First Korea-Japan Workshop on Pattern Recognition.
2006/11/23-24, Jeju Island, Korea.

THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

TECHNICAL REPORT OF IEICE.

Scene Clustering with Multiple Non-Calibrated Cameras

Takashi NISHIZAKI†, Kouji KANARI†, Yoshinari KAMEDA†, and Yuichi OHTA†

† Graduate School of Systems and Information Engineering, University of Tsukuba 1-1-1 Tennoudai, Tsukuba, Ibaraki, 305-8573 Japan

E-mail: †{tanishi,kanari}@image.esys.tsukuba.ac.jp, ††{kameda,ohta}@iit.tsukuba.ac.jp

Abstract We propose a new scene clustering method for multiple fixed cameras that spread widely over daily scenes. The cameras do not need to be calibrated or to overlap with each other. Our new method utilizes regions of interest (ROIs) in multi-view images extracted using our previously described method. The scene clustering method discriminates between all possible events by checking snapshots observed in the extracted ROIs in the multiple views of the cameras. The advantage of our method is that it only requires a set of video data, and does not need to know the camera locations. We implemented a preliminary surveillance camera system with 35 cameras, and experiments confirmed the utility of the proposed method.

Key words ROI, surveillance camera, uncalibrated cameras, multiple cameras, event clustering

1. Introduction

Recently, there has been increasing social demand for observation and recognition of usual and/or unusual events automatically with surveillance camera systems in various environments. This is helpful for security, traffic monitoring, nursing care, etc., so the number of surveillance cameras in daily scenes will increase in the future. Advanced examples include the Aware Home [1], the Ubiquitous Home [2], etc. [3] [4].

Basic vision approaches for multiple cameras usually require camera calibration or at least camera location databases, such as "which cameras are filming the same area" or "which cameras are in the same room and which are not." However, as the number of cameras increases, their calibration becomes difficult, and it will not be an easy task to manage the camera location database. In addition, when cameras are installed in daily scenes, target objects are often large, and sometimes only part of the object can be found in the images. A typical snapshot of a multiple camera images is shown in Fig. 1. Therefore, a new vision approach for scene clustering that does not depend on this type of information is required.

Stauffer *et al.* proposed a method to build a correspondence model for the entire set of cameras by estimating a planar tracking correspondence [5]. They assume that camera regions overlap with each other and the trajectory is obtained accurately. Tieu *et al.* tackled the problem of multi-camera geometric correspondence [6]. In their approach, it

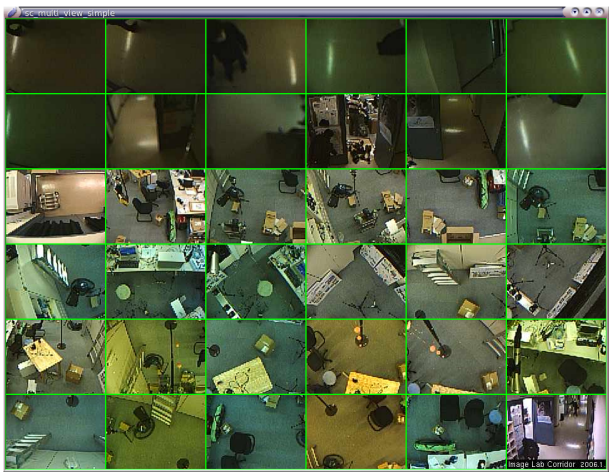

Figure 1 Example of Multi-view videos

is assumed that two cameras are connected; if an object is departing from the imaging area of one camera, it should be observed by the other camera although their method does not require overlapping camera regions. Both methods [5] [6] utilize trajectories and so the results are dependent on the quality of the trajectory, which is sometimes not good in daily scenes because objects are too close to the cameras.

Hammadi *et al.* proposed a method to recognize living room activities for overhead tracking [7]. Foresti *et al.* proposed a method to discriminate events by the scale of danger for human tracking in a parking lot [8]. These types of event recognition [7] [8] use templates. The recognition is accurate and reliable to the extent that the templates can be prepared or defined.

To deal with typical recognition problems of complicated human and/or object activities, advanced methods have been proposed [9] [10] [11]. However, these methods are customized for a particular situation and so are not flexible for various environments.

In this paper, we propose a new clustering method that can be applied in a wide variety of situations. It does not have any assumptions about object size, object color, or object behavior. It works even in cases where no accurate trajectories can be obtained. It requires neither preparation of templates nor limitations regarding every possible event in a scene. Another advantage of our method is that it works with non-calibrated cameras without requiring any information about the camera layout. The cameras can be set in an arbitrary way in the scene. Our method relies only on global analysis of the video data set obtained from multi-view cameras supplied to the system in advance.

The rest of the paper is organized as follows. In section 2., we present a brief review of the method to obtain compact expression of video data, which we proposed in [12] and show some experimental results obtained with our preliminary system. Section 3. explains the proposed method for scene classification with multiple non-calibrated cameras. In section 4., we show experimental results of scene clustering and describe experimental verification of the utility of our method. Concluding remarks and future prospects are presented in section 5..

2. ROI Extraction by Video Data Analysis

2.1 ROI Extraction

Our clustering method utilizes regions of interest (ROIs) in multi-view camera images. ROIs are regions in which it is valuable to see objects in motion. Suppose each video image is divided into R regions, there are C cameras, and the video dataset for ROI extraction is taken for a given time period T. We select ROIs among $C \times R$ regions if they are significant to describe the video dataset. The method described here uses PCA for the feature vectors of the video data. The ROI extraction method is described in detail in [12].

The extracted $N (N \leqq C \times R)$ ROIs can describe the given video dataset efficiently. Note that the time T should be set such that the video dataset can include human actions (events) that occur frequently in the scene. We denote the ROIs as a vector $\mathbf{v}(t)$ of N elements in the following equation. The module $v_i(t)$ represents feature quantity defining mean intensity in the foreground region taken at time t.

$$\mathbf{v}(t) = \{\ v_1(t),\ \cdots\ ,\ v_i(t),\ \cdots\ ,\ v_N(t)\ \}$$

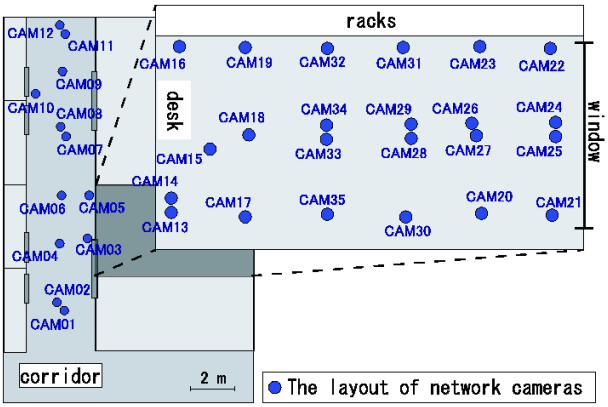

Figure 2 Camera Layout

2.2 Experiment

We set up an experimental scene and preliminary multi-camera system. A scene in an office is selected here. Thirty-five cameras are hung on the ceiling of a room and in the corridor next to the room. Fig. 1 shows a set of snapshots taken simultaneously. We used 12 of the 35 cameras due to space limitations of the paper. The layout of cameras is shown in Fig. 2.

We performed the experiment for a scene in which a subject is working in the room. The subject worked mainly at the desk in the center of the room, sometimes walked around, opened and closed the doors of the rack, moved some objects, and went out and came back through the corridor several times. We took video data for about 3 hours at an average of 7.6 fps. R was set to 64.

The extracted ROIs are shown in Fig. 3 and Fig. 4. The regions marked with bright frames are the selected ROIs. A total of 178 blocks were selected from 768 blocks in this case using 12 cameras. Therefore, only 23.3% of the area was worth checking for further scene analysis. In the case of 4 cameras (CAM 25, 26, 27, and 28), the number of ROIs was 87. Note that some ROIs used in the 12-camera case were not selected in the 4-camera case. This was because the selection was done so that the ROIs well described the whole video dataset based on PCA, and so the amount of motion detection was not counted directly.

The computation of ROI extraction was performed on a Pentium4 2.8 GHz PC with 1GB memory. The experiment with 12 cameras took 2157.85 seconds; 474.38 seconds to calculate the mean vector, 1677.75 seconds to obtain the variance-covariance matrix, and 5.72 seconds to conduct PCA and final ROI selection. In the case of 4 cameras, the total time was 1467.57 seconds, including 142.16 seconds, 1324.68 seconds, and 0.73 seconds for mean vector calculation, to obtain the variance-covariance matrix, and to conduct PCA and final ROI Selection, respectively.

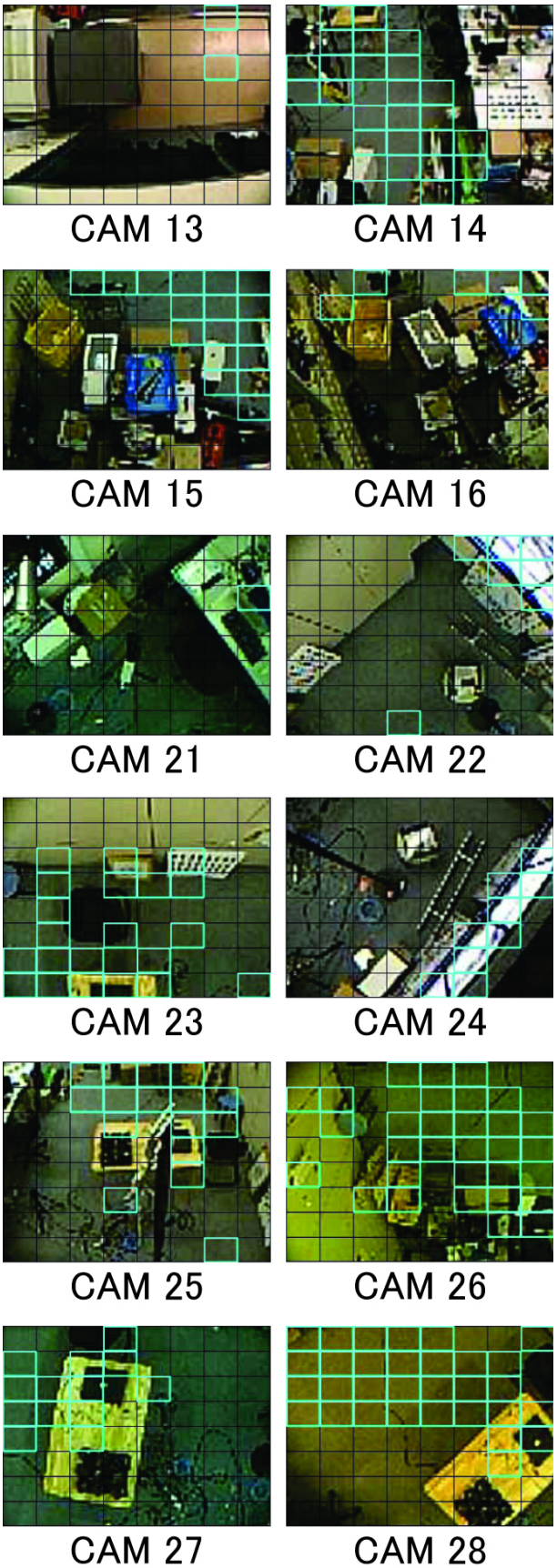

Figure 3 Resultant ROIs (12 cameras)

3. Scene Clustering

We defined scene clustering so as to make clusters of similar

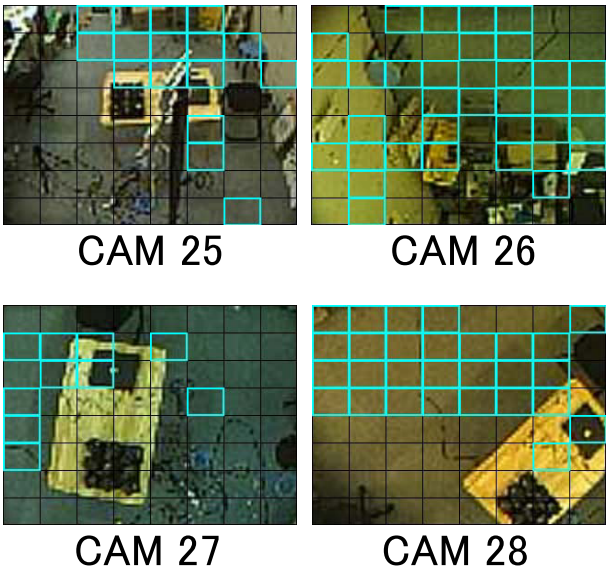

Figure 4 Resultant ROIs (4 cameras)

$\mathbf{v}(t)$ as video data are fed to the system on line. A member $\mathbf{v}(t)$ of a cluster corresponds to an instance of an event in the scene at time t, observed in the N ROIs. In a sense, a cluster will represent a certain event that appears similarly in the N ROIs. The discrimination distance between clusters is determined by preparatory off-line clustering of the video dataset used to extract ROIs.

3.1 Discrimination Distance

We do preparatory off-line clustering to determine the Euclidean distance between clusters to use it as a discrimination threshold in on-line clustering of the scene.

We apply K-means clustering for a set of $\mathbf{v}(t), (0 \leqq t \leqq T)$. K can be set arbitrarily as it defines the grain of the clusters of the scene classification. It is a good guess to set K as the number of types of event found in the scene over the time period T. After K clusters are obtained, we exploit the minimum distance between the clusters as the discrimination threshold D_{th} of the on-line clustering.

3.2 On-line Clustering

During on-line clustering, the input vector is given in the shape of $\mathbf{v}(t)$ with N elements. At the beginning of on-line clustering, we set no clusters in the space. On-line clustering is done by repeating steps 1–3 outlined below.

Step 1: Discrimination

Given a new input vector $\mathbf{v}(t_k)$, calculate the Euclidean distance to the center of each cluster and exploit the minimum distance among them. If the input vector $\mathbf{v}(t_k)$ is a null vector, wait until a non-zero vector is obtained.

Step 2: Cluster Update and Insertion

If the distance is smaller than D_{th}, the input vector $\mathbf{v}(t_k)$ is merged into the nearest cluster. Otherwise, a new cluster is inserted for the input vector.

Step 3: Cluster Reorganization

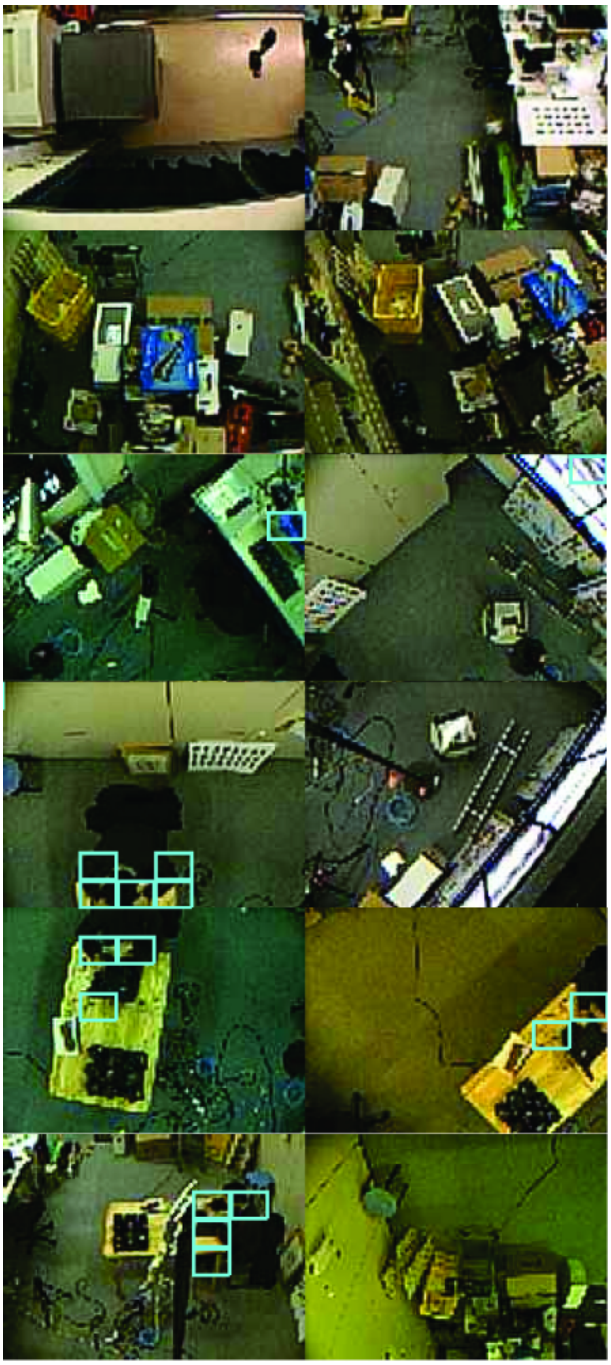

Figure 5 Major ROIs of Cluster α

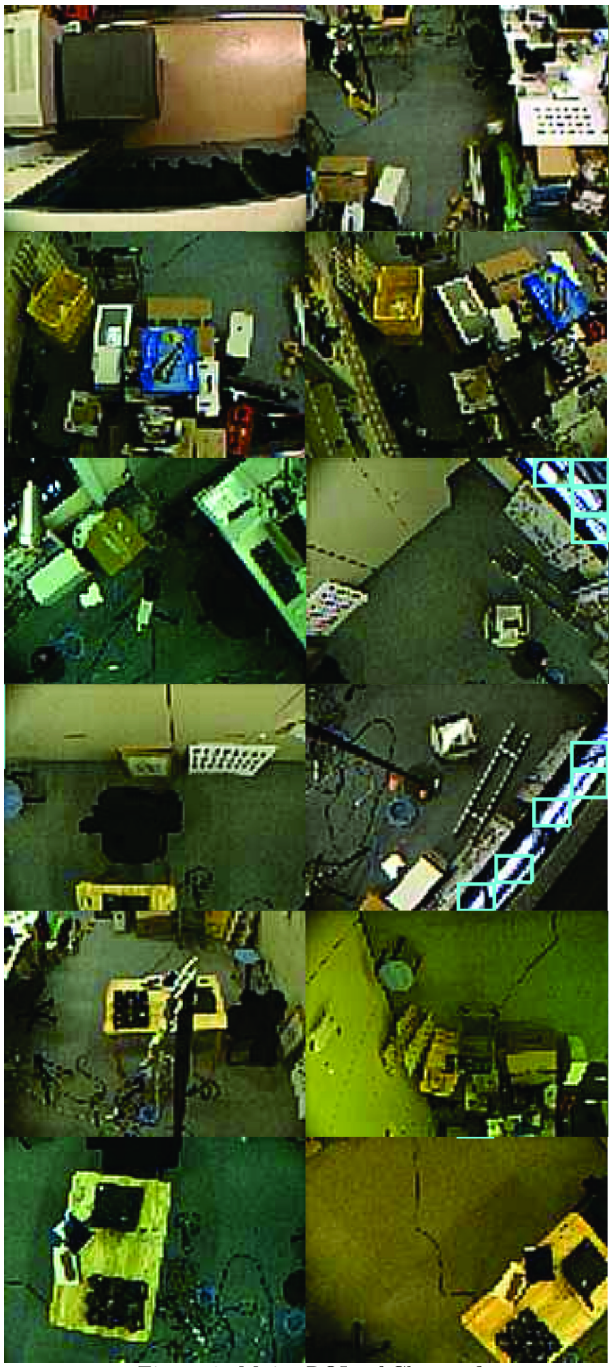

Figure 6 Major ROIs of Cluster β

As the center of the cluster that has obtained the new input vector is moved accordingly, the members of the cluster and the members of the adjacent cluster should be re-examined. Therefore, for these members, the Euclid distances to the cluster centers are re-calculated and they are re-classified into the clusters.

4. Experimental Results

We conducted a scene clustering experiment on our preliminary system. We obtained the discrimination distance D_{th} by setting $K = 20$. In this study, we fed the same video dataset of about 3 hours used in ROI extraction and determination of D_{th} in the experiment although it can also work in real-time. Among the 137,351 input vectors, 44,122 were none-zero vectors in the experiment.

We used the same PC for D_{th} determination and on-line clustering. It took 1341.77 seconds to calculate D_{th}. The on-line clustering for the video dataset took 2457.49 seconds. Looking at video data taken for 10800 seconds (about 3 hours), this shows that our system can conduct on-line clustering in real-time.

We obtained 762 clusters by on-line clustering at the end of the three-hour data. Three examples of the resultant clusters are shown in Fig. 5, Fig. 6, and Fig. 7. Fig. 5 corresponds to

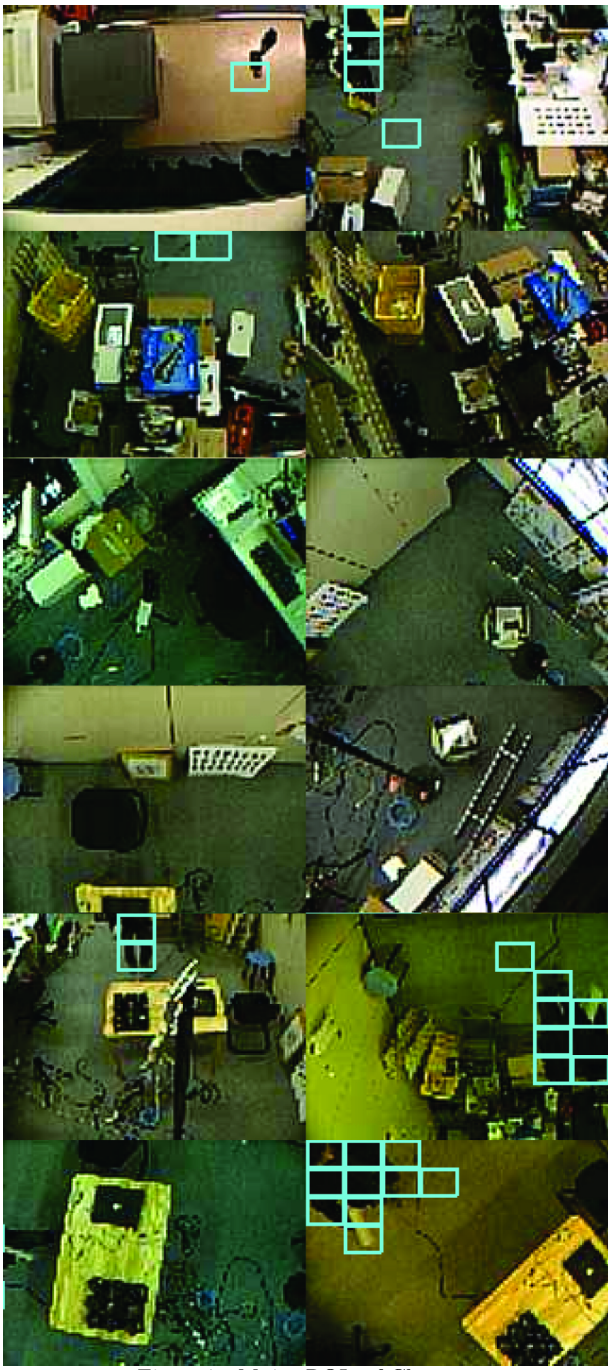

Figure 7 Major ROIs of Cluster γ

a cluster of 665 members, which we call cluster α. The cluster shown in Fig. 6 has 2,422 members and we call it cluster β. And the cluster of Fig. 7 has 2 members and we call it cluster γ. The bright frames in the figures indicate that a large amount of the foreground region is found there. Note that the images in Fig. 5, Fig. 6 and Fig. 7 are snapshots of the scene at a certain time, so the images do not contain foreground objects.

Cluster α represents a scene showing desk work and cluster β represents a changing illuminant environment in the uninhabited room. Cluster α, which corresponds to Fig. 5, can be confirmed to show the subject at a wooden desk working on a laptop on the desk. Cluster β, which corresponds to Fig. 6, confirmed the foreground regions around the windows, and we confirmed that the intensity showed wide variations in the investigation. This is appropriate because we conducted the experimentation in the early-evening when there was a great deal of sunlight fluctuation. As described previously, we confirmed that our method is capable of recognizing variance in the environment as variance in the scene. The number of members in cluster α was large, but smaller than that in cluster β. The number of members in cluster β was the largest, and so the changing illumination environment in the uninhabited room was often observed. In fact, the subject went out and came back through the corridor several times and worked in locations other than the experimental room, so he did not remain the whole time in the room. On the other hand, the number in cluster γ was extremely small. This cluster represents a scene in which the subject moved around the desk in the room. Thus, our method is not good at summing up members in such a cluster. This is because our method doe not consider spatial-temporal consistency, and so clusters with "moving" cannot be bundled together.

As shown in these results, our method allows clustering of possible events in the scene without any knowledge of camera layout or behaviors of the object. However, as the discussion in this paper is subjective, further experiments are required to validate our method.

5. Conclusion

We propose a new method of scene clustering in a daily scene under an environment in which a number of non-calibrated fixed surveillance cameras are being operated.

Our method first divides all camera images into small blocks and selects significant blocks as ROIs to observe the events in the scene. After estimation of the discrimination distance by the off-line process of K-means clustering, on-line clustering is performed.

We implemented a preliminary system with 35 cameras and showed the experimental results obtained with 12 non-calibrated cameras in an office scene. In the experiments, we succeeded in clustering the events in the scene.

Future studies should be performed to examine the relevance of the clustering algorithm. In addition, it is necessary to verify the validity of clusters and the proposed method with a wide variety of events and scenes. Longer-term experiments are also necessary for validation.

References

[1] Abowd, G. A. Bobick, I. Essa, E. Mynatt, and W. Rogers: The aware home: Developing technologies for successful aging. Workshop held in conjunction with American Association of Artificial Intelligence (AAAI) Conference 2002, July 2002.

[2] T. Yamazaki, H. Ueda, A. Sawada, Y. Tajika and M. Minoh: Networked appliances collaboration on the ubiquitous home. 3rd International Conference On Smart homes and health Telematic (ICOST 2005), 15:135–142, July 2005.

[3] G. C. de Silva, T. Yamasaki, T. Ishikawa and K. Aizawa: Video handover for retrieval in a ubiquitous environment using floor sensor data. IEEE Int. Conf. on Multmedia and Expo (ICME2005), July 2005.

[4] T. Mori, H. Noguchi, A. Takada and T. Sato: Sensing room: Distributed sensor environment for measurement of human daily behavior. First International Workshop on Networked Sensing Systems(INSS2004), (7):40–43, June 2004.

[5] C. Stauffer and K. Tieu: Automated multi-camera planar tracking correspondence modeling. the IEEE Computer Vision and Pattern Recognition, pp.259–266, July 2003.

[6] K. Tieu, G. Dalley, and W. Eric. L. Grimson: Inference of non-overlapping camera network topology by measuring statistical dependence. International Conference on Computer Vision 2005, October 2005.

[7] H. Nait-Charif and S. J. McKenna: Activity summarisation and fall detection in a supportive home environment. ICPR2004, 4:138–141, August 2004.

[8] G. L. Foresti, C. Micheloni and L. Snidaro: Event classification for automatic visual-based surveillance of parking lots. ICPR2004, 3:314–317, August 2004.

[9] M. Higuchi, S. Aoki, A. Kojima, and K. Fukunaga: Scene recognition based on relationship between human actions and objects. ICPR2004, 3:73–78, August 2004.

[10] A. H. Kam., K. A. Toh, H. L. Eng, W. Y. Yau and J. Wang: Automated recognition of highly complex human behavior. ICPR2004, 4:327–330, August 2004.

[11] M. Leo, T. D'Orazio, I. Gnoni, P. Spagnolo, A. Distante: Complex human activity recognition for monitoring wide outdoor environments. ICPR2004, 4:913–916, August 2004.

[12] T. Nishizaki, Y. Kameda, Y. Ohta: Visual surveillance using less rois of multiple non-calibrated cameras. Asian Conference on Computer Vision 2006 (ACCV 2006), 3851:317–327, January 2006.