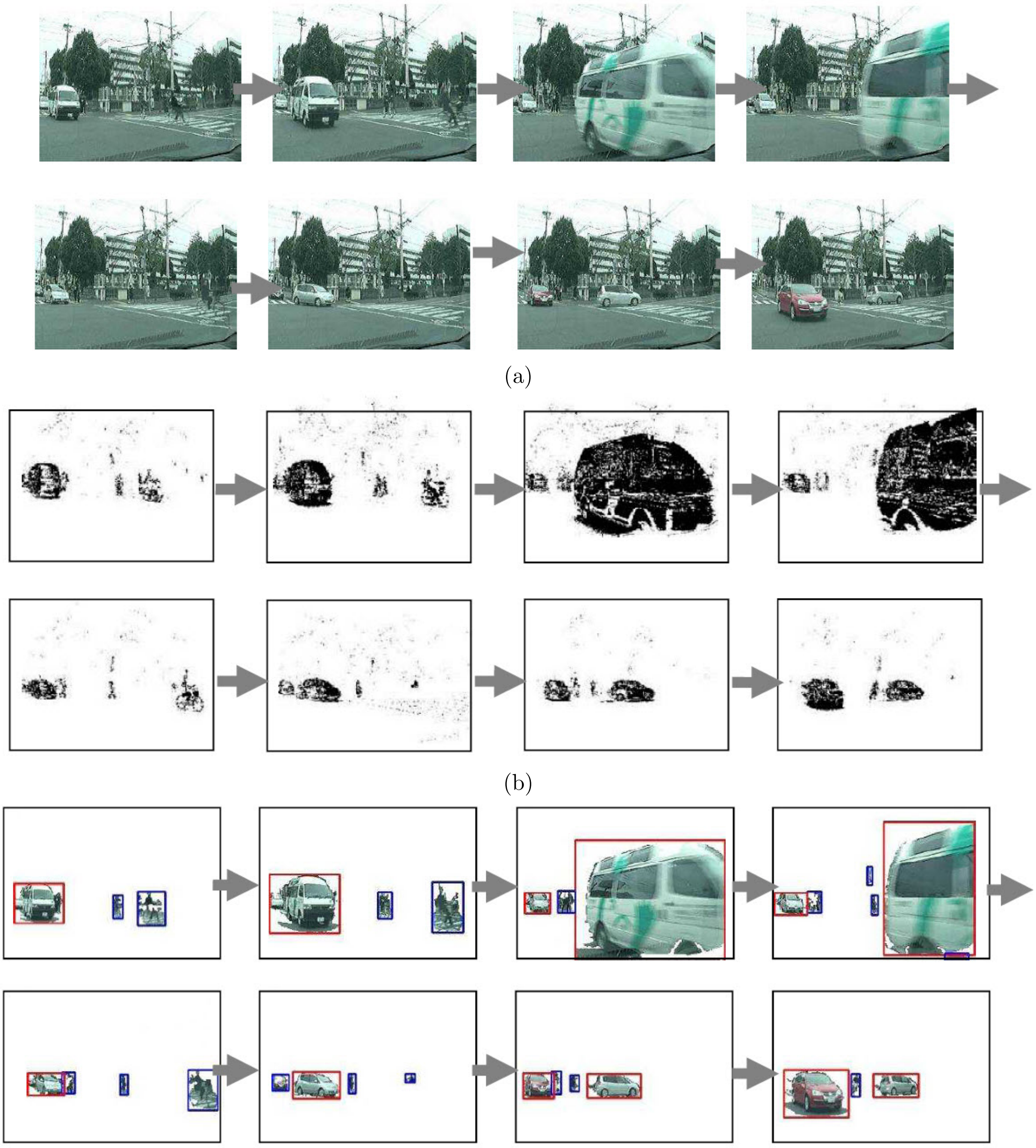


Fig.9: *Examples of Incorrect Segmentation with Fig. 5. (a) Original Image Frames, (b) Result of the Segmentation.*



(a)

(b)

(c)

Fig.10: *Experimental Results on Turning-right Video which Contains 112 Image Frames each Having 320×220 Pixels (a) Original Video Scene,(b) Result of Foreground Extraction, (c) Result of Objects segmentation (Blue Rectangle: Pedestrians, Red Rectangle: Vehicle)*

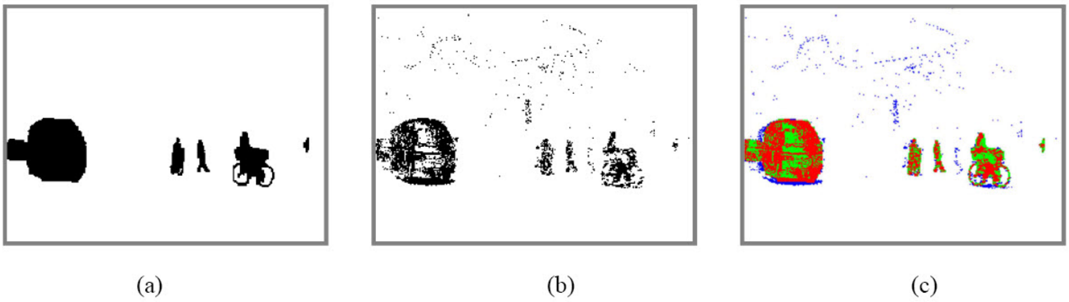


Fig.11: *Evaluation of the Experimental Result. (a) The Ground Truth Image, (b) The Extracted Foreground Image, (c) Result of the Comparison Between (a) and (b) (Red Pixels: True Positive; Green Pixels: False Negative; Blue Pixels: False Positive)*

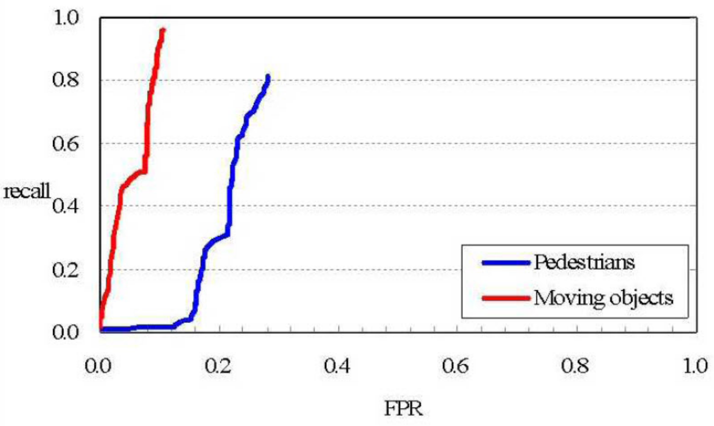


Fig.12: *Evaluation Curve of the Segmented Objects with respect to the Result in Fig. 11*

Joo Kooi Tan received B.E. and M.E. degrees in Computer Science and Ph.D degree in Control Engineering from Kyushu Institute of Technology. She is presently with Department of Mechanical and Control Engineering in the same university as an Associate Professor. Her current research interests include three-dimensional shape/motion recovery, human motion analysis, and human activities recognition. She received the SICE Kyushu Branch Young Authors Award in 1999, the AROB10th Young Authors Award in 2004, Young Authors Award from IPSJ of Kyushu Branch in 2004 and BMFSA Best Paper Award in 2008 and 2010. She is a member of IEEE, The Society of Instrument and Control Engineers, and The Information Processing Society of Japan.

Seiji Ishikawa obtained B.E., M.E., and D.E. from The University of Tokyo, where he majored in Mathematical Engineering and Instrumentation Physics. He joined Kyushu Institute of Technology and he is currently Professor of Department of Control & Mechanical Engineering, KIT. Professor Ishikawa was a visiting research fellow at Sheffield University, U.K., from 1983 to 1984, and a visiting professor at Utrecht University, The Netherlands, in 1996. He was awarded BMFSA Best Paper Award in 2008 and 2010. His research interests include threedimensional shape/motion recovery, and human detection and its motion analysis from car videos. He is a member of IEEE, The Society of Instrument and Control Engineers, The Institute of Electronics, Information and Communication Engineers, and The Institute of Image Electronics Engineers of Japan.

Shinichiro Sonoda received the B.E. degree in Mechanical and Control Engineering from Kyushu Institute of Technology in 2010. He is now a Master student in the same university. He research interests include computer vision and image processing. He is a student member of The Society of Instrument and Control Engineers.

Makoto Miyoshi received the M.E. degree in Mechanical and Control Engineering from Kyushu Institute of Technology in 2009. His research interests include computer vision and image processing. He is a student member of The Society of Instrument and Control Engineers.

Takashi Morie received the B.S. and M.S. degrees in physics from Osaka University, Osaka, Japan, and the Dr.Eng. degree from Hokkaido University, Sapporo, Japan, in 1979, 1981 and 1996, respectively. From 1981 to 1997, he was a member of the Research Staff at Nippon Telegraph and Telephone Corporation (NTT). From 1997 to 2002, he was an associate professor of the department of electrical engineering, Hiroshima University, Higashi-Hiroshima, Japan. Since 2002 he has been a professor of Graduate School of Life Science and Systems Engineering, Kyushu Institute of Technology, Kitakyushu, Japan. His main interest is in the area of VLSI implementation of neural networks, mixed/merged analog-digital circuits, new functional devices, and image processing. Dr. Morie is a member of IEEE, IEICE, IEEJ, the Japan Society of Applied Physics and the Japanese Neural Network Society.