3

Study on Data-driven Methods for Image and Video Understanding

Tatsuya Yamazaki
National Institute of Information and Communications Technology
Japan

1. Introduction

Owing to progress of broadband Internet communication infrastructure, people can enjoy multimedia services. Among the services, images and videos are making an appeal and people desire to find out what they want to see. But they do not always make a success of searching and it sometimes takes plenty of time to reach their targets. A solution is to attach information tags according to the image or video content. Since image and video submission into Internet is increasing day by day, manual tag attachment is almost impossible. Development of automatic tag attachment is an urgent theme for future Internet service.

Another imaging technology in the real world is environmental cameras. The environmental cameras mean the cameras set in the environment such as a ceiling or a street corner. One can easily imagine surveillance cameras as an example of the environmental cameras and sometimes they are considered to be identical. In this paper, I want to segregate them because the purpose of the environment cameras is not only to keep a watch on accidents or crimes but to understand peoples' situations and behaviours. Namely the surveillance cameras are a subset of the environmental cameras.

For both of the above cases, the key point is understanding image and video. It is one of the issues that have been discussed for a long time and related with the artificial intelligence technology. Therein image clustering and object extraction are most essential technologies.

Image clustering means to divide an image into several segmented regions in a way of unsupervised, which was used for object extraction (Meier & Ngan, 1998), image compression (Kunt, 1988), or image categorization. Since it is hard to say how many regions are included in an image in general, there are a few studies that estimated the number of regions. Usually the number of regions has been assumed to be known. As a previous research in which the number of region was estimated, Won and Derin (Won & Derin, 1992) proposed to use the Akaike Information Criterion to determine the number of regions. But these are model-based approaches, that is, how to select a suitable model is important.

Regarding with the environmental cameras, there have been previous works to extract moving object in order to capture human behaviours. Satake and Shakunaga (Satake & Shakunaga, 2004) proposed an appearance-based condensation tracker, which is composed of a condensation tracker and a sparse template matching method to detect the movement of people with a camera. The template-based condensation tracker is stabilized for tracking even in the case of object occlusion. Thonnat and Rota (Thonnat & Rota, 1999) used low-

level image processing techniques to detect and track mobile objects. Their aim was rather to understand images, namely, to generate alarms automatically for operators when interesting scenarios had been recognized by the system. In both of the above works, they used only one camera. On the contrary, Matsuyama (Matsuyama, 1999) proposed a protocol for negotiation among multiple environmental cameras. Because of the protocol, they could synchronize several cameras in real time and grab the human behaviours more smoothly and more seamlessly.
As I described in the above paragraphs, automatic processing of image and video media should be deployed more in future Internet services for the information explosion era. Technically adaptation of processing based on the observed data, which is called data-driven, is necessary. Therefore, in this paper, data-driven image clustering and object detection methods are proposed.

2. Data-driven Image Clustering

In this section, a data-driven clustering algorithm for colour images is proposed based on a multi-dimensional histogram. A statistical model is introduced and the image data is assumed to be derived from a mixture of multi-variant distributions.

2.1 Multi-dimensional histogram

Although the R-G-B colour space is assumed to make a discussion simple, the proposed method can be applied to other colour spaces. The R-G-B colour image data is represented as $\boldsymbol{y}=(\boldsymbol{y}_1, \dots, \boldsymbol{y}_{Np})$, where $\boldsymbol{y}_i=(y_i^R, y_i^G, y_i^B)$ (i=1, … , N_p) is the observed element at the i-th pixel. N_p is the total number of pixels, y_i^X is a scalar value observed on the X plane at the i-th pixel, and y_i^X is assumed to range from G_{min} to G_{max}, where X is R, G, or B.
The multi-dimensional histogram is formed easily. First, distribute the observed data into the R-G-B color space. Second, construct a complete set of nonoverlapping intervals, called bins, by dividing the cube ($[G_{min}, G_{max}]^3$) equally with a width h. Finally, count the number of elements in each bin. Fig. 1 shows a construction of a multi-dimensional histogram with G_{min}=0. The histogram width h is an important parameter, and relates to the data distribution. When the data comes from a single density, several rules have been proposed to determine h. When the data are obtained a mixture of densities and the number of densities, N_c, is unknown, it is difficult to determine h. The multi-dimensional histogram of the colour image data deal with in this study corresponds to the latter case.

2.2 Data-driven clustering algorithm

It is assumed that there are a set of candidates for the histogram width, that is, $\{H=h_1, h_2, \dots, h_C\}$. Here, a novel algorithm is proposed to determine h and N_c.

Step 1) Select a candidate h_j (j=1, … , C) from H. Construct a histogram with a width h_j. Select the bins that have at least one element in the histogram, and sort them in the order of the number of elements in each bin. The sorted bins are numbered in the order of the number of elements as $b_1^j, b_2^j, \dots, b_{nz}^j$, where b_{nz}^j is the number of bins having at least one element. The k-th bin, b_k^j, has n_k^j elements (Fig. 2).

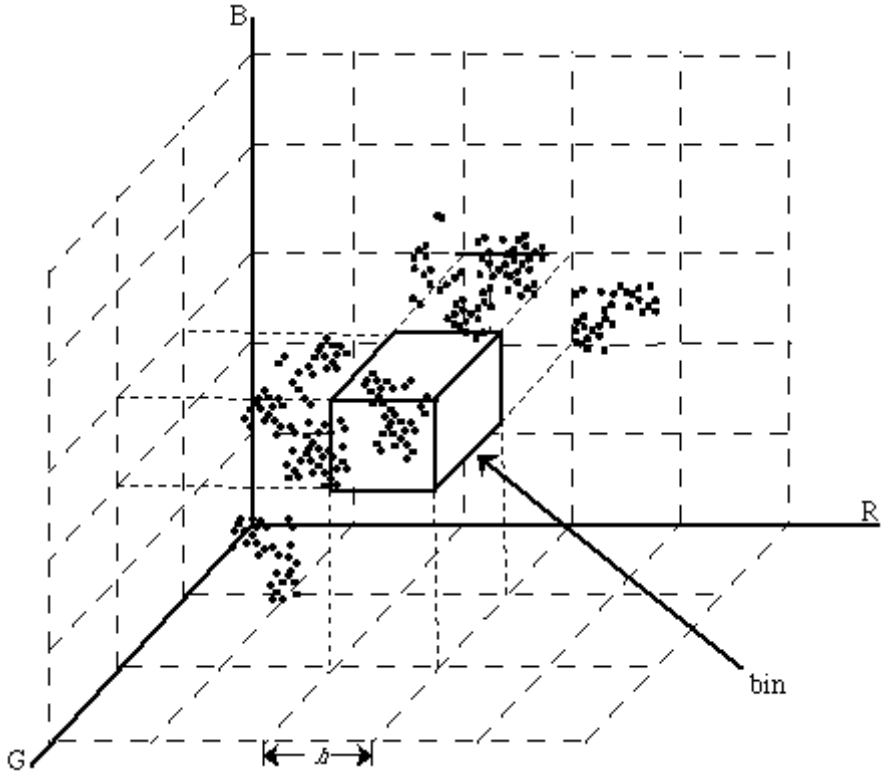

Fig. 1. Multi-dimensional histogram

Step 2) If $n_k^j < n_{cut}^j$, then b_k^j is removed. n_{cut}^j is a threshold calculated as

$$\mathrm{n_{cut}} = \alpha_{\mathrm{h_j}} \frac{\mathrm{b_{nz}}}{\mathrm{N_p}} \tag{1}$$

where $\alpha_{\mathrm{h_j}}$ is a control parameter to be determined for each h_j heuristically. n_{cut} is the threshold that divides explicitly insignificant bins. Eventually b_{cut}^j bins remain.

Step 3) Extract the top b_{sig}^j bins as significant bins from b_{cut}^j bins, where b_{sig}^j is determined as follows.

$$\mathrm{b_{sig}^j} = arg \max_{\mathrm{k=1,\ldots,b_{cut}^j}} \mathrm{Cr_{sig}}(k) \tag{2}$$

$$\mathrm{Cr_{sig}}(k) = \sum_{\mathrm{l=1}}^{\mathrm{k}} \frac{\sum_{\mathrm{m=1}}^{\mathrm{b_{cut}^j}} \mathrm{n_m^j}}{\mathrm{n_l^j}} - \frac{\mathrm{b_{cut}^j}}{\mathrm{k}} \tag{3}$$

$Cr_{sig}(k)$ is a criterion to select significant bins that include as many elements as possible considering suppression of the number of selected bins.

Step 4) If b_{sig}^{j} for every histogram-width candidate h_j is calculated, then go to Step 5. Otherwise, go to Step 1 and pick up another candidate whose b_{sig}^{j} has not been calculated.

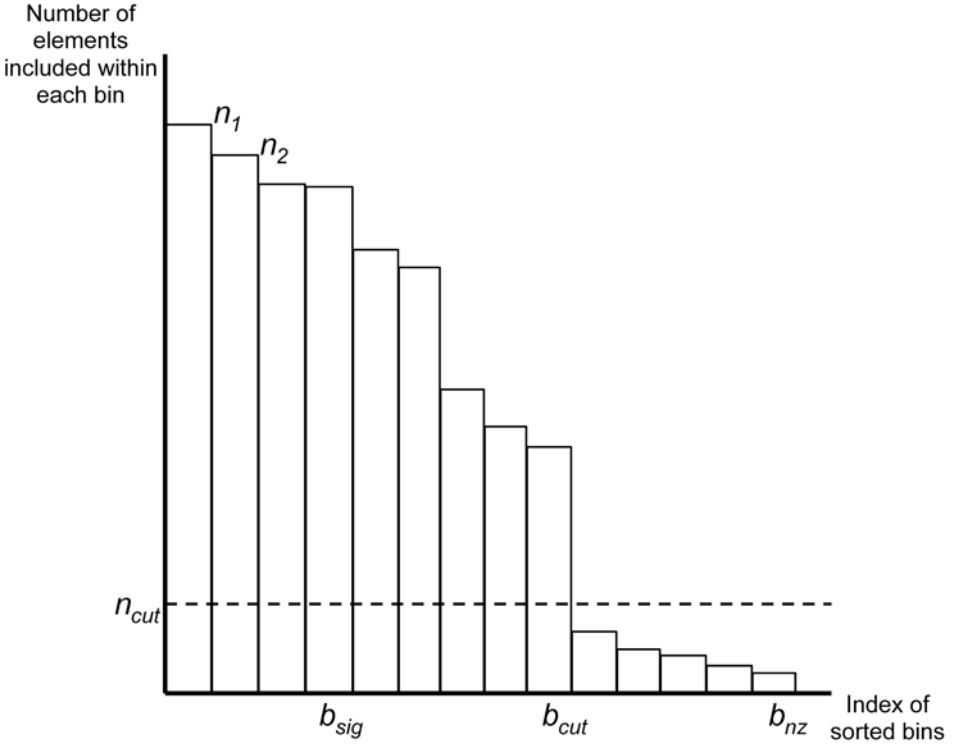

Fig. 2. Sorted frequency distribution of elements within each bin

Step 5) Calculate the optimal histogram width h^* as

$$h^* = arg \max_{h_j \in H} \sum_{l=1}^{b_{sig}^{j}} n_l^j \tag{4}$$

where b_{sig}^{j} corresponds to h_j, and b_{sig}^{j} corresponding to h^* is denoted as b_{sig}^{*}.
Consequently h^* is selected from H as the optimal width from the perspective of extracting as many significant elements as possible to compute the distribution statistics.

The significant b_{sig}^{*} bins are determined beforehand without considering the mutual relationships. Therefore, the adjacent bins that are supposed to belong to the same density must be merged to determine the final cluster count N_c. The clustering criterion is as follows. Calculate the average $\left(\mu_k^R, \mu_k^G, \mu_k^B\right)$ and the standard deviation $\left(\sigma_k^R, \sigma_k^G, \sigma_k^B\right)$ for the k-th bin (k=1, ..., b_{sig}^{*}).

Select any two bins, say b_l and b_m, from b_{sig}^* bins. If $\left|\mu_l^R - \mu_m^R\right| < \beta\overline{\sigma^R}$, $\left|\mu_l^G - \mu_m^G\right| < \beta\overline{\sigma^G}$ and $\left|\mu_l^B - \mu_m^B\right| < \beta\overline{\sigma^B}$, then the two bins belong to the same cluster, where $\overline{\sigma^X} = \sum_{i=1}^{b_{sig}^*} \sigma_i^X / b_{sig}^*$ for X=R, G, or, B and β is a parameter. Finally, cluster the b_{sig}^* bins into N_c clusters.

2.3 Proposed algorithm application to real data

The proposed algorithm was applied to real image data. Fig. 3 shows one of the original images, which is called the "Lady with a rose" image in this paper. The image size is 480 × 480. Although it is shown in grey scale, the original colour space is the RGB colour space. The histogram width candidates set used in the algorithm is H={4, 8, 16, 32}. The values of α_{hi} corresponding to each histogram width are α_4 =1.0, α_8=0.7, α_{16}=0.35 and α_{32}=0.05. In this experiment, β is set to 0.0; this means that the merging process is skipped.

Applied the proposed algorithm, the histogram width was determined to be 32 and the number of clusters was 6 finally. These values were determined in a data-driven way.

Then, the statistics of each cluster were computed by the minimum distance method, and the conventional maximum likelihood method with the estimated statistics, under the normal distribution assumption, was performed to obtain the segmented image. The final result followed by the 3×3 mode filtering operation is shown in Fig. 4.

Fig. 3. Original image "Lady with a rose"

Fig. 4. Segmentation result of "Lady with a rose"

The final estimates of the means and the proportions are shown in Table 1.

	Cluster 1	Cluster 2	Cluster 3
R	37.4	17.7	73.5
G	45.7	12.1	84.0
B	52.1	6.6	104.5
Proportion	0.282	0.166	0.172
	Cluster 4	Cluster 5	Cluster 6
R	29.3	83.1	110.3
G	37.9	60.0	100.8
B	41.1	47.3	103.9
Proportion	0.132	0.157	0.091

Table 1. Final parameter estimates of the means and the proportions for "Lady with a rose"

3. Data-driven object detection

In the real world, cameras are becoming popular to detect and track moving objects not only for surveillance or security but also for digital signage or spoken dialogue systems. The background subtraction method, which can be used to detect objects moving in the foreground by determining the difference between the current frame and an image of the scene's static background, is still one of the useful methods to detect moving objects in video sequences. Although the background subtraction method is a simple and effective method to detect moving objects, it occasionally suffers from illumination changes and unexpected background changes such as shadow. To improve the original background subtraction method, we propose that knowledge application and data-driven parameter adaptation techniques be adopted.

3.1 Detection of People by Background Subtraction Method

In order to cope with the illumination changes, we adopt the normalized distance method. Here, a unit vector is defined as the projection onto the unit sphere of a vector whose elments are the intensity values of pixels in a target region. The normalized distance is defined as a distance between two unit vectors. Let $\boldsymbol{\tau}$ and $\boldsymbol{\beta}$ as vectors consiting of intensity values of pixels in an observed image and in a background image respectively. Then the distance $\boldsymbol{\delta}$ and normalized distance $\boldsymbol{\delta}'$ are shown as follows,

$$\boldsymbol{\delta} = |\boldsymbol{\tau} - \boldsymbol{\beta}| \tag{5}$$

$$\boldsymbol{\delta}' = \left| \frac{\boldsymbol{\tau}}{|\boldsymbol{\tau}|} - \frac{\boldsymbol{\beta}}{|\boldsymbol{\beta}|} \right| \tag{6}$$

Supposing that we process image frames during a period of T_{int}. We calculate $\boldsymbol{\delta}$ for each frame in T_{int}, then $Max(\boldsymbol{\delta})$, $Min(\boldsymbol{\delta})$, and $Ave(\boldsymbol{\delta})$ are calculated as maximum value, minimum value, and averaged value of $\boldsymbol{\delta}$. In the same way, $Max(\boldsymbol{\delta}')$ and $Min(\boldsymbol{\delta}')$ are calculated as maximum value and minimum value of the normalized distance $\boldsymbol{\delta}'$. Consequently, the following three discriminant functions are defined,

1. discriminant function for scene change
   $\max \boldsymbol{\delta} - \min \boldsymbol{\delta} > Th_{scene}$

2. discriminant function for background change
   Ave $\boldsymbol{\delta} > Th_{bs}$

3. discriminant function whether environment change or illumination change
   $\max \quad \boldsymbol{\delta}' - \min \quad \boldsymbol{\delta}' > Th_{ill}$

where Th_{scene}, Th_{bs}, and Th_{ill} are thresholds to be determined.
Using these discriminant functions, whether there is a moving object detection or a background updating in T_{int} can be judged as follows:

- If both (3) and (5) are true, there is a scene change by a moving object.
- If (3) is true and (5) is false, there is a scene change by an illumination change.
- If (3) is false and (4) is true, there is a background change.
- If both (3) and (4) are false, there is nothing. It is a normal background.

A challenging point in this method is adaptively setting the threshold value to differentiate foreground objects from the background image in spite of environmental changes.
To determine the threshold value, Wren et al. (Wren et al., 1997) modeled the background using a Gaussian distribution and estimated the parameters adaptively. Grimson et al. (Grimson et al., 1998) also set up parameters according to the statistical analysis of training samples of the background images. Stauffer and Grimson used a mixture model of Gaussian distributions of the images to cope with multimodal background distributions (Stauffer & Grimson, 1999).

3.2 Knowledge Application and Parameter Adaptation

We apply two techniques to the original background subtraction method in order to cope with unexpected moving objects and adaptive threshold parameter setting. Our aim is to detect and track people as moving objects. There are, however, other unexpected moving objects in the scene, such as an automatic door. To avoid detection of such an unexpected moving object, we introduce knowledge about special spots as the first technique. The positions of special spots are assumed to be known and masking is applied not to detect the unexpected moving object. This is a simple but effective technique.

To set the threshold values adaptively, we introduce a kind of steepest descent method as the second technique. The algorithm is depicted in Fig. 5.

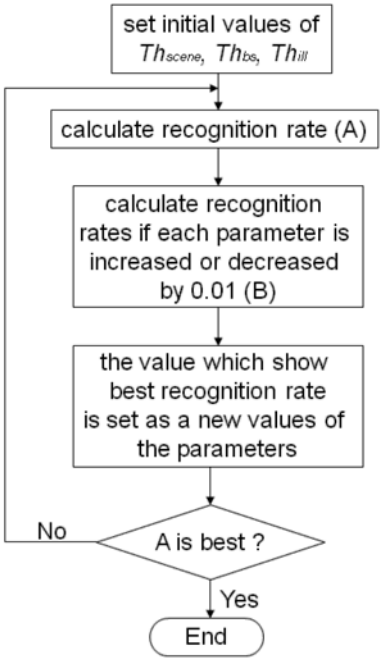

Fig. 5. Flowchart of threshold adaptation by steepest descent

In the first step of the algorithm shown in Fig. 5, initial values of Th_{scene}, Th_{bs} and Th_{ill} are set. Then the recognition rate A with the current threshold values that are the same as the initial values at the very first stage of the algorithm. After the threshold values are increased or decreased by a small value, the new recognition rates B are calculated. 0.01 is set as the small value in Fig. 5. B is a set of the recognition rates because increasing and decreasing of each threshold value are tried. A is compared with the best recognition rate in B. If A is superior to the best recognition rate, the algorithm is terminated. Otherwise, the best threshold values that correspond to the best recognition rate are substituted to the current threshold values and the same steps are carried out.

3.3 Experimental results at the real world test bed

We constructed such a test bed in the entrance of our research laboratory. In the ceiling of the test bed, there are five cameras, and the area covered by the cameras is the meshed region in Fig. 6. The camera can take a 768 × 494 pixel image and has remote pan, tilt, and zoom functions.

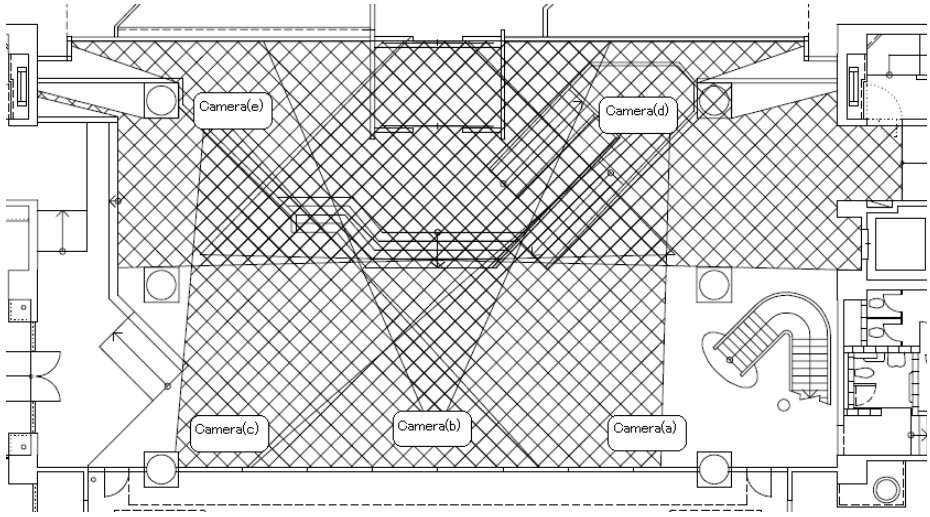

Fig. 6. Area monitored by cameras

In the test bed, we tried to detect moving objects using camera images and the background subtraction method. In the background subtraction method, first of all, we selected the initial image without any moving objects as the background image to be subtracted. Then, the difference between the current and the background images is calculated and pixels that have a difference larger than the threshold are registered as candidate pixels of an image of moving objects. This difference calculation is operated for each small image block of 80 × 60 pixels. The judgment described in the previous subsections is applied. The adjacent small image blocks of candidate pixels are merged into a larger image block. To track the moving objects, a two-dimensional histogram with hue and saturation values in an HSV color space is constructed to calculate the correspondence between objects in the current and previously captured images. When the difference between two images in the two-dimensional histogram is smaller, the probability that objects belong to the same object is higher.

We applied the above background subtraction method to a ten-second video captured in an actual situation. The number of frames captured from each camera was different because the time consumed for image compression was different.

Two experiments were carried out. The first experiment was moving object detection by the background subtraction method with knowledge application only. The second one was moving object detection by the background subtraction method with parameter adaptation as well as knowledge application. The first and the second are referred as Experiment I and Experiment II respectively. The applied knowledge is that the position of the automatic door is known.

In Experiment I, the threshold parameters were fixed as Th_{scene}=0.10, Th_{bs}=0.25 and Th_{ill}=0.05. One result of detecting a moving object is presented in Fig. 7. The people in the image

should be recognized as moving objects, and rectangles are drawn as a result. Two larger rectangles (shown in red) are hand-made markings that indicate correct answers. Several smaller rectangles in the left larger rectangle (shown in yellow) indicate detected results obtained by the background subtraction method. In the right larger rectangle, no object was detected by the background subtraction method. We call the larger rectangles ground truth rectangles and the smaller rectangles are called detected rectangles.

Fig. 7. Result of detecting moving objects in Experiment I

Although an identical match between ground truth and detected rectangles is desirable, detected rectangles are almost always included in or overlapped on ground truth rectangles. Here, we define two kinds of error: the type one error and the type two error. The type one error is that in which no detected rectangle is drawn where there was a ground truth rectangle. The type two error is that in which detected rectangles appeared where there was no ground truth rectangle. The total error rate can be calculated by averaging these two types of error. The rates of occurrence of two types of error and the total error rate are shown in Table 2 for cameras (a) - (e). The positions of the cameras are shown in Fig. 6.

	Type one error rate (%)	Type two error rate (%)	Total error rate (%)
Camera(a)	29.38	0.32	14.85
Camera(b)	54.35	27.94	41.15
Camera(c)	28.39	0.31	14.35
Camera(d)	10.23	16.19	13.21
Camera(e)	10.06	8.38	9.22

Table 2. Error rates with knowledge application and fixed parameters

Next, we applied threshold parameter adaptation presented in Fig. 5 as well as the knowledge application. This is Experiment II. The initial threshold parameters were set as Th_{scene}=0.10, Th_{bs}=0.25 and Th_{ill}=0.05. One moving object detection result with the parameter adaptation is presented in Fig. 8, that corresponds to Fig. 7. By comparing two images, it is found that the person in the right side was detected in Experiment II, who was missed in Experiment I. The rates of occurrence of two types of error and the total error rate are shown in Table 3.

Fig. 8. Result of detecting moving objects in Experiment II

	Type one error rate (%)	Type two error rate (%)	Total error rate (%)
Camera(a)	14.69	0.27	7.48
Camera(b)	17.39	0.00	8.97
Camera(c)	19.36	1.97	10.67
Camera(d)	1.14	14.70	7.92
Camera(e)	10.06	0.08	5.07

Table 3. Error rates with knowledge application and parameter adaptation

Almost all error rates were improved. Especially improvement for camera (b) was splendid. It can be considered that the reason is owing to knowledge application. As the result of parameter adaptation, the final obtained parameters, Th_{scene}, Th_{bs} and Th_{ill}, are shown in Table 4.

	Th_{scene}	Th_{bs}	Th_{ill}
Camera(a)	0.10	0.21	0.04
Camera(b)	0.10	0.20	0.04
Camera(c)	0.10	0.21	0.04
Camera(d)	0.07	0.23	0.04
Camera(e)	0.12	0.26	0.04

Table 4. The final threshold parameters in Experiment II

4. Conclusion

Among multimedia, the roles of image and video media are becoming more important both in the cyber and real worlds. The cyber world means the information world structured by computer networks such as Internet. Videos and images are accumulated in the cyber world and the users are wandering to search for what they want. Also in the real world, cameras are becoming popular to collect the users' and environmental information. How to analyse these more efficiently is one of the issues to be solved urgently.

Image and video understanding has been studied and several approaches have been developed. In the parametric approaches, how to set the parameters is a challenging problem. In this paper, data-driven parameter adaptation was applied to image clustering and object extraction for still and video images. Although the experimental results were limited, definite improvement has been attained.

In the future, it is desired to extract contextual information from image and video media by applying image and video understanding techniques including the methods proposed in this paper. It must contribute to realize a personalized, adaptive, situation-aware service in a ubiquitous network society.

5. References

Meier, T. & Ngan, K.N. (1998). Automatic Segmentation of Moving Objects for Video Object Plane Generation. *IEEE Trans. on Circuits Syst., Video Technol.,* Vol. 8, No. 5, (Sept. 1998) pp. 525-538

Kunt, M. (1988). Progress in High Compression Image Coding. *Int. J. Pattern Recognition and Artificial Intelligence,* Vol. 2, No. 3, (1988) pp. 387-405

Won, C.S. & Derin, H. (1992). Unsupervised Segmentation of Noisy and Textured Images Using Markov Random Fields. *CVGIP: Graphical Models and Image Processing,* Vol. 54, No. 4, (July 1992) pp. 308-328

Satake, J. & Shakunaga, T. (2004). Multiple target tracking by appearance-based condensation tracker using structure information, *Proceedings of the 17th International Conference on Pattern Recognition (ICPR2004)*, Vol.1ap, No. We-ii, pp. 537-540, 2004

Thonnat, M. & Rota, N. (1999). Image Understanding for Visual Surveillance Applications, *Proceedings of Third International Workshop on Cooperative Distributed Vision (CDV-WD'99)*, No. 3, pp. 51-82, Nov. 2004

Matsuyama, T. (1999). Dynamic Memory: Architecture for Real Time Integration of Visual Perception, Camera Action, and Network Communication, *Proceedings of Third International Workshop on Cooperative Distributed Vision (CDV-WD'99)*, No. 1, pp. 1-30, Nov. 2004

Wren, C.; Azarbayejani, A.; Darrell, T. & Pentland, A. (1997). Real-time Tracking of the Human Body, *IEEE Trans. on Patt. Anal. and Machine Intell.,* Vol. 19, No.7, (1997) pp.780-785

Grimson, W.E.L.; Stauffer, C.; Romano, R. & Lee, L. (1998). Using adaptive tracking to classify and monitor activities in a site, *Proceedings of 1998 Conference on Computer Vision and Pattern Recognition (CVPR '98)*, pp. 22-29, 1998

Stauffer, C. & Grimson, W.E.L. (1999). Adaptive background mixture models for real-time tracking, *Proceedings of 1999 Conference on Computer Vision and Pattern Recognition (CVPR '99)*, pp. 246-252, 1999

Multimedia
Edited by Kazuki Nishi

ISBN 978-953-7619-87-9
Hard cover, 452 pages
Publisher InTech
Published online 01, February, 2010
Published in print edition February, 2010

Multimedia technology will play a dominant role during the 21st century and beyond, continuously changing the world. It has been embedded in every electronic system: PC, TV, audio, mobile phone, internet application, medical electronics, traffic control, building management, financial trading, plant monitoring and other various man-machine interfaces. It improves the user satisfaction and the operational safety. It can be said that no electronic systems will be possible without multimedia technology. The aim of the book is to present the state-of-the-art research, development, and implementations of multimedia systems, technologies, and applications. All chapters represent contributions from the top researchers in this field and will serve as a valuable tool for professionals in this interdisciplinary field.

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821