Proceedings of
IEEE International Conference on
Multisensor Fusion and Integration for Intelligent Systems
Seoul, Korea, August 20 - 22, 2008

Foot Position Estimations for Moving Objects using a Mixture Model

First T. Hirose, Second S. Morishita, and Third H. Asama

***Abstract*—As a sensing technology, the use of monocular cameras for security is widespread. This paper presents a proposal of a novel method for positioning moving objects in footage obtained from monocular cameras. The proposed method, which estimates foot position by calculating the mean and the variance of the figure field, is inapplicable to plural persons. We presume Gaussian mixture models in the image and estimate each distribution, which enables position estimation of many pedestrians.**

I. Introduction

To offer appropriate service, for example, directory enquiries, along with the intention of pedestrians in large place such as public spaces, it is useful to estimate their intentions by analyzing the person's movement [1]. The CCD sensor is suitable to get a movement trace because the setting of a new sensor is unnecessary by the spread of surveillance cameras. Two procedures are necessary to acquire a trajectory rapidly in footage obtained from monocular cameras.

–First, extract a moving object domain from an animated image.
–Second, calculate the foot position from the domain.

The first requires a method of extracting a moving objects domain that is robust to lighting conditions [2]. Furthermore, the second would benefit from a method of estimating the foot position from the mean of a person's figure field [3]. This method proposes that the foot position be estimated by assuming the major axis' direction as an oval encircling humans as a step position. Unfortunately, this technique is inapplicable for images that include plural people.

From the points described above, this report proposes sensing the position of plural walking people for extracting foot traces, along with a method for estimating respective mean positions of many pedestrians by presuming Gaussian mixture models in the image using the EM algorithm.

II. Related Works

A. Classification of sensors used in methods to trace pedestrians

Sensors of movement traces are categorizable into the following three types.

–First, non-vision sensors
–Second, sensor fusion
–Third, vision only

Non-vision sensors use infrared sensor networks [4], pyroelectric infrared detectors [5] [6], optical ID tags and floor pressure sensors [7], laser range-finders [8], those overlapping with the fusion domain, and so on.

Sensor fusion methods use unified systems of multi-directional cameras and floor pressure sensors [9], a monocular camera and a laser range finder [10], pyroelectric detector and a Fresnel lens array [11], and so on. Generally, such sensors themselves are very expensive.

For vision only sensors, various methods are suggested, including plural cameras [12] and multi-directional cameras [13]. However, these sensors are not generally available. To install new equipment is costly. Therefore, monocular cameras were adopted for this study because their use for security purposes is widespread, new installations are probably unnecessary, and low cost installation is possible if necessary.

B. Methods using monocular cameras

[14] [15] [16] [17] describe conventional techniques using monocular cameras. A tracking-based S-T MRF Model [17] is a representative method, but it poses problems: it cannot distinguish plural pedestrians who are mutually adjacent and moving similarly. Furthermore, calculation costs are not low.

Tracking methods based on background subtraction algorithms (BSAs) are simple and easily applicable for high-speed moving object detection [18] [19]. Notwithstanding, problems of vulnerability to separation of labels from people and occlusion occur: on balance, their reliability is low.

In light of the points described above, this report presents a proposal of a high-speed and reliable measurement method of the movement trace of plural pedestrians using a monocular camera

III. Proposed Method

A. Premises

We consider the coordinate values of the pixels of a

Manuscript received March 1, 2008.

T. Hirose is with the Faculty of Engineering, The University of Tokyo, Kashiwanoha 5-1-1, Kashiwa-shi, Chiba 277-8568, Japan (phone: +81-4-7136-4260; fax: +81-4-7136-4242; e-mail: hirose@race.u-tokyo.ac.jp).

S. Morishita is with Research into Artifacts, Center for Engineering (RACE), The University of Tokyo, Kashiwanoha 5-1-1, Kashiwa-shi, Chiba 277-8568, Japan (phone: +81-4-7136-4260; fax: +81-4-7136-4242; e-mail: mori@race.u-tokyo.ac.jp).

H. Asama is with Research into Artifacts, Center for Engineering (RACE), The University of Tokyo, Kashiwanoha 5-1-1, Kashiwa-shi, Chiba 277-8568, Japan (phone: +81-4-7136-4255; fax: +81-4-7136-4242; e-mail: asama@race.u-tokyo.ac.jp).

978-1-4244-2144-2/08/$25.00 ©2008 IEEE

background differential image that has been binarized to its constituent statistical elements. The method proposed by Morishita [3], presumes that a region of a person is the set of points of a normal distribution; it then calculates a step position from a center of gravity and the dispersion of the region of a person, as depicted in Fig. 1.

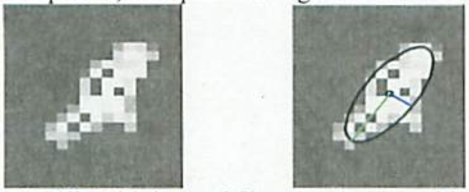

Fig. 1. Differential image (left) and a human-modeled-oval calculated using the mean and the divergence (right).

It is necessary to calculate centers of gravity and dispersions of the region of a person severally when plural people exist in a background differential image. Therefore, we presume the region of a person to be the set of points generated using a normal mixture distribution. For example, the following figures show the region of moving objects and a histogram calculated from the image where two pedestrians exist, as presented in Fig. 2.

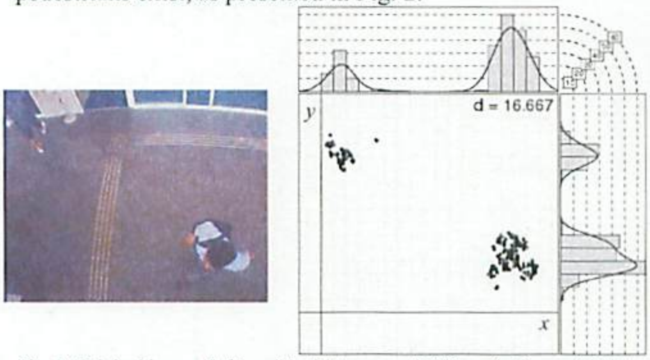

Fig. 2. Original image (left) and the histogram of differential image (right).

Figure 2 shows a background differential image containing two people: it is generated by the mixture distribution of two normal distributions. By a calculation of the center of gravity and the dispersion of each class, humans should be positioned.

However, estimation of the mixture distribution to estimate which class each point belongs to is necessary for a background differential image which does not include class information. For that reason, we adopt the EM algorithm for general estimation of the mixture distribution.

B. Gaussian mixture models

The density function of the normal distribution ϕ is defined as type (1).

$$\phi(x;\mu,\Sigma)=\frac{1}{2\pi|\Sigma|}\exp\left(-\frac{(x-\mu)^T\Sigma^{-1}(x-\mu)}{2}\right) \quad (1)$$

Here, x represents a coordinate value, μ is the mean of the normal distribution, Σ is a covariance matrix, and Σ^{-1} is the inverse matrix of Σ.

In the following equation, K is the number of classes of a mixed normal distribution when the mixture degree of the distribution is equal. The probability that sample a certain point x was generated by the normal mixture distribution of K unit is given as

$$p(x;\boldsymbol{\mu},\boldsymbol{\Sigma})=\frac{1}{K}\sum_{k=1}^{K}\phi(x;\mu_k,\Sigma_k)\,. \quad (2)$$

It is assumed that sample marks of N units were generated by the normal mixture distribution of expression (2); the likelihood of the data is expressed as

$$p(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma})=\frac{1}{K}\prod_{n=1}^{N}\sum_{k=1}^{K}\phi(x_n;\mu_k,\Sigma_k) \quad (3)$$

Parameters of this normal mixture model are decided with maximization of the likelihood that sample points x are observed. Then, we adapt an Expectation-Maximization (EM) algorithm. An EM algorithm performs maximum likelihood estimation in the situation that some of variables are unobserved.

An EM algorithm is one of the methods for parameter estimation of contaminated normal distribution. It was generally formulated by Dempster [20], and commonly used in statistics. Because it performs between an expectation step (E step) and a maximization step (M step), it is called an EM algorithm.

In the next section, E step is described in detail as follows.

C. E step

According to Dempster [20], given initial values of model parameters (μ, Σ), the EM algorithm guarantees calculation of more plausible model parameters using maximum likelihood method. By repeating the estimation and updating model parameters, EM algorithm acquires the limited part most suitable solution dependent on the initial values.

The probability $\overline{Q}_n(k)$ that each sample point x_n is generated by the Gaussian distribution of k joint is expressed with a suitable initial value $(\overline{\mu},\overline{\Sigma})$, as

$$\overline{Q}_n(k)=\frac{\phi(x_n;\overline{\mu}_k,\overline{\Sigma}_k)}{\sum_{k=1}^{K}\phi(x_n;\overline{\mu}_k,\overline{\Sigma}_k)}\,. \quad (4)$$

This $\overline{Q}_n(k)$ is calculated for each sample point x_n using a suitable initial value $(\overline{\mu},\overline{\Sigma})$ and is not afterward probability itself that x_n generated using the normal distribution of the k joint. In the M step, each parameter is estimated using this result.

D. M step

Each parameter is calculated as follows using maximum likelihood method, using the result of the E step.

$$\mu_k = \frac{\sum_{n=1}^{N} \overline{Q}_n(k) x_n}{\sum_{n=1}^{N} \overline{Q}_n(k)} \qquad (5)$$

$$\Sigma_k = \frac{\sum_{n=1}^{N} \overline{Q}_n(k)(x_n - \mu_k)(x_n - \mu_k)^T}{\sum_{n=1}^{N} \overline{Q}_n(k)} \qquad (6)$$

Using the proposed method, *E*-step and *M*-step are applied to every frame to calculate a model parameter online. Consequently, the results of estimation in the time are the model parameters in the time being updated every frame.

E. Application of the EM algorithm

Target tracking is enabled using a background differential image depicting plural people by computing each person's center of gravity and dispersion. Means and divergences of each person's figure field in the target frame are estimated (updated) using the EM algorithm with initial values of dispersion and a center of gravity that were updated in a previous frame, as depicted in Fig. 3.

However, this method requires input of the number of classes and the number of people. Therefore, it is assumed that increases and decreases of the number of people are acquired using an appropriate measurement method of people [21]. Then addition and deletion are necessary for the increase and decrease of the number of people in subsequent images over time.

F. Procedures used for addition of a class

When a person enters the image, the region of the person should be far from all classes being tracked. The distance between a class and a pixel is evaluated using the likelihood calculation used for the EM algorithm. Here the likelihood is the probability that pixel *x* belongs to a normal distribution of class *k*; it is shown as expression (1).

Based on the points described above, with pixels for which the sum of the likelihood between the pixel and all classes is smaller than threshold *q* (the pixels distant from all classes) are divided; these belong to a new class. Then addition of the class was achieved. An evaluation function of the sum of likelihood sets is shown as follows.

$$F_{add}(n) = \sum_{k=1}^{K} \phi(x_n ; \mu_k , \Sigma_k) \qquad (7)$$

Let x_n $\{n \mid A(n) \le \theta\}$ be elements of a new class. The result of the addition of a class is shown in Fig. 4. The histogram of $F_{add}(n)$ and the result of the classification are shown in Fig. 5.

G. Procedure of deletion of a class

When a person exits from the image, the surroundings of the class where the person belonged is expected not to have a pixel. The class for which the sum of the likelihood between the class and all pixels is the smallest (i.e., no pixel exists around the class) is divided and deleted. Thereby, deletion of the class was achieved. The evaluation function of the sum of likelihood sets is shown as the following expression.

$$F_{del}(k) = \sum_{n=1}^{N} \phi(x_n ; \mu_k , \Sigma_k) \qquad (8)$$

Delete class $k_{\min}$ s.t. $F_{del}(k_{\min}) = \min F_{del}(k)$. The result of the deletion of a class is portrayed in Fig. 6.

H. Flow of the proposed method

The flow of the algorithm is presented in Fig. 7.

Step 1: Setting an initial state
A screen in an initial state is expected to show no people. However, it can support the existence of plural people by application of the *k*-means method or if the number of people on the screen is known.

Step 2: Observation of data
The pixels of the differential images are randomly sampled fixed numbers, as presented in Fig. 8. The calculation of the information is enabled by about 1/10 to 1/100 of the original information.

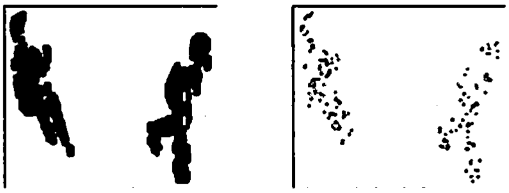

Fig. 8. Differential image (left) and sample points (right).

Step 3: *E* step
The likelihood is evaluated using expression (4).

Step 4: Update the number of classes
The change of the number of people is given and the number is updated for the number of the classes depending on a case.

Step 5: *M* step
The model parameters are updated using expressions (5) and (6). The series of operations is returned to (2) and repeated by the end.

IV. Experiment

We performed an experiment to apply the suggested method to an animated sequence acquired using a camera set above the first floor entrance of the Kashiwa Library of Tokyo University to assess the effectiveness of the proposed method. The sampling points were 400; θ was set to 1.0×10^{-6}.

A. Results

Position sensing was possible for most frames and yielded good results, as depicted in Fig. 9. The result shows that our proposing method has high reliability and robustness for occasion of occlusion. Regarding execution speed, as shown in a reference experiment, high-speed handling of 3–5 μs degree was ensured.

B. Evaluation

The 15 fps animations, cut off to eight scenes in which plural people appeared, were evaluated by viewing every frame. Failures are those cases in which a direction and an oval size were remarkably inappropriate; a step position was not identified. In fact, 357 of 441 frames were evaluated successfully; the distinction rate was therefore 80.9%.

C. Consideration

Almost all frames were successful in instances where there was no pedestrian on the screen edge and pedestrians were mutually separated. The proposed method is robust to occlusion, and is able to distinguish plural people who move similarly.

Failure examples are portrayed in Fig. 10. The following are factors that distinguished failures.

(1) Problems of a background difference

A background difference method can not distinguish some clothes with similar color to that of the background from the background; non-human moving objects are indistinguishable from people.

(2) Problems of the screen edge

The form of a person collapses and fails to allow modeling of the oval when it appears at the screen edge. Coverage of the method in the domain is limited to an area that is smaller than the frame: only people in that domain are detectable. This problem can be solved.

(3) Problems of sampling

Because the number of sampling points is fixed, if there are many people in an image, the pixels per person are expected to be few. For that reason, they are difficult to estimate correctly. An appropriate sampling can reduce computational complexity, but it might cause such a false distinction. The number of sampling points by the experiment in this report is set from experience, and the setting of an appropriate value can be quantitatively examined. In addition, there might be deflection in an extracted points because they chosen at random. There is room for the examination in a choice method of extraction of points, such as to select apart between extracted points.

(4) Problem of occlusion

It is impossible with a single camera to alleviate occlusion. However, the proposed method is comparatively robust to occlusion, as judged from the result.

Most false distinction occurs because the four described above are mutually occurring. Further improvement of the distinction rate is expected from solving these problems, although problems exist that cannot be settled fundamentally.

V. CONCLUSIONS

To offer appropriate service for estimating the pedestrians' intentions, this report described a high-speed and reliable measurement method of movement tracing of plural pedestrians using a monocular camera.

As a future subject, implementations of the method performing online and on a real-time basis are considered along with improvement of the distinction rate. Concretely,

(1) Examination of the number of appropriate sampling points

(2) Examination of a method assuming the same distribution, which would be proper as a presupposition of the region of humans than normal distribution

(3) Reflection of the prediction of movement using means such as Karman filters.

VI. REFERENCE EXPERIMENT

For an animation image photographed in the Kashiwa Library of Tokyo University, we measured the speed of the EM algorithm under the following conditions.

(1) CPU: Intel Pentium M 1.2 GHz

(2) Number of sampling points, 400

(3) An initial value is estimated using k-means method for every frame

Non-use of calculated results of the previous frame should be given attention.

TABLE I
RESULTS OF REFERENCE EXPERIMENT

Number of classes	2	3	4	5
Average [ms]	3.529	3.949	4.645	5.344
Standard deviation [ms]	1.564	0.559	0.954	0.896

A. Results

The results are as presented in Table 1. The average and the standard deviation were calculated for 150 frames.

B. Consideration

Convergence judgment using a threshold is unnecessary. Therefore, high-speed processing is realized uniformly. In addition, processing in $O(M)$ is possible for the number of people M because the number of the samplings is constant. When the calculation result of the previous frame is not used for the EM algorithm, the execution speed is sufficiently rapid, but high-speed processing can be expected when the calculated result of the previous frame is used because an initial value and an estimate are similar.

REFERENCES

[1] A. Nishimura, S. Morishita, and H. Asama, "Estimation of Destination from Walking Patterns using Hidden Markov Model," *Proceedings of the 2007 JSME Conference on Robotics and Mechatronics*, Akita, Japan, May 10-12 , pp. 2P1-C10 (1)-2 P1-C10 (3), 2007

[2] J. Kamibata, S. Morishita, and H. Asama, "Study on Shadow Detection using Correlation Analysis for Moving Object Extraction, " *Proceedings of the 2007 JSME Conference on Robotics and Mechatronics*, Akita, Japan, May 10-12, pp. 2P1-C09 (1)-2 P1-C09 (4) , 2007

[3] S. Morishita and H. Asama, "Study on Identification of Position of Foot from Barycentre of Figure in the Footage, " *7th SICE System*

Integration Division Annual Conference, Sapporo, Dec. 14-16, pp.1380-1381, 2006

[4] Zhiqiang Zhang, Xuebin Gao, Jit Biswas, Jian Kang Wu, "Moving targets detection and localization in passive infrared sensor networks," *Information Fusion, 2007 10th International Conference on* , vol., no., pp.1-6, 9-12 July 2007

[5] J. -S. Fang, Q. Hao, D. J. Brady, M. Shankar, B. D. Guenther, N. P. Pitsianis, and K. Y. Hsu, "Path-dependent human identification using a pyroelectric infrared sensor and fresnel lens arrays," *Opt. Express 14*, 609-624, 2006

[6] Burchett, J. and Shankar, M. and Hamza, A.B. and Guenther, B.D. and Pitsianis, N. and Brady, D.J., "Lightweight biometric detection system for human classification using pyroelectric infrared detectors," *Applied Optics*, vol.45, no.13, pp.3031-3037, 2006

[7] S. Ota, R. Sakamoto, K. Kogure, and T.Fujinami, "Multiple human tracking by integrating pixel-wise Optical ID sensors and floor sensors," *The Institute of Electronics, Information and Communication Engineers*, Vol.105, No.674, pp. 137-142, 2006

[8] H. Yamada, R. Kurazume, K. Murakami, and T. Hasegawa, "Robot Town Project: Target Tracking using Level Set Tracking and Multiple Laser Range Finders," *Nippon Robotto Gakkai Gakujutsu Koenkai Yokoshu (CD-ROM)*, Vol. 24, pp. 2N16, 2006

[9] T. Murakita, T. Ikeda, and H. Ishiguro, "Dynamic Fusion of Visual Features Based on Multisensor Data," *The 19th Annual Conference of the Japanese Society for Artificial Intelligence*, 2005

[10] G. Monteiro, C. Premebida, P. Peixoto, and U. Nunes. Tracking and classification of dynamic obstacles using laser range finder and vision In *Proc. of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2006

[11] Brady, D.J. and Guenther, B.D. and Feller, S. and Shankar, M. and Fang, J.S. and Hao, Q, Sensor system for identifiying and tracking movements of multiple sources, 2007

[12] H. Kato, A. Nakazawa, and S. Inokuchi, "Realtime Human Tracking Using Ellipsoid Model," *Transactions of Information Processing Society of Japan*, Vol.40, No.11, pp. 4087-4096, 1999

[13] Y. Sato, Y. Maita, K. Hashimoto, and Y. Shibata, "Face Image Tracking System for Surveillance using an Omni-directional Camera," *IPSJ SIG Notes*, Vol.2007, No.58, pp. 13-18, 2007

[14] Arulampalam, Sanjeev; Clark, Martin; Vinter, Richard, "Performance of the shifted Rayleigh filter in single-sensor bearings-only tracking," *Information Fusion, 2007 10th International Conference on* , vol., no., pp.1-6, 9-12 July 2007

[15] Ulmke, M. Erdinc, Ozgur Willett, Peter, *Gaussian mixture cardinalized PHD filter for ground moving target tracking*, Information Fusion, 2007 10th International Conference on, Quebec City, QC, Canada, pp. 1-8, 9-12 July 2007

[16] Arthur E. C. Pece and Anthony D. Worrall, *Tracking with the EM Contour Algorithm*, Computer Vision — ECCV 2002, Volume 2350/2002, pp. 3-17, 2002

[17] S. Kamijo et al. "Traffic Monitoring and Accident Detection at Intersections," *IEEE TRANSACTION on ITS*, Vol.1, No.2, pp. 108-118, Jun. 2000

[18] L.M. Fuentes and S.A. Velastin. "People Tracking in Surveillance Applications," *In Second IEEE International Workshop on Performance Evaluation of Tracking and Surveillance, Hawaii*, December 2001

[19] A. Nishimura, S. Morishita, and H. Asama, "Estimation of Destination from Walking Patterns using Hidden Markov Model, " *7th SICE System Integration Division Annual Conference*, Sapporo, Dec. 14-16, pp.772-773, 2006

[20] A.P.Dempster, N.M.Laird, and D.B.Rubin, "Maximum Likelihood from Incomplete Data via The EM Algorithm," *Journal of The Royal Statistical Society (B)*, vol.39, no.1, pp.1-38 , 1977

[21] S.Morishita and H.Asama, "Estimation of Distributed Parameter of Moving Object Region for a Person's Determination," *8th SICE System Integration Division Annual Conference*, pp.1276-1277, 2007

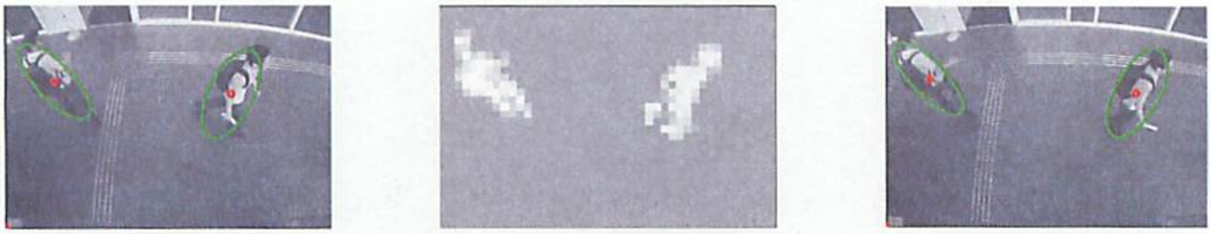

Fig. 3. Application of EM algorithm: Left, Result of computation in previous frame; Center, Differential image; Right, Result of computation in target frame

Fig. 4. Addition of a class: Left, Result of computation in previous frame; Center, Differential image; Right, Result of computation in the target frame.

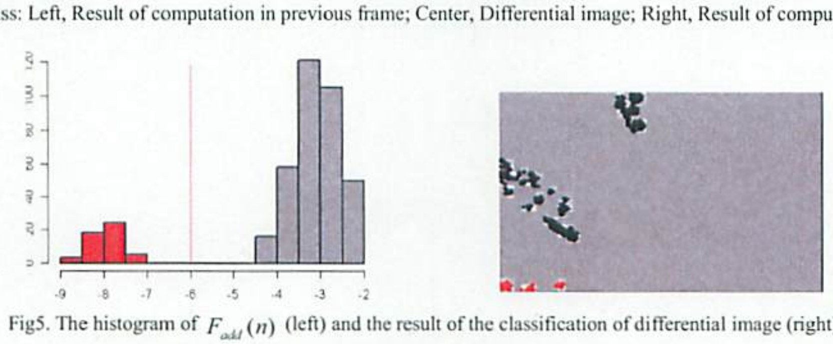

Fig5. The histogram of $F_{add}(n)$ (left) and the result of the classification of differential image (right)

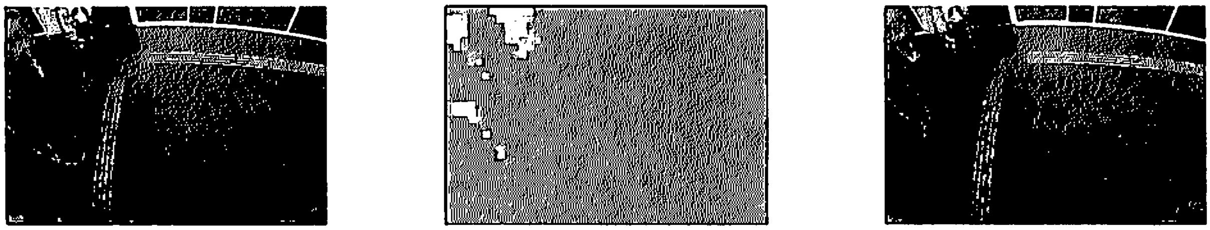

Fig. 6. Deletion of a class: Left, Result of computation in the previous frame; Center, Differential image; Right, Result of computation in the target frame.

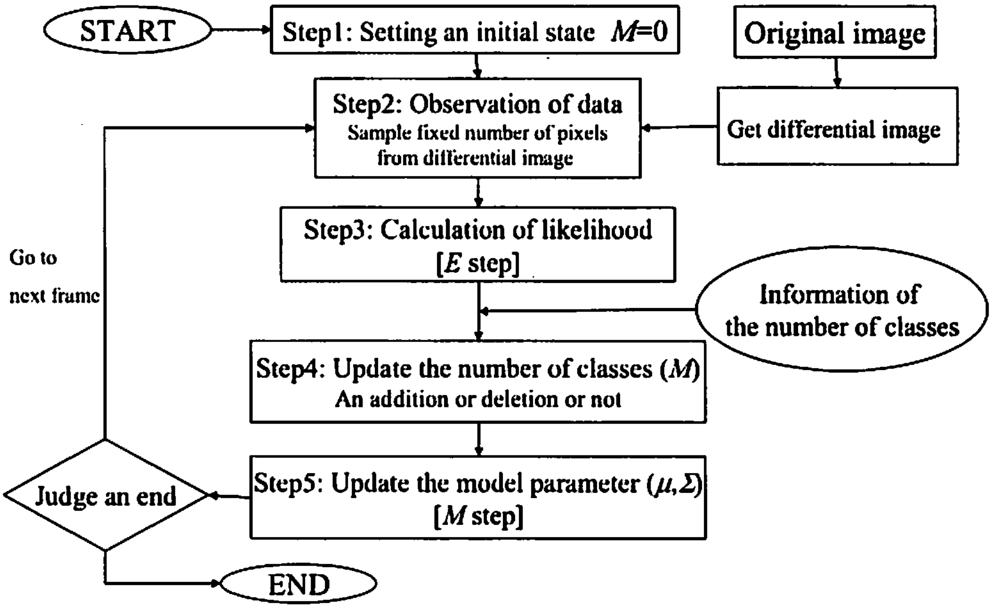
START
Step1: Setting an initial state M=0
Original image
Step2: Observation of data
Sample fixed number of pixels from differential image
Get differential image
Step3: Calculation of likelihood [E step]
Go to next frame
Information of the number of classes
Step4: Update the number of classes (M)
An addition or deletion or not
Judge an end
Step5: Update the model parameter (μ,Σ) [M step]
END

Fig. 7. Flow chart for the proposed method.

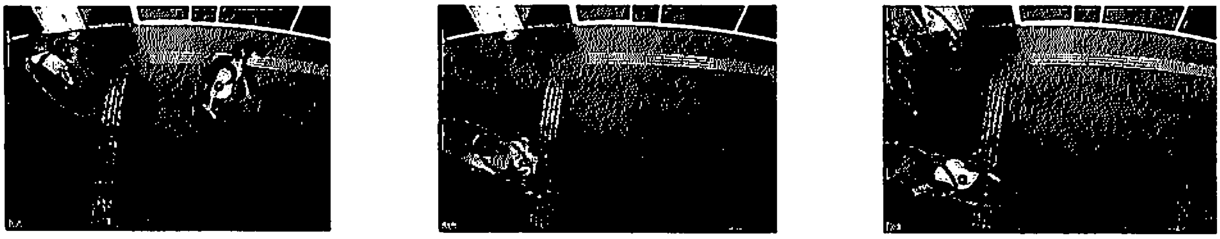

Fig. 9. Results of an experiment.

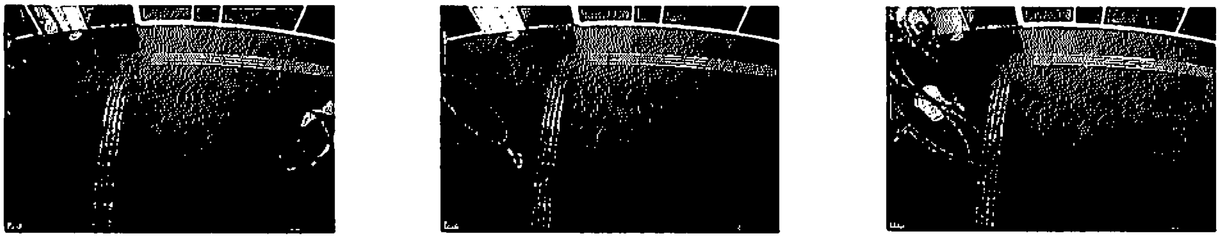

Fig. 10. Failure samples.