Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2010, Article ID 525026, 19 pages
doi:10.1155/2010/525026

Research Article

A Macro-Observation Scheme for Abnormal Event Detection in Daily-Life Video Sequences

Wei-Yao Chiu and Du-Ming Tsai

Department of Industrial Engineering and Management, Yuan-Ze University, 135 Yuan-Tung Road, Nei-Li, Tao-Yuan 32026, Taiwan

Correspondence should be addressed to Du-Ming Tsai, iedmtsai@saturn.yzu.edu.tw

Received 19 October 2009; Revised 4 March 2010; Accepted 8 April 2010

Academic Editor: Robert W. Ives

We propose a macro-observation scheme for abnormal event detection in daily life. The proposed macro-observation representation records the time-space energy of motions of all moving objects in a scene without segmenting individual object parts. The energy history of each pixel in the scene is instantly updated with exponential weights without explicitly specifying the duration of each activity. Since possible activities in daily life are numerous and distinct from each other and not all abnormal events can be foreseen, images from a video sequence that spans sufficient repetition of normal day-to-day activities are first randomly sampled. A constrained clustering model is proposed to partition the sampled images into groups. The new observed event that has distinct distance from any of the cluster centroids is then classified as an anomaly. The proposed method has been evaluated in daily work of a laboratory and BEHAVE benchmark dataset. The experimental results reveal that it can well detect abnormal events such as burglary and fighting as long as they last for a sufficient duration of time. The proposed method can be used as a support system for the scene that requires full time monitoring personnel.

1. Introduction

Activity recognition has played an important role in video surveillance for security, traffic-monitoring, homecare, and healthcare applications. An activity recognition system generally involves the following four steps: low-level detection of moving objects from the background with a still camera, spatiotemporal representation of motions in an image sequence, extraction of motion features from the representation, and high-level classification.

There are two major approaches for activity recognition in video sequences: micro-observation and macro-observation. The micro-observation approach analyzes the motions based on the local detailed parts of individual moving objects. In human motion analysis, this means the body parts such as head, torso, and limbs must be identified first, followed by poses assignment based on the extracted body parts. The poses then construct a specific action, and finally a sequence of actions gives a meaningful behavior. This approach requires a bottom-up process to construct a representation from the low-level primitives of foreground objects. The macro-observation approach does not describe the motion of an object by the local details. Instead, it describes the motion from a global aspect using an abstract representation of time-space changes in a video sequence. Human beings have the remarkable ability to recognize the behavior of a single isolated person, or the interaction between multiple people from a far distance without knowing the detailed motions of individual persons.

In this paper, we propose a fast macro-observation surveillance scheme that can detect abnormality in our daily life that involves distinct activities of a single person or a group of people. A video surveillance system that can monitor abnormal events in daily life is very complicated to construct due to unanticipated or indefinable activities.

1.1. Micro-Observation Approach. The micro-observation approach for activity recognition can well describe the details of a motion and provides a good discrimination between individual activities with subtle changes. However, it generally requires an accurate segmentation of foreground

objects from the background and precise identification of the individual body parts. An inaccurate extraction and description of details in a lower level causes the failure in a higher level process.

Appearance-based methods [1–6] used appearance models that combine shape, color, and texture to analyze the moving objects. Model-based methods constructed a human body as articulated/kinematic or skeleton models [7–13]. The poses identified from the object models were considered as individual states in space, and then hidden Markov models (HMMs) [14–18] were generally used to describe the state changes over time. Bayesian networks and neural networks [19–21] were also commonly used for high-level activity recognition. W^4 [5] is a well-known system using such an approach to recognize events between people and objects. This approach is also well applied to gesture recognition [22, 23] and gait recognition [24, 25]. Shah et al. [26] presented a surveillance system, called KNIGHT, that used rule-based algorithms to detect single object activities and multiobject interactions. Speed, direction, and orientation of object silhouettes and their interobject distances were used as features to detect activities such as falling, running, and meeting.

1.2. Macro-Observation Approach. Spatiotemporal representation of an image sequence is critical for recognizing different activities using the macro-observation approach. Optical flow [27–30] that describes each pixel in two consecutive frames by a velocity vector has been popularly used as a motion representation. Efros et al. [31] recognized human actions of individuals in a low resolution video sequence. Their algorithm started by tracking individual human figures and forming a figure-centric sequence. Then the optical flow vector field was calculated from the figure-centric sequence, and a set of motion descriptors were derived from 4 channels of the optical flow. The K-nearest neighbor classifier was finally used to recognize various human actions in sport videos.

Trajectory [32–36] is a commonly-used representation for describing moving objects from a far distance. In order to construct the trajectory of moving objects in an image sequence, object tracking is generally applied first, and then the centroid of a tracked object is marked as a point on the trajectory. The position, speed, direction, and curve/shape of the motion trajectory are used to analyze the intended behaviors of moving objects in the scene. The trajectory representation has been mostly applied to traffic monitoring. Stauffer and Grimson [37] used vector quantization to cluster trajectories for parking lot monitoring. The clusters were identified by a hierarchical analysis of the vector cooccurrences in the trajectories. The trajectory is good for the representation of a widely open scene, but may fail to describe the people interaction in a room-sized scene.

Eigenspace [38–41] derived from principal component analysis is also used for motion representation in video sequences. Rahman and Ishikawa [42] recognized human motion using an eigenspace. A 2D spatial image was first arranged as a column vector. Then, a series of a fixed number of consecutive images for every possible motion to be recognized were organized as a matrix. The eigenvectors with dominant eigenvalues of the covariance matrix formed the eigenspace. A human posture in a frame was then represented by a point in the eigenspace, and a motion was described by a set of successive points in the eigenspace. Distance measures were finally used to match the lines of the observed motions and those of the reference motions. The eigenspace approach is computationally expensive and can only describe specific activities.

Motion Energy Image (MEI) and Motion History Image (MHI), first proposed by Davis and Bobick [43], are a global spatiotemporal representation of motion. They are treated as temporal templates for the match of human movement [44]. MEI is defined as the sum of object silhouettes in every image frame over a fixed duration. The result of MEI is a binary image of motion shape. While MEI is used to record the "shape" of a motion, the intensity of MHI is a function of recency of motion. The effectiveness of the MEI and MHI representations is critically determined by the fixed duration value. Bradski and Davis [45] extended the MHI for motion segmentation and pose recognition by extracting additional pose and directional motion information in MHI. The gradient orientation at each pixel is derived from the spatial derivatives along the y- and x-axis of MHI. Wong and Cipolla [46] also used the gradient directions in MHI for gesture recognition. Davis and Bobick [47] used MHI for recognizing aerobic movements. The temporal templates of MEI and MHI were also used for hand gesture recognition [48]. The temporal template has shown to be a good global representation of motions. However, it is currently only verified for simple activities such as hand gestures and aerobic exercises that have a fairly steady motion duration and is only tested for single isolated object in a simple background.

1.3. Unusual Event Detection. There were a few methods proposed to tackle abnormal/rare event detection in specific domains. Vaswani et al. [49] presented a system that learned the pattern of normal activities and detected abnormal events from a very low-resolution video where the moving objects were small enough to be modeled as point objects. The activity of moving objects was modeled by a polygonal "shape" of the configuration of the tracked points using Kendall's statistical shape theory. The expected log likelihood of the represented Kendall's shape for an observed sequence of fixed length was then used as the change detection statistic. The system was applied to monitor passengers getting out a plane and moving towards the terminal from a very far observation distance. It is basically a trajectory-based method and is only applicable to the monitoring of a widely open scene. Piciarelli and Foresti [36] proposed an on-line trajectory clustering for anomalous event detection, and applied it to traffic behavior monitoring on a highway. The trajectory is represented by a series of position coordinates and is matched to the clusters of a training set by a distance measure. Hu et al. [50] proposed a self-organizing method to learn activity patterns for anomaly detection and activity prediction. The activity patterns were represented by trajectories, where object position, velocity and size were

used as the features. A fuzzy self-organizing neural network was then presented to classify the activity patterns. The system was applied in traffic monitoring to detect abnormal driving trajectories.

Fleet et al. [51] and Andrade et al. [52] used optical flow patterns to detect emergency events in crowded scenes. They first computed the optical flow for the whole frame, and retained only the flow information in the foreground region. Principal component analysis was then performed on the optical flow fields for a series of image frames of fixed duration. The dominant eigenvectors of the training data matrix were used to form bases for the projection. The projected optical flow vectors were then used as features. A mixture of Gaussian hidden Markov model was trained with the feature vectors for each video segment in the training set, and the spectral clustering was used to determine the number of HMMs to represent various flow sequences. For day-to-day behavior analysis, an extremely large set of training image sequences may be required. The covariance matrix of such a large training set could be prohibitively large for PCA computation. Adam et al. [53] proposed an optical flow-based method for unusual event detection in cluttered and crowded environments. The abnormality was mainly detected by evaluating the probability distribution of flow magnitude and direction in the optical flow fields. An unusual event without radical motion changes cannot be detected with this method.

Zhong et al. [54] presented a technique for detecting unusual activities in video sequences. Moving objects in each image frame were detected first. The simple spatial histogram of the detected objects was used as image features and, therefore, the observed activities were location-dependent. They divided the video into equal-length segments and classified the extracted features into prototypes. A cooccurrence matrix between the video segments and prototype features was constructed for similarity comparison. The correspondence between prototypes and video segments was then solved as a graph editing problem.

The abnormal event detection methods aforementioned generally consider the motions of objects with a fixed observation duration (i.e., a predetermined number of image frames) in video sequences, and require well-controlled environments or well-defined patterns of activities. Most of the currently available activity recognition methods only deal with very simple activities, and are domain specific such as aerobic exercises [44] and tennis strokes [55]. In this study, we propose a macro-observation approach to detect abnormal events observed, especially, in a room-sized scene from a still camera. The scene of a room may involve complicated day-to-day behaviors such as an older person staying alone at home (for homecare monitoring), multiple people with/without interaction in a nursing home (for healthcare monitoring), and cashier-customer interaction in a shop (for security monitoring). The observed objects in such scenes have moderate sizes in the image. The trajectory representation of an object as a point may lose meaningful interaction between people. The proposed method does not take the micro-observation approach since it requires complicated object tracking, object segmentation and body part extraction, and state-space modeling for all possible events to detect. The abnormal events in a daily life are very difficult to define semantically, and the normal events are too numerous to model individual day-to-day activities.

1.4. Overview of the Proposed Method. With the macro-observation approach, the proposed method first segments moving objects from the background for each input scene image. The foreground objects shifting in spatial images over time are globally represented by an energy map, where the movement strength of each pixel in the current scene image is exponentially increased/decreased based on the state changes of the pixel over time. The length of image frames for different activities does not have to be explicitly specified, and the energy map can be promptly updated for each new scene image. The shape of the energy map and movement strength of every pixel in the map carry meaningful time-space interaction of single person or multiple people with the environment. A set of discriminative features can then be effectively extracted from the energy map of each new scene image.

Abnormal event detection in daily life can be considered as a very special case of one-class classification problem. No all possible abnormal event in daily life can be foreseen. It is also very difficult to collect all possible conditions of a specific abnormal event. Conversely, normal behaviors in daily life can be easily collected for learning. The behaviors repeated daily can be grouped into many clusters, and all clusters belong to the same class, that is, the normality class. Because the types of normal behaviors in a daily life could be numerously large and quite different from each other, the images are randomly sampled from a long image sequence that can sufficiently represent the cyclical day-to-day activities of the observed scene. An unsupervised clustering subject to distance constraints is proposed to group various normal activities into a manageable number of clusters so that the computation in the recognition process can be efficiently carried out and all normal events can have distances from their cluster centroids within very tight control limits (distance thresholds). The video images with distinct feature distances lasting for an extended period of time are then declared as an abnormal event.

The proposed macro-observation method mimics the human observer who can easily recognize abnormal events from a far distance without knowing the detailed movements of individual persons. The global representation of complicated motions in a scene can well detect abnormal events as long as they can last for tens of seconds. Since the detailed motions of individual body parts are not separately extracted, the proposed method cannot be responsive to the events with subtle motion changes and the events spanning only a few seconds. The proposed monitoring system can be used as a supplement for the personnel that requires intensive and constant manual monitoring of scenes for unpredictable events from multiple cameras.

This paper is organized as follows. Section 2 first discusses the foreground segmentation method to extract moving objects in video images. The energy map used to represent the spatiotemporal motion is then described,

followed by the extraction of discriminative features from the energy map. The proposed clustering mechanism is then presented to group similar energy maps sampled from image sequences of normal daily life. Section 3 describes the experimental results of daily activities in a laboratory over a long period of observation and the BEHAVE benchmark dataset [56]. Section 4 concludes the paper and discusses future work.

2. Abnormal Event Detection

This section discusses the abnormal event detection scheme that comprises the processes of moving object detection, exponential energy map for spatiotemporal representation of motions, extraction of motion features, and the constrained clustering model for classification.

2.1. Moving Object Detection. The objective of the paper is to detect abnormal events in daily life in a scene such as an office or a nursing home, where nonstationary background changes such as movements of a chair, placing of cups and newspapers on tables, revolving of ceiling fans, opening/closing of doors or curtains, and switching on/off room lights are not uncommon. Since the proposed method does not rely on the accurate detail parts of moving objects for the detection, any background subtraction techniques such as the ones in [3, 57–59] can be directly applied to foreground segmentation as long as it is computationally fast.

In background updating models, each pixel of the background image over time has been simply modeled with a single Gaussian model [3]. A more robust background modeling technique is to represent each pixel by a mixture of Gaussians [37, 57]. In order to promptly detect moving objects for nonstop monitoring of day-to-day activities, we adopt a single-Gaussian background updating approach, instead of the more complicated mixture Gaussian model, to extract foreground objects with a high processing rate.

In the single Gaussian model for each individual pixel in the image, the parameters are represented by the gray-level mean $\mu_T(x, y)$ and standard deviation $\sigma_T(x, y)$ of the pixel (x, y) over a limited time duration. Different from the Gaussian background updating models [3, 57] that estimate the parameter values by a linear filtering technique, these two statistical values of the single Gaussian model can be easily and precisely calculated by deleting the last image in the series of the historical image frames and adding the current image frame for nonstop monitoring.

Let $\{f_t(x, y),\ t = T, T-1, \ldots, T-N+1\}$ be a series of N consecutive image frames, where T denotes the current time frame. The gray-level mean and variance of the single Gaussian background model for pixel (x, y) at time frame T is given by

$$\begin{aligned}\mu_T(x,y) &= E[f(x,y)] = \frac{1}{N}S_T(x,y),\\ \sigma_T^2(x,y) &= E[f^2(x,y)] - \{E[f(x,y)]\}^2\\ &= \frac{1}{N}\cdot S_T^2(x,y) - \mu_T^2(x,y),\end{aligned} \tag{1}$$

where

$$\begin{aligned}S_T(x,y) &= \sum_{i=0}^{N-1} f_{T-i}(x,y)\\ &= S_{T-1}(x,y) - f_{T-N}(x,y) + f_T(x,y),\\ S_T^2(x,y) &= \sum_{i=0}^{N-1} f_{T-i}^2(x,y)\\ &= S_{T-1}^2(x,y) - f_{T-N}^2(x,y) + f_T^2(x,y).\end{aligned} \tag{2}$$

Note that $S_T(x, y)$ and $S_T^2(x, y)$ can be efficiently updated by dropping the last image frame $f_{T-N}(x, y)$ in the image series and adding the current image frame $f_T(x, y)$ to the image series. Therefore, the updating computation involves only two simple arithmetic operations. A very high process rate of image frames is achieved accordingly. Note also that the mean and variance updating processes in (1) are invariant to the number of image frames N in the series.

In motion detection, the multiple temporal images of the background will present approximately the same gray value with a small variance. The gray value of a foreground pixel will be distinctly different from that of the background. The upper and lower control limits for foreground-pixel detection in the current image frame $f_T(x, y)$ can be given by $\mu_{T-1}(x, y) \pm \kappa \cdot \sigma_{T-1}(x, y)$, where κ is a control constant. If the gray-level of $f_T(x, y)$ is out of the control limits, pixel at (x, y) is then considered as a foreground point. Otherwise, it is classified as a steady background point. The detection result is represented by a binary image $B_T(x, y)$, where

$$B_T(x,y) = \begin{cases} 0\ (\text{background}), \text{ if } |f_T(x,y) - \mu_{T-1}(x,y)| \\ \qquad\qquad \le \kappa\cdot\sigma_{T-1}(x,y), \\ 1\ (\text{foreground}), \text{ otherwise.}\end{cases} \tag{3}$$

Since the gray values between foreground and background points are generally distinctly different, the control constant κ is set at 5 in this study.

2.2. Spatiotemporal Representation. The goal of this subsection is to construct a global representation of motion that can describe the changes in both temporal and spatial dimensions. The existing spatiotemporal representations of motions aforementioned generally describe the temporal context with a fixed duration in a video sequence. The motion representation from a fixed number of image frames may not sufficiently capture the salient and discriminative properties for a large variety of activities encountered in daily life. Short observation duration cannot describe a full cycle of an activity. In contrast, excessively long observation duration may mix two or more different activities or reduce the significance of a unique activity in the spatiotemporal representation.

In order to construct a more responsive spatiotemporal representation for the scene that may involve the motions of a single person or multiple people with varying time spans

of activities, we construct the motion energy map using an exponential time update process, which is defined as

$$E_T(x, y) = M_T + E_{T-1}(x, y) \cdot \gamma, \tag{4}$$

where γ is the energy update rate, $0 < \gamma < 1$, and

$$M_T(x, y) = \begin{cases} T_{\text{energy}}, \text{if } B_T(x, y) \in \text{foreground}, \\ 0, \qquad \text{otherwise.} \end{cases} \tag{5}$$

The initial value of energy is set to zero at time frame 0, that is, $E_0(x, y) = 0$ for all pixels. The energy of a pixel $E_T(x, y)$ will be increased if it remains as a foreground point. It is only decayed when it becomes a background point. In (4) above, T_{energy} is a predetermined energy constant, and is assigned to each foreground pixel. Assume that the current energy value of a pixel (x, y) is E. If pixel (x, y) is a foreground point and lasts for a period of N_f frames, then the energy at (x, y) is increased up to

$$T_{\text{energy}} \cdot \sum_{i=0}^{N_f - 1} \gamma^i + E \cdot \gamma^{N_f}. \tag{6}$$

Conversely, if pixel (x, y) is changed from a foreground point to a background point and lasts for N_b frames, the energy at (x, y) is then exponentially decreased to

$$E \cdot \gamma^{N_b}. \tag{7}$$

The choice of T_{energy} value is not critical at all as long as it is larger than zero for foreground points and equal to zero for background points. The value of T_{energy} affects only the visual representation of the energy map in the image. It does not change the detection results.

The exponential energy updating of foreground pixels assigns larger weights to the most recent image frames. The energy update rate γ gives an exponential decrement of the energy. A large γ value gives a slow decrement of the energy, and the long-term history of the pixel is taken into account for spatiotemporal representation. In contrast, a small γ value results in an accelerated decrement of energy, and only the short-term history of the pixel is used to represent the motion. The exponential decrement of energy allows flexible adjustment of the observed period for the historical status of each pixel. The proposed exponential energy map of motions prevents the explicit choice of a predetermined number of image frames for the construction of spatiotemporal representation. It can be thus effectively used to represent activities that last for various durations. By detecting each individual pixel as a foreground or a background point in the video sequence, the energy of the pixel can be easily updated according to (4) without knowing its associated moving part of an object. If the motion of the pixel continues (i.e., foreground point), the energy of the pixel will be exponentially accumulated. Otherwise, the energy of the pixel (i.e., background point) will be decreased. In the macro-observation approach, two (or multiple) movements within a scene is simply interpreted as an event in the energy map. They do not have to be separated into different moving parts.

Figure 1 displays the motion energy maps of various video sequences of one single person from daily activities in a laboratory, in which the energy constant T_{energy} is set at 10 for visual display, and the update rate γ is given by 0.999 for the normal walking speed of people in the room. The video images were taken at 10 frames per second. Figure 1(a) shows the original video sequence at varying time frames. The scenario in the sequence is that a single person walked towards the door from the lower-left to the upper-right in the scene. The resulting energy map is shown in the bottom row of Figure 1, where the brightness is proportional to the energy value. Figure 1(b) presents another single person walked from the upper-right door to the lower-left corner in the opposite direction. By closely observing the two corresponding energy maps in Figures 1(a) and 1(b), both display similar representations in shape. The energy values in the upper-right are higher than those in the lower-left in the energy map of Figure 1(a), whereas the energy values in Figure 1(b) show the reverse trend. Therefore, the representative shape of the energy map describes various spatiotemporal activities, and the changes of energy values in the map implicitly indicate the moving direction. Figure 1(c) displays a single person working on a computer. The resulting energy map, as seen in the bottom row of Figure 1(c), shows that only the sitting area of the person gives bright energy values. The historical data of the movement from the lower-left to the upper-right corners were responsively decayed to very small energy values.

Figure 2(a) shows a group of people discussing in the middle-right area for a prolonged period of time, and then walked back to their seats. The bottom row in Figure 2(a) gives the corresponding energy map, in which the middle-right area is brighter than the remaining regions in the image and the energy values for pixels in the walking paths are larger than those of the background. Figure 2(b) shows two people chatting around the public desk in the laboratory, and Figure 2(c) displays two people separately working on the computers. The corresponding energy maps are presented in the bottom row of Figures 2(b) and 2(c), which show that different moving frequencies of multiple people generate different energy maps. Based on the representative samples in Figures 1 and 2, the exponential energy maps can represent different day-to-day activities that involve single or multiple people.

The proposed exponential energy map can well represent spatiotemporal activities from a macro-observation view. It requires no complicated segmentation and object recognition techniques to identify the detailed parts of individuals in a group. It can well represent activities that last a sufficient period of time. In order to prevent false alarms, it is suggested in this paper that an activity last for only a few seconds is interpreted as noise. The restriction of the exponential energy map is that it cannot be effectively used to describe activities that involve only subtle movements of the body or last only a very short period of time.

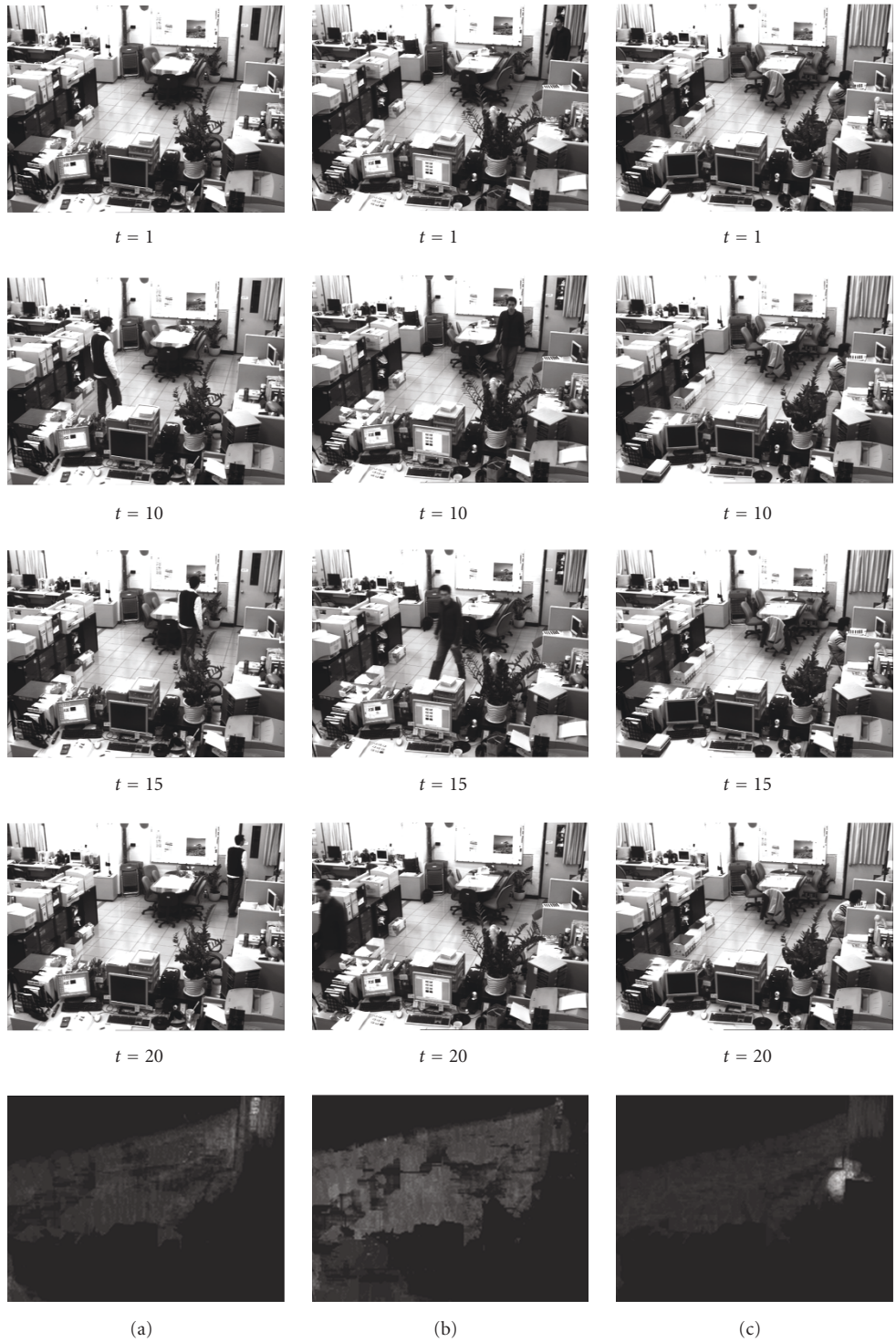

Figure 1: Video sequences involving different activities of a single person and their corresponding energy maps: (a) single person walking from lower-left to upper-right; (b) single person walking from upper-right to lower-left; (c) single person working on a computer. The corresponding energy map of each column sequence is shown in the bottom row.

2.3. Discriminative Features. The proposed exponential energy map gives spatiotemporal representation of an activity. To construct a classification system for identifying abnormal events, we need to design and extract discriminative features from the energy map. The shape and energy statistics of the energy map are used as descriptors. Currently, we use up to 12 discriminative features and they are described in detail as follows.

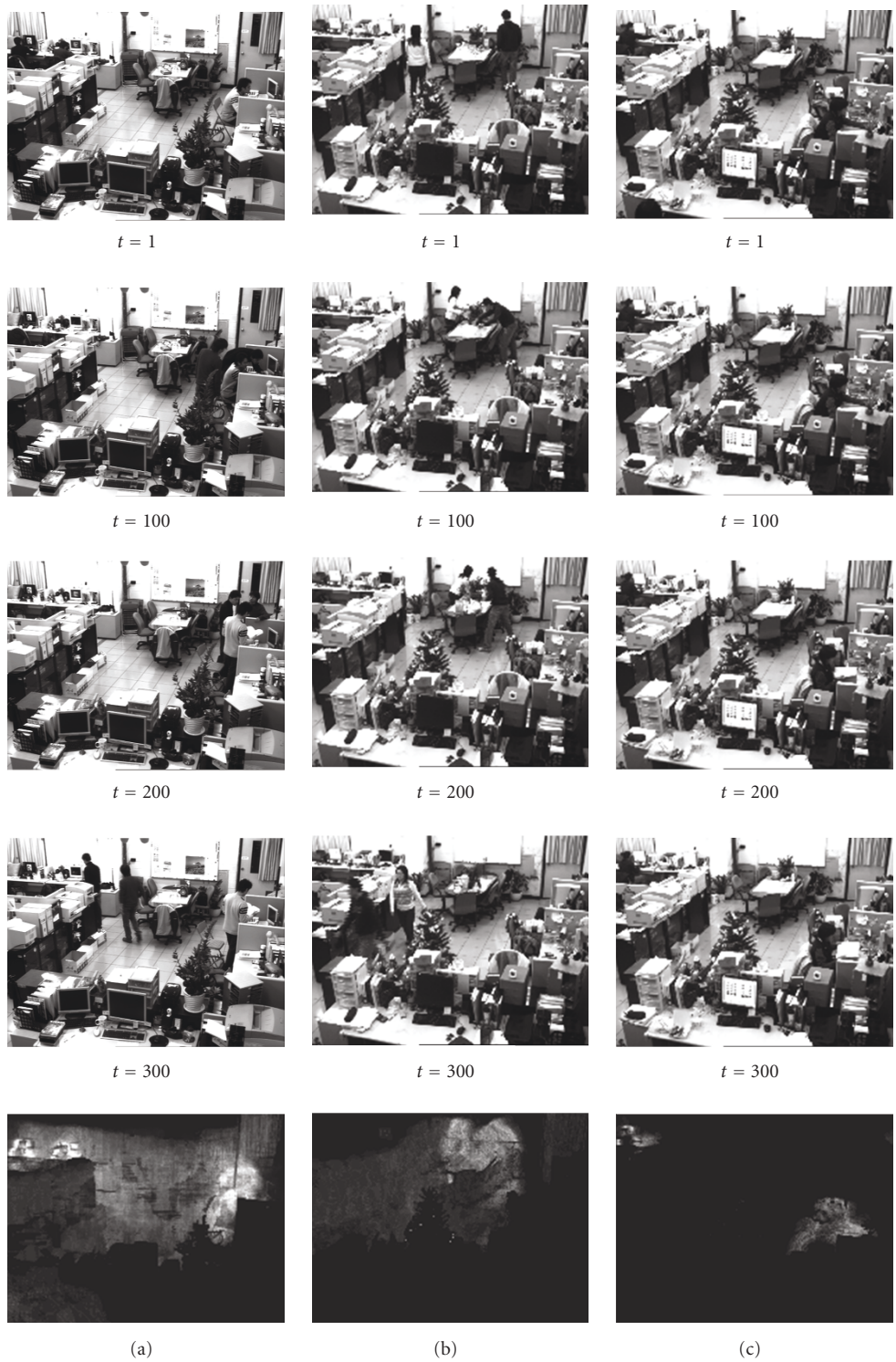

Figure 2: Video sequences involving different activities of multiple people and their corresponding energy maps: (a) a group of people discussing in the middle-right area; (b) multiple people chatting around the public desk; (c) multiple people working on the computers. The corresponding energy map of each column sequence is shown in the bottom row.

Invariant Moments f_1 to f_7. an event in the energy map forms a specific shape with the energy magnitude of each pixel as the weight. The extracted features from the energy map should be independent of location, orientation and size of an activity in the image. The first seven discriminative features are, therefore, based on Hu's invariant moments [60]. Features $f_1 \sim f_7$ are invariant to position, rotation and scale changes. The seven invariant moments used in this

study are not merely computed from the binary shape, but use the energy value $E_T(x, y)$ as the density for each pixel in the energy map.

Entropy f_8. let $E'_T(x, y)$ be the normalized energy value into integer in the range between 0 and 255 (for an 8-bit display). Thus

$$E'_T(x, y) = \left[\frac{E_T(x, y) - \mathrm{Min}_{u,v} E_T(u, v)}{\mathrm{Max}_{u,v} E_T(u, v) - \mathrm{Min}_{u,v} E_T(u, v)} \times 255 \right]. \tag{8}$$

Denote by P_i the probability that $E'_T(x, y) = i, i = 0, 1, 2, \ldots, 255$. The entropy of the energy map is therefore defined as

$$f_8 = -\sum_i P_i \cdot \log P_i. \tag{9}$$

The entropy feature describes the complexity of movements in a scene. A still scene will have an entropy value approximate to zero. A single person sitting in a chair for study will yield a small entropy value, whereas a scene involving interaction and movements of multiple people will result in a large entropy value.

Maximum Energy f_9. the maximum energy is defined as

$$f_9 = \mathop{\mathrm{Max}}_{x,y} E_T(x, y). \tag{10}$$

This feature gives the maximum energy value in the energy map. A foreground object that keeps moving for a prolonged period of time will have a larger feature value of f_9, compared to that for a short period of time.

Total Energy f_{10}. the total energy is defined as

$$f_{10} = \sum_x \sum_y E_T(x, y). \tag{11}$$

A scene with a group of people generally yields a larger total energy value, compared to the scene with a single person. A person who keeps moving around in the scene will generate a larger total energy value, whereas a person sitting for study will result in a smaller total energy value.

Area of Nonzero Energy f_{11}. this feature is defined as

$$f_{11} = \sum_x \sum_y b(x, y), \tag{12}$$

where

$$b(x, y) = \begin{cases} 1, \text{if } E_T(x, y) > 0 \\ 0, \text{otherwise.} \end{cases} \tag{13}$$

A wide moving area will result in a larger feature value of f_{11}, even if the movement lasts for only a very short period of time, whereas a limited moving area will have a smaller feature value even if the movement lasts for a prolonged period of time.

Mean Energy f_{12}. the mean energy is defined as

$$f_{12} = \frac{\sum_x \sum_y E_T(x, y)}{\sum_x \sum_y b(x, y)} = \frac{f_{10}}{f_{11}}. \tag{14}$$

This feature gives the mean energy value in the region of nonzero energy. The total energy f_{10} for highly repetitive activities in a small limited area may be similar to that for nonrepetitive activities in a wide area. The mean energy can be used to describe the relationship between the repetitive motions and the moving area.

As demonstration examples, Figures 3(a1)–3(a3) present the energy maps of people sitting in chairs for study, Figures 3(b1)–3(b3) display the energy maps of a single person walking in different directions, and Figures 3(c1)–3(c3) are the energy maps of the interaction between multiple people. The corresponding features values of $f_1 \sim f_{12}$ for the individual energy maps are summarized in Table 1. It shows that similar activities yield similar feature values, and different activities result in distinct feature values.

2.4. Classification. The discriminative features extracted from the motion energy maps can now be used to identify abnormal events from the normal activities in daily life. Monitoring of abnormality in daily life is not possible to be restricted only to the recognition of prestudied and premodeled events. As aforementioned, there could have numerous distinct daily-life activities in an observed scene. It is extremely difficult to apply a supervised classification system, where each input sample must be manually assigned a class index. The selected classification system should be computationally efficient in the detection stage so that it can be easily implemented for on-line, real-time monitoring. The fuzzy C-means (FCM) clustering [61] has been a widely used technique for unsupervised classification. However, the conventional clustering technique only partitions samples into clusters such that the weighted mean distance of each sample to its centroid is minimized. There is no control of the distance variance in each cluster. It cannot handle clusters of different sizes and densities. It is extremely difficult to find a fixed global distance threshold for each cluster to separate normal and abnormal events in video images.

In this paper, a constrained clustering method is applied for training with the objective that the distance of every cluster member to its own cluster center meets adaptively a distance constraint. In order to collect sufficient representative samples of daily life under observation, the training energy maps and, thus, their corresponding discriminative features are randomly sampled from a video image sequence that spans the sufficient period for all possible day-to-day activities. Note that each single input scene image $f_T(x, y)$ has its own corresponding energy map $E_T(x, y)$. Each training sample in this study means the feature vector $(f_1, f_2, \ldots, f_{12})$ of an energy map. Based on our experiments, a range between 15% and 20% of the total image frames in the video image sequence is sufficient to train the classifier. The classification system involves two processes, learning process and detection process, which are individually described in the following two subsections.

Table 1: Feature values for the demonstrative energy maps in Figure 3.

Features	Energy maps in Figure 3								
	3(a1)	3(a2)	3(a3)	3(b1)	3(b2)	3(b3)	3(c1)	3(c2)	3(c3)
f_1	1.381	1.636	1.466	2.111	2.108	2.136	2.354	2.496	2.476
f_2	2.951	3.627	3.260	4.401	4.577	4.553	5.309	5.507	5.588
f_3	4.678	4.630	4.798	7.227	8.090	8.711	7.555	8.126	8.534
f_4	4.922	4.647	4.840	7.081	7.111	8.315	8.738	9.571	9.499
f_5	9.752	9.286	9.700	14.238	14.871	16.981	16.993	20.433	20.102
f_6	6.402	6.461	6.471	9.285	9.412	10.595	11.454	13.215	12.294
f_7	10.175	10.829	10.041	15.189	14.854	16.978	17.089	18.421	18.517
f_8	0.717	0.708	0.740	1.085	1.309	1.421	1.603	1.747	1.755
f_9	186	256	340	253	249	313	361	325	640
f_{10}	103727	106136	161232	381675	435401	513579	860751	1335763	1281152
f_{11}	13084	13248	14248	13231	17102	20284	20063	20177	21571
f_{12}	7	8	11	28	25	25	42	66	59

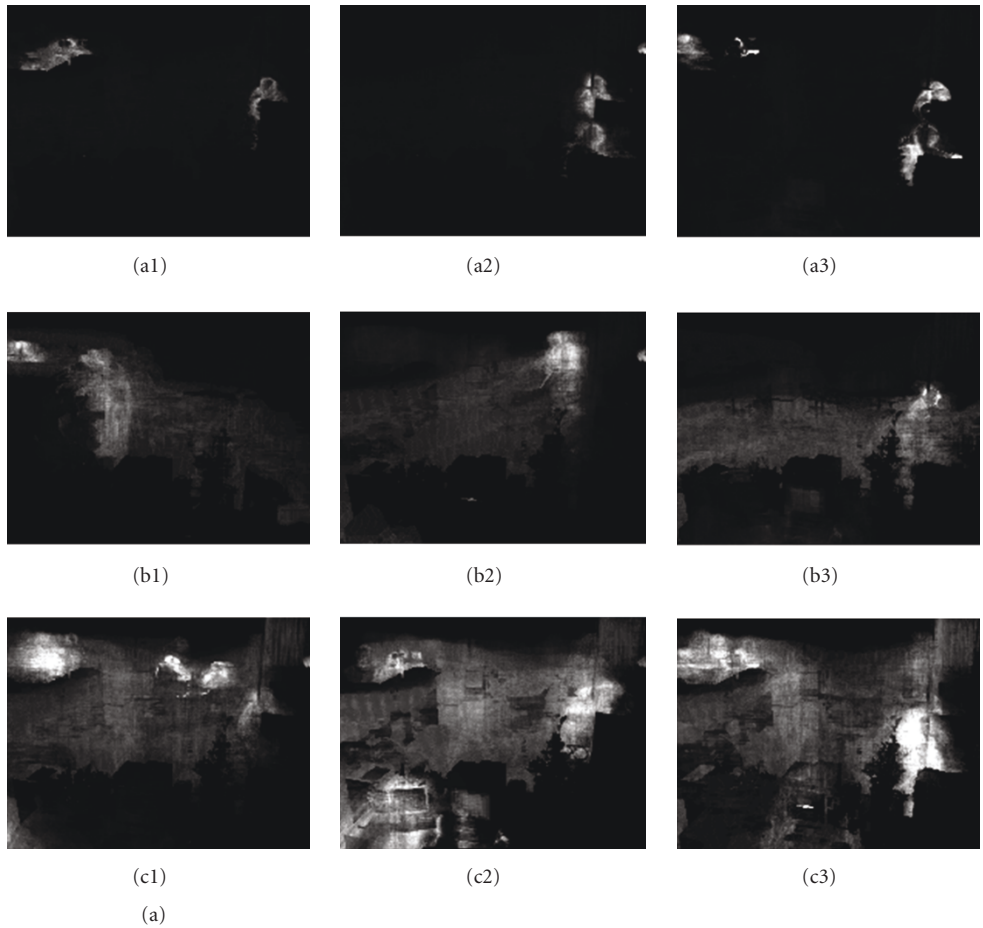

Figure 3: Demonstration examples of energy maps: (a1)–(a3) people sitting in chairs; (b1)–(b3) single person walking in different directions; (c1)–(c3) interaction between multiple people.

2.4.1. Learning Process. In this paper, we are only interested in the classes of normal and abnormal events. Since the abnormal events are unpredictable beforehand, all training samples are normal activities collected from a video sequence of daily life. They all belong to the same class, that is, there is only one class to identify. However, different normal activities may have very distinct representations of energy maps. We would like to group similar activities that have similar energy maps and, thus, similar feature vectors into the same cluster. The goal of clustering for this one-class

classification with distinct patterns problem is to assign similar training samples to the same cluster so that the distance of every member in the cluster to the cluster center meets a minimum distance threshold.

Let $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_K\}$ be a set of K training samples, and $\mathbf{v}_i$ the centroid of cluster i. The distance between sample $\mathbf{x}_k$ and the centroid $\mathbf{v}_i$ is denoted by $d(\mathbf{x}_k, \mathbf{v}_i) = \|\mathbf{x}_k - \mathbf{v}_i\|^2$. The objective of the proposed clustering is given by

$$\begin{aligned} &\text{Min } \beta \\ &\text{s.t.} \quad d(\mathbf{x}_k, \mathbf{v}_i) \le \mu_{d_i} + \beta \cdot \sigma_{d_i}, \\ &\qquad \text{if } \mathbf{x}_k \in \mathbf{v}_i, \quad \forall k = 1, 2, \ldots, K, \end{aligned} \tag{15}$$

where μ_{d_i} and σ_{d_i} are the mean and standard deviation of the distances $d(\mathbf{x}_k, \mathbf{v}_i)$ for all members in cluster i, and β is a control constant. The upper control limit $\mu_{d_i} + \beta \cdot \sigma_{d_i}$ is used as an adaptive distance threshold T_{d_i} for each individual cluster i. Each member in its own cluster must meet the distance constraint, and the control limit should be as tight as possible.

In this aper, we use a hierarchical clustering technique to group similar energy maps into clusters. In each hierarchical level of the clustering, a small number of clusters C is given, and then the standard fuzzy C-means clustering process is carried out. In the resulting clusters, the Euclidean distance between each assigned member of the cluster and the cluster centroid is calculated so that the mean μ_{d_i} and standard deviation σ_{d_i} for each cluster i can be determined. If the distance is less than the distance threshold T_{d_i}, the sample member is retained in the cluster. Otherwise, it is removed from the cluster. This procedure is repeated for every cluster. At the end of the process, all removed samples are considered as a new set of training data, and the fuzzy C-means clustering with C as the number of clusters is performed in the next hierarchical level. The clustering process is expanded to the lower hierarchical levels until the distance of every member in individual clusters is less than its distance threshold T_{d_i}, or the maximum total number of clusters $C_{\max}$ is met.

At the end of the hierarchical clustering process, the control constant β will be reduced to tighten the distance thresholds if the distance constraints for all training samples under a given total number of clusters $C_{\max}$ are satisfied. Otherwise, it is increased to loosen the distance thresholds. The hierarchical clustering process is then repeated with the new control constant. The minimum value of the feasible control constant β can be efficiently obtained by a binary search. The total number of clusters $C_{\max}$ is predetermined, which is related to the complexity of daily activities in question. Experiments on various scene scenarios have shown that the total number of clusters around 50 and 60 is sufficient to represent different patterns of normal activities in daily life. The number of clusters C in each hierarchical clustering level is given by 10 in this study. The detailed algorithm of the constrained clustering model with minimum adaptive distance thresholds is presented as follows.

Input. The number of clusters C in each level, maximum number of clusters $C_{\max}$, and training data set $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_K\}$.

Step 1. Normalize the feature values.

Let $\mathbf{x}_k = (f_{k,1}, f_{k,2}, \ldots, f_{k,12})$ be the feature vector, and $f_{k,j}$ be the jth feature of sample k, for $k = 1, 2, \ldots, K$

$$f'_{k,j} = \frac{f_{k,j} - \mu_j}{\sigma_j}, \quad j = 1, 2, \ldots, 12, \tag{16}$$

where μ_j and σ_j are the mean and standard deviation of feature j for all training samples.

Let $\mathbf{x}'_k = (f'_{k,1}, f'_{k,2}, \ldots, f'_{k,12})$.

Step 2. Perform the standard fuzzy C-means clustering.

Let $\mathbf{v}_i$ be the centroid of cluster i, $i = 1, 2, \ldots, C$, and

$$\mathbf{v}_i = \frac{\sum_{k=1}^{K} w_{ik}^p \cdot \mathbf{x}'_k}{\sum_{k=1}^{K} w_{ik}^p}, \tag{17}$$

where w_{ik} is the weight for training sample k in cluster i, p is weighting exponent ($p = 2$ in this study).

In each iteration, w_{ik} is updated by

$$w_{ik} = \frac{1}{\sum_{j=1}^{C} \left(d\left(\mathbf{x}'_k, \mathbf{v}_i\right)/d\left(\mathbf{x}'_k, \mathbf{v}_j\right)\right)^{1/p-1}}, \tag{18}$$

where $d(\mathbf{x}'_k, \mathbf{v}_i) = \|\mathbf{x}'_k - \mathbf{v}_i\|^2$. Then the centroid $\mathbf{v}_i$ is updated using the new assigned weight w_{ik}. The updating procedure is repeated until convergence.

Let $\mathbf{V}^r = \{\mathbf{v}_i^r\}_{i=1}^{C}$ be the resulting set of cluster centroids at hierarchical level r (Initially, set $r = 1$).

Step 3. (a) Assign sample $\mathbf{x}'_k$ to cluster $\mathbf{v}_i^r$, for $k = 1, 2, \ldots, K$, where $i = \arg\min_c d(\mathbf{x}'_k, \mathbf{v}_c^r)$.

(b) Set the distance threshold of each cluster $\mathbf{v}_i^r$, $i = 1, 2, \ldots, C$, to

$$T_{d_i} = \mu_{d_i} + \beta \cdot \sigma_{d_i}, \tag{19}$$

where μ_{d_i} and σ_{d_i} are the distance mean and standard deviation of cluster i.

(c) Let $\mathbf{X}_i^r = \phi$ and $\widetilde{\mathbf{X}}_i^r = \phi$, for $i = 1, 2, \ldots, C$. Given that $\mathbf{x}'_k \in \mathbf{v}_i^r$, $k = 1, 2, \ldots, K$, if $d(\mathbf{x}'_k, \mathbf{v}_i^r) < T_{d_i}$, then assign

$$\mathbf{X}_i^r \longleftarrow \mathbf{X}_i^r \bigcup \left\{\mathbf{x}'_k\right\}, \tag{20}$$

otherwise,

$$\widetilde{\mathbf{X}}_i^r \longleftarrow \widetilde{\mathbf{X}}_i^r \bigcup \left\{\mathbf{x}'_k\right\}. \tag{21}$$

At the end of the assignment, $\mathbf{X}_i^r$ contains all the members that meet the distance constraints, that is, $d(\mathbf{x}'_k, \mathbf{v}_i^r) < T_{d_i}$, in cluster i. The centroid of cluster $\mathbf{v}_i^r$ is updated by

$$\mathbf{v}_i^r = \frac{1}{|\mathbf{X}_i^r|} \sum_{\mathbf{x}'_k \in \mathbf{X}_i^r} \mathbf{x}'_k, \tag{22}$$

where $|\mathbf{X}_i^r|$ is the cardinality of cluster $\mathbf{X}_i^r$.

$\tilde{\mathbf{X}}_i^r$ records the samples with $d(\mathbf{x}'_k, \mathbf{v}_i^r) > T_{d_i}$. Let $\tilde{\mathbf{X}}^r = \bigcup_{i=1}^{C} \tilde{\mathbf{X}}_i^r$, which is the set that contains all the samples that violate the distance constraints at iteration r. It is passed along to the next hierarchical level $r + 1$ as a new training set.

Step 4. Cluster in the lower hierarchical level.

Take $\tilde{\mathbf{X}}^r$ as the set of new training data. Let $r \leftarrow r + 1$.

Repeat Steps 2 and 3 until $r \cdot C > C_{\max}$ (max. number of clusters is violated), or $\tilde{\mathbf{X}}^r = \phi$ (all samples meet the distance constraints).

Step 5. Find the minimum control constant β.

If $r \cdot C > C_{\max}$ and $\tilde{\mathbf{X}}^r \neq \phi$, the current control constant β is too tight, and must be increased by setting

$$\beta \longleftarrow \frac{1}{2}\left(\beta + \beta_{\text{upper}}\right), \tag{23}$$

else setting

$$\beta \longleftarrow \frac{1}{2}\left(\beta + \beta_{\text{lower}}\right). \tag{24}$$

Repeat Steps 2 to 5 unit $\Delta\beta < 0.1$, where $\Delta\beta$ is the difference between the old and the new β values. Currently, the lower bound and upper bound of β are set at $\beta_{\text{lower}} = 0.0$ and $\beta_{\text{upper}} = 2.0$.

The proposed clustering model can effectively assign similar activities into the same cluster that all adaptively meet a tight distance threshold. It is expected that normal activities similar to the sampling ones in the training set will also have a corresponding cluster that meets the distance threshold, whereas an abnormal event (the one not observed in the training set) will not find any cluster that yields a distance less than the threshold.

2.4.2. Detection Process. Let $\Omega = \bigcup_r \mathbf{V}^r$ be the set of the final cluster centroids obtained from the training process. For a new scene image at current time frame T with the feature vector $\mathbf{x}_T = (f_{\mathbf{x},1}, f_{\mathbf{x},2}, \ldots, f_{\mathbf{x},12})$, the feature value is first normalized with respect to the mean μ_j and standard deviation σ_j for each feature j of the training samples, that is,

$$f'_{\mathbf{x},j} = \frac{f_{\mathbf{x},j} - \mu_j}{\sigma_j}, \quad j = 1, 2, \ldots, 12, \tag{25}$$

and let $\mathbf{x}'_T = (f'_{\mathbf{x},1}, f'_{\mathbf{x},2}, \ldots, f'_{\mathbf{x},12})$. The minimum distance of $\mathbf{x}'_T$ to the cluster centroids in Ω is given by

$$d(\mathbf{x}'_T, \mathbf{v}_{i^*}) = \min_{\mathbf{v}_i \in \Omega} \left\{ \left|\left| \mathbf{x}'_T - \mathbf{v}_i \right|\right|^2 \right\}, \tag{26}$$

where $\mathbf{v}_{i^*} = \arg\min_{v_i \in \Omega} d(x'_T, v_i)$.

In the training process, the distance threshold T_{d_i} of each cluster i is adaptively given by $\mu_{d_i} + \beta \cdot \sigma_{d_i}$. In the detection process, the same distance threshold of each cluster is also applied to detect abnormal events. If $d(\mathbf{x}'_T, \mathbf{v}_{i^*}) > T_{d_{i^*}}$, a suspected abnormal event is declared. Otherwise, it is classified as a normal activity in daily life. Since an abnormal event will generally last for an extended period of time, a single alarm of $\mathbf{x}'_T$ is treated as noise. When the motion energy maps have $d(\mathbf{x}'_T, \mathbf{v}_{i^*}) > T_{d_{i^*}}$ and prolong for a sufficient duration, an abnormal event is evidently detected. Since the detection process involves only simple Euclidean distance computation from a small set of cluster centroids, it is computationally very fast.

3. Experimental Results

This section evaluates the performance of the proposed abnormality detection scheme from two image sequences, one involving the scene of a laboratory and the other obtained from the BEHAVE benchmark dataset. The proposed algorithms were implemented using the C++ language on a Pentium 4, 3.0 GHz personal computer. The test images in the experiments were 200×150 pixels wide with 8-bit gray levels. The total computation time from foreground segmentation to abnormality detection for an input image is 0.132 seconds, of which the computation of the seven invariant-moments takes 0.121 seconds. It achieves a mean of 7.6 fps for real-time detection of abnormal events.

The first activity monitoring example is the daily work in a laboratory, which involves various activities of a single person and multiple people. Some of the demonstration activities in the laboratory are displayed in Figures 1 and 2. The training image sequences were collected for two days, and there are a total of 100,216 energy maps. 15% of the total energy maps were randomly sampled, which corresponds to 15,030 energy maps used in training. In the experiments, the two parameters used to construct the motion energy maps were set with $T_{\text{energy}} = 10$ and $\gamma = 0.999$ for the relatively slow activities in the laboratory. The total number of clusters $C_{\max}$ is given by 50. The resulting minimum feasible value of the control constant β is 0.1.

In the experiments, we simulated three abnormal events including burglary, fighting and moving furniture out of the room. All these three activities are very difficult to define explicitly and model beforehand. Scenario 1 involves only the actions of a single person. Scenarios 2 and 3 involve interactions between two people. For the burglary scenario, a person was asked to find a wallet hidden in the room as fast as possible. No further instructions on how to find the wallet were given to the pretended burglar. Figure 4(a) displays the original video sequence at varying time frames for the burglary scenario, and Figure 4(b) shows the corresponding energy maps. It can be seen that the energy in the map is weak in the early stage of the burglary activity. The energy is then accumulated and the shape in the map becomes stable after a sufficient period of time.

For the fighting scenario, two people were fighting each other in the room. Figures 5(a) and 5(b) present, respectively, the video sequence and the corresponding energy maps for the fighting scenario. For the moving furniture scenario, two people sequentially moved a chair, a computer monitor and other laboratory objects out of the room. Figures 6(a) and 6(b) show, respectively, the video sequence and the corresponding energy maps for the moving furniture

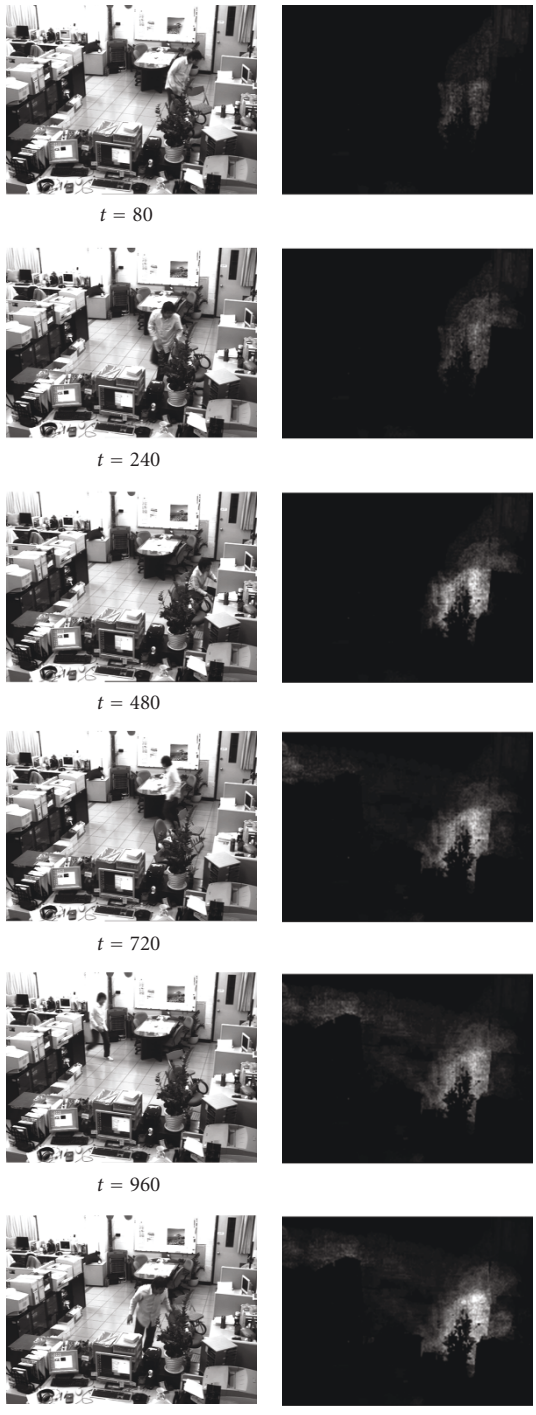

Figure 4: Abnormal event of a burglary scenario: (a) discrete image frames in the sequence; (b) corresponding energy maps. (symbol t represents the frame number in the sequence with fps = 10).

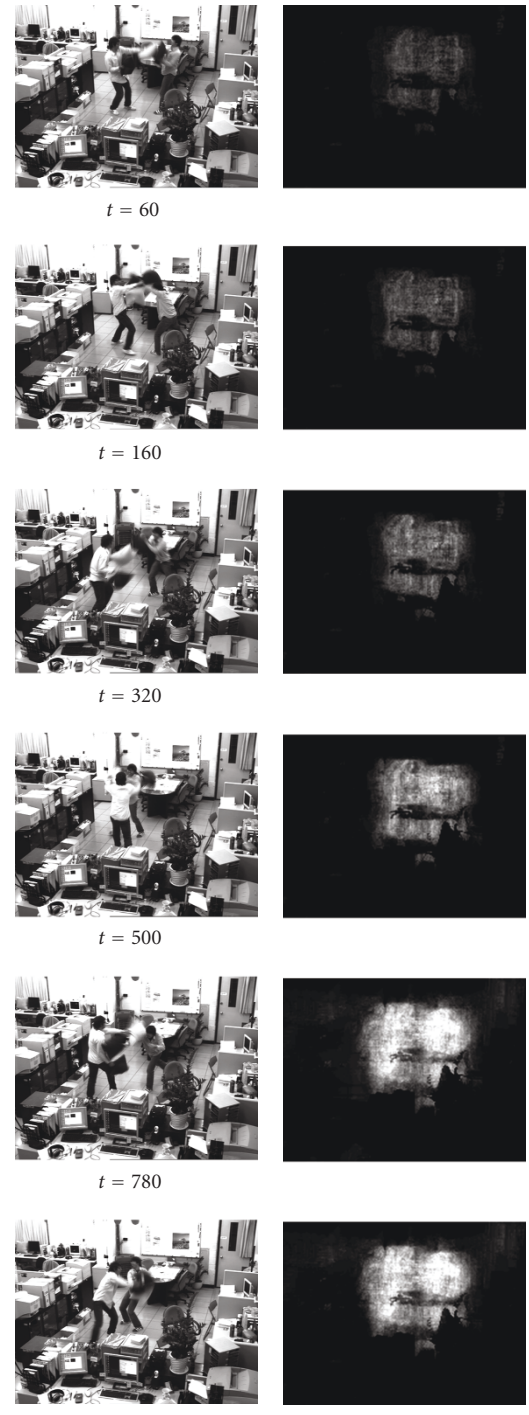

Figure 5: Abnormal event of a fighting scenario: (a) discrete image frames in the sequence; (b) corresponding energy maps.

scenario. The energy is accumulated and the shape becomes clear in the map as the activity proceeds.

When the training is done, 85% of the untrained image frames (a total of 85,186 frames) from the two-day video sequence are used to test for the similarity measurement. In the total of 85,186 frames, only 27 events that have distances $d(\mathbf{x}_T, \mathbf{v}_{i^*})$ larger than the threshold $T_{d_{i^*}}$ are falsely alarmed, and the detection results are displayed in Figure 7. Since each individual input image $\mathbf{x}_T$ has its own corresponding cluster i^* and, thus, different distance threshold

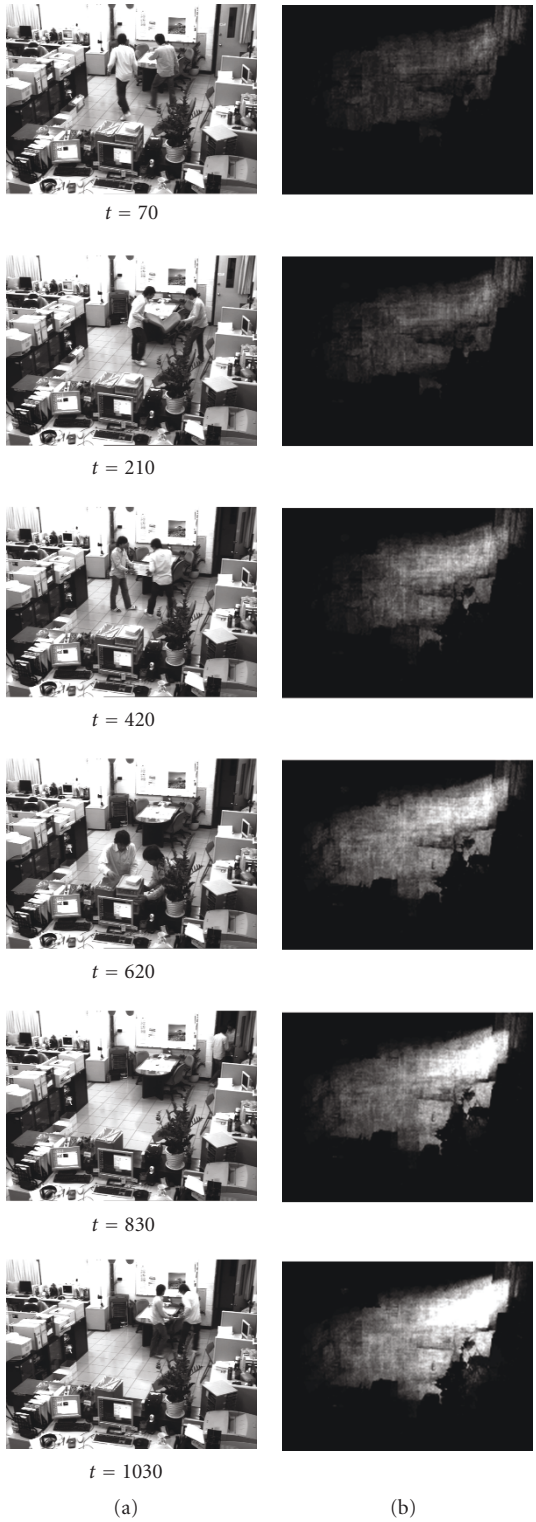

Figure 6: Abnormal event of a moving-furniture scenario: (a) discrete image frames in the sequence; (b) corresponding energy maps.

$T_{d_{i^*}}$, the plot in Figure 7 displays only the difference between the distance $d(\mathbf{x}_T, \mathbf{v}_{i^*})$ and the threshold $T_{d_{i^*}}$, that is, $\Delta d(\mathbf{x}_T, \mathbf{v}_{i^*}) = \max\{d(\mathbf{x}_T, \mathbf{v}_{i^*}) - T_{d_{i^*}}, 0\}$. In the figure, the x-axis presents the event number and the y-axis is the excessive

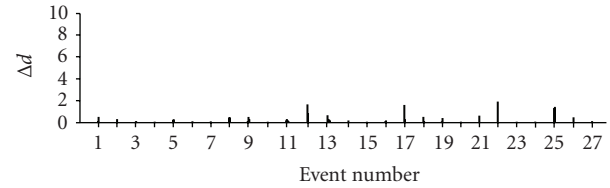

Figure 7: Excessive distances Δd over the threshold $T_{\Delta d}$ for the 27 detected events with distances beyond the control limits in the normal 2-day laboratory video sequence. (The length of each event in the x-axis represents the event duration.)

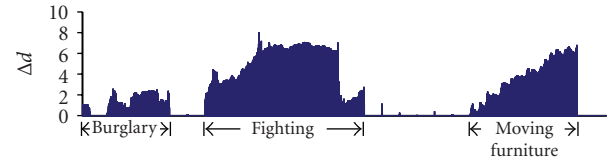

Figure 8: Excessive distances Δd of the three abnormal events in the laboratory: burglary, fighting and moving furniture out of the room. (Note that the duration scales in the x-axis of both Figures 7 and 8 are the same.)

distance Δd. The results show that most of the 85,186 frames have the distances within the control limits. All the falsely detected events last only a very few frames (as seen in the x-axis) and the excessive distance Δd is very small and less than 2 (as seen in the y-axis). The duration for the 27 falsely-detected events is from a minimum of 0.2 seconds (1 frame) to a maximum of 4 seconds (20 frames) with a mean of 0.72 seconds (3.6 frames). Because an activity must last for some duration (i.e., a sufficient number of consecutive image frames), the isolated image frames can be classified as noise.

Figures 8 illustrates the measured distances over time for the three abnormal events of burglary, fighting and moving furniture. The plot only displays the distance differences $\Delta d(\mathbf{x}_T, \mathbf{v}_{i^*})$. The scale on the x-axis in Figure 8 is exactly the same as that in Figure 7, that is, the length of an event in the x-axis of the figure represents also the duration of the activity. The results show that the abnormal activity at the beginning gives small distance values. As the abnormal activity continues, the resulting distances become distinctly large and prolong for a long duration, as seen in the x-axis and the y-axis in Figure 8. Table 2 summarizes the resulting statistics of duration and excessive distance Δd for the 2-day normal image sequence and the three abnormal events. It again reveals that the proposed detection scheme can well identify the prolonged abnormal activities with distinctly large distances Δd. The falsely-alarmed events give only a very short duration with very small excessive distances and therefore, can be effectively eliminated by introducing additional decision rules based on the event duration.

In order to further test the robustness of the proposed method for abnormal event detection in daily life, the same laboratory scene was continuously monitored for 31 days. There are a total of 26,784,000 images frames observed. The trained cluster centroids based on the two-day sampled images, as described previously, are also used for

Table 2: Statistical analysis for the two-day normal video sequence of the laboratory scene and the three abnormal events.

Image sequence	Total events detected	Excessive distance Δd		Alarm duration (sec.)		
		Mean	Std.	Min.	Max.	Aveage
Normal image sequences for 2 days (85,186 frames)	27	0.33	0.39	0.2	4	0.72
Abnormal image sequence	3	3.75	2.18	71.4	182	96

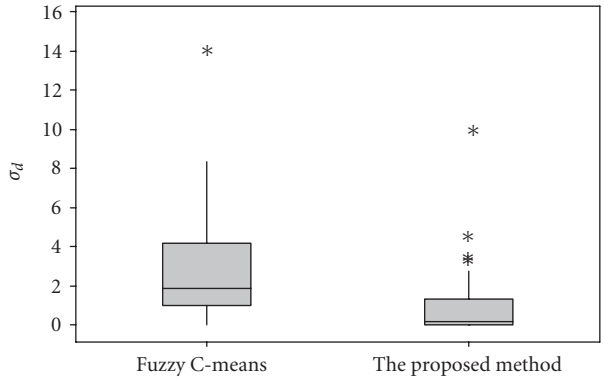

Figure 9: Box-plots of σ_d for the 50 clusters from the constrained clustering model and the conventional FCM.

Table 3: False positive measures under varying event durations for the 31-day laboratory video sequence.

Category of eventduration (seconds)	NFAPA	MTBFA
	Number of false alarms (per day)	Mean time between false alarms (hours)
Event duration > 1 s	3.6	6.6
Event duration > 5 s	2.3	10.4
Event duration > 30 s	1.4	17.1
Event duration > 60 s	1.0	24.0
Event duration > 90 s	0.7	34.2
Event duration > 120 s	0.6	40.0

Table 4: The definition of nine activities in the BEHAVE dataset.

Activity	Definition
In-Group	The people are in a group and not moving very much
Approach	Two people or groups with one (or both) approaching the other
Walk Together	People walking together
Ignore	Ignoring of one another
Split	Two or more people splitting from one another
Following	Being followed
Meet	Two or more people meeting one another
Fight	Two or more groups fighting
Run together	The group is running together

abnormal event detection in this long-observation sequence. The performance of the proposed method on the 31-day image sequence is measured by the false positive rate (false alarms of normal events) given that all the three abnormal events (burglary, fighting and moving furniture) are correctly identified. Because the distances are calculated for individual image frames and an event lasts a number of consecutive image frames, the false positive rate is therefore measured by the mean number of false alarms per day (NFAPD) and its corresponding "mean-time-between-false-alarms (MTBFA)". MTBFA is the average time between two consecutive events alarmed by the monitoring system. The higher the MTBFA is, the higher the reliability of the monitoring system. Table 3 summarizes NFAPD and MTBFA measures for the 31-day image sequence. The detected events are grouped into 6 categories according to their time durations. For the detected events lasting longer than 5 seconds, the mean number of false alarms is only 2.3 events per day. It indicates the mean time between false alarms is 10 hours, and is quite tolerable for a monitoring support system. By analyzing the falsely detected events in detail according to their durations in seconds, we found that the falsely detected events with prolonged durations are generally traceable, that is, there are assignable causes to those events, such as installing a new air conditioner in the laboratory, assembling new computer equipment by a vendor, and tour visit to the laboratory. None of them were observed in the two-day video sequence used in training.

In order to show the effectiveness of the constrained clustering model with respect to the standard fuzzy C-means (FCM) method, the distances $d(\mathbf{x}_T, \mathbf{v}_{i^*})$'s of the two methods for the 15,030 training image frames described previously are evaluated. The total number of clusters was 50 for both methods. Let σ_d be the standard deviation of $d(\mathbf{x}_T, \mathbf{v}_{i^*})$ for all members in a cluster. The value of σ_d should be as small as possible for a more reliable monitoring. Figure 9 presents the box-plot that shows the maximum, minimum, median, and the lower and upper quartiles of σ_d values for the resulting 50 clusters of individual methods. It indicates that the standard FCM method generates a high variation of distances (with a mean σ_d of 2.93) and the proposed clustering model results in a smaller and more stable variation (with a mean σ_d of 0.93).

Surprisingly, the laboratory work can be trained in two days with a very limited number of sampled images, and the trained cluster centroids can be used to describe most of the daily work in the laboratory for over a month. It is believed the false positive rate can be further improved by including more sampling images from sufficient observation days in the learning process.

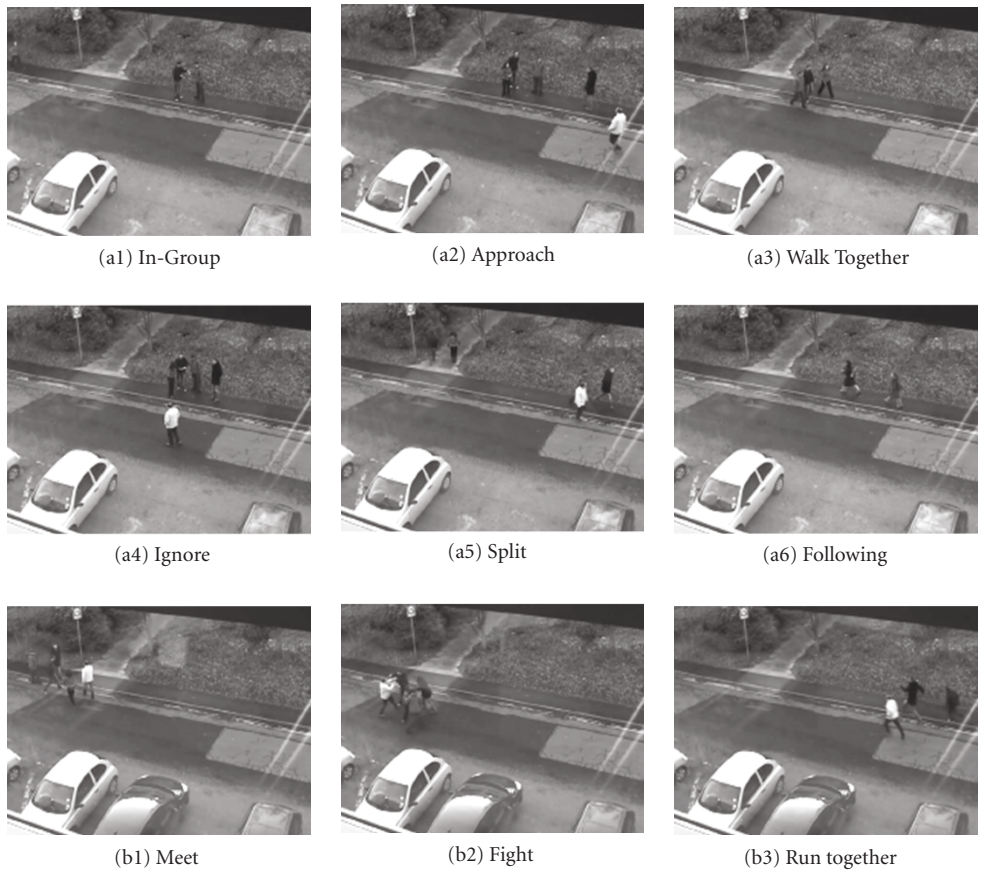

Figure 10: Activity examples in the BEHAVE dataset: (a1)–(a6) scenarios in the Sequence 0; (b1)–(b3) three activities in Sequence 5, which are abnormal with respect to Sequence.

Table 5: The scenarios of the learning and testing sequences for the BEHAVE dataset.

Video clips	Scenarios	Frame number	Video length
Sequence 0	In-Group, Approach, Walk Together, Ignore, Split, Following	1–11200	7 min. 27 sec.
Sequence 5	In Group, Approach, Walk Together, Split, Meet, Run Together, Fight	47300–58400	7 min. 24 sec.

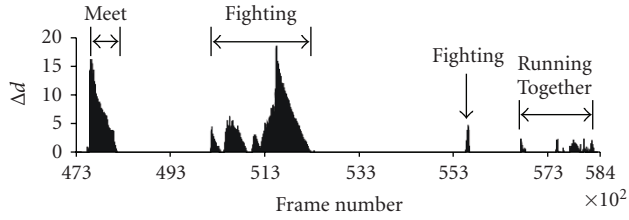

Figure 11: Excessive distances Δd of the four abnormal events in Sequence 5 of the BEHAVE dataset.

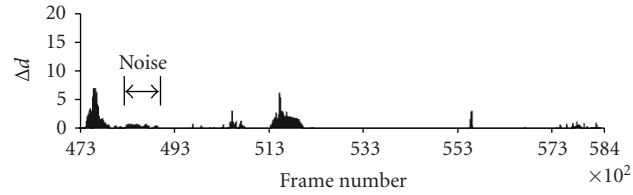

Figure 12: Detection results with the 7 moment-based features f_1 to f_7.

The second evaluation dataset is a street scene obtained from the BEHAVE Interactions Test Case Scenarios [56]. BEHAVE is funded by the UK's Engineering and Physical Science Research Council project. It involves nine different activities such as Walk Together and Run Together in the image sequences. The definitions of these nine activities are listed in Table 4. The BEHAVE dataset has eight video sequences, each containing a different combination of activities. The training image sequence is Sequence 0 from the BEHAVE dataset, which contains six activities of In-Group, Approach, Walk Together, Ignore, Split, and Following. The demonstration images for these six activities are shown in Figure 10(a1)–10(a6). The testing video is Sequence 5 that contains seven events of Approach, Ignore, Walk Together, Split, Meet, Run together, and Fight. In Sequence 5, the three

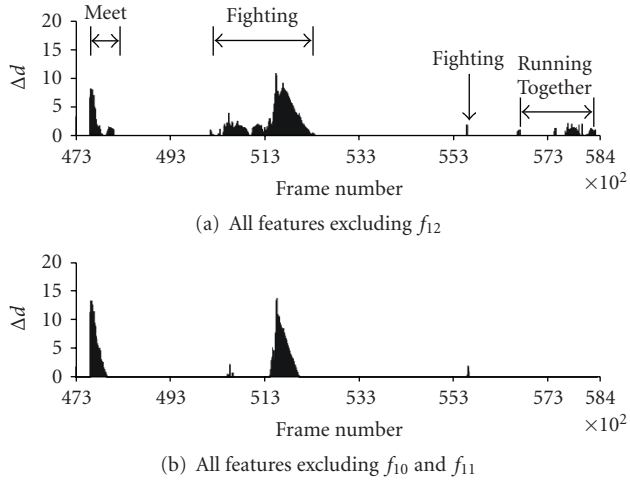

Figure 13: Detection results based on: (a) all features excluding f_{12}; (b) all features excluding f_{10} and f_{11}.

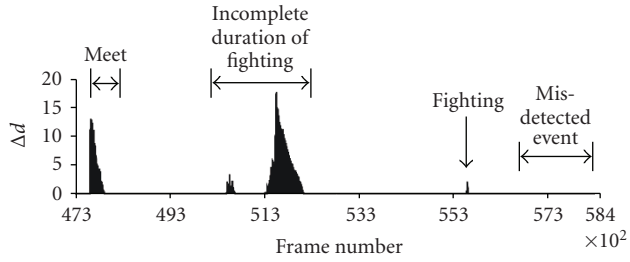

Figure 14: Detection results by K-means for Sequence 5 in the BEHAVE dataset.

activities of Meet, Fight, and Run Together are not included in the training sequence 0. Therefore, these three activities are treated as abnormal events. The demonstration images of these three abnormal events are shown in Figure 10(b1)–10(b3).

The BEHAVE video images are captured at 25 frames per second. The activities and video lengths of the training and testing sequences are listed in Table 5. There are a total of 11,200 frames (7 minutes and 27seconds) in Sequence 0, of which 50% (i.e., 5,600 image frames) are randomly sampled and used as the training samples. The update rate γ is set at 0.999 to construct the motion energy maps. The total number of clusters $C_{\max}$ is given by 40. The resulting minimum control constant β from the training process is 0.1. The test video of Sequence 5 has a total of 11,100 frames (7 minutes and 24 seconds).

When the training is done, the whole video images of Sequence 5 are used to test for the detection performance. Figure 11 illustrates the testing result of Sequence 5. In the figure, the x-axis presents the frame number and the y-axis is the excessive distance Δd. The results show that all the normal activities of In-Group, Approach, Walk Together, and Split in Sequence 5 are within the control limits. There are four major abnormal events detected in Sequence 5, that is, one long fighting event, one short fighting event, one meeting event and one running together event. The resulting distances of the four abnormal events are distinctly large and prolong for their corresponding durations, as seen in the x-axis and the y-axis in Figure 11. The running together event includes many discrete running activities where people abruptly enter and exit from the street scene and, therefore, the resulting distances Δd are not continuous.

We have also conducted additional experiments on the BEHAVE dataset with various combinations of features. Figure 12 shows the detection results using only the seven moment-based features. It fails to detect the subtle activity of Running Together. Noise is also created. The long Fighting activity is not alarmed at the beginning of the duration, and the overall discrimination magnitudes for the abnormal activities are reduced.

Since feature 12 is the ratio of f_{10} and f_{11}, we also evaluate the detection performance without including f_{12} in the feature set. As seen in Figure 13(a), the four abnormal activities Meet, long Fighting, short Fighting and Running Together are also well detected without the use of feature f_{12}. Comparing the detection results between Figures 11 and 13(a), the use of 12 full features gives higher discrimination magnitudes, especially in the case of short Fighting. We have also performed the detection task by excluding features f_{10} and f_{11} (and retaining all the remaining 10 features for classification). Figure 13(b) shows the detection results. The Running Together event is misdetected, and the whole duration of the long Fighting is not fully detected.

We have also used principal component analysis (PCA) for feature selection. It finds the eigenvalues of the 12 features, and sorts the features in descending order of their corresponding eigenvalues. Then 12 feature sets, each containing the dominant features from 1 to 12, are individually evaluated. The detection results consistently indicate that the use of 12 full features generates the highest discrimination magnitudes. The discrimination power is significantly reduced when less number of features is used for classification.

In order to evaluate the clustering performance between Fuzzy C-means and K-means, we have also used K-means for clustering and classification, and tested it on the BEHAVE dataset. We replicated 10 times with different random initial solutions for both FCM and K-means. Figure 14 shows representative detection results of sequence 5 in the BEHAVE dataset. The K-means technique is less responsive to abnormal activities. Compared to the detection results of FCM in Figure 11, K-means clustering procedure misdetects the subtle activity of Running Together. The duration of the long Fighting activity is not fully detected, and the discrimination magnitude of the short Fighting is less significant. Under the same termination criterion, K-means needs additional 40% computation time to converge.

4. Conclusions

Analysis of events has been conventionally based on the recognition of a set of predefined activities. In a scene of daily life such as home and office, it is extremely difficult to define and model every possible activity in advance. In this paper, we have proposed a macro-observation approach to

detect abnormal events such as burglary and fighting in daily life. The proposed motion energy map can simultaneously represent both spatial context and temporal context of an activity. All historical image frames are taken into account with exponential weights to construct the energy map. It alleviates the limitation on the use of a fixed duration for various activities with different paces. The constrained clustering model can effectively divide numerous activities in daily life into groups based on their similarity in energy maps. By training a sufficient number of randomly sampled energy maps in a video sequence that spans sufficient repetition of day-to-day activities, all normal events can be effectively represented by the cluster centroids. It allows fast computation of similarity measure for each new scene image. The proposed method can therefore be applied for on-line, real-time monitoring of unpredictable abnormal events in daily life.

The merit of this paper is to show the feasibility of the easily-implemented macro-observation approach for abnormality detection in daily life. The proposed scheme in its present form can well detect abnormal events with prolonged durations, especially those lasting tens of seconds or more. It is not highly responsive to the events that last only a few seconds. It is worth further investigation on the spatiotemporal representation and similarity metric for the analysis of short-term activities in daily life.

References

[1] J. Yamato, J. Ohya, and K. Ishii, "Recognizing human action in time sequential images using hidden Markov models," in *Proceedings of the International Conference on Pattern Recognition (ICPR '92)*, pp. 379–385, 1992.

[2] R. Polana and R. Nelson, "Low level recognition of human motion (or how to get your man without finding his body parts)," in *Proceedings of the IEEE Workshop on Motion of Non-Rigid and Articulated Objects*, pp. 77–82, Austin, Tex, USA, November 1994.

[3] C. R. Wren, A. Azarbayejani, T. Darrell, and A. P. Pentland, "Pfinder: real-time tracking of the human body," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 780–785, 1997.

[4] H. Roh, S. Kang, and S.-W. Lee, "Multiple people tracking using an appearance model based on temporal color," in *Proceedings of the International Conference on Pattern Recognition (ICPR '00)*, pp. 643–646, 2000.

[5] I. Haritaoglu, D. Harwood, and L. S. Davis, "W^4: real-time surveillance of people and their activities," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 809–830, 2000.

[6] Y. Chen, Y. Rui, and T. S. Huang, "JPDAF based HMM for real-time contour tracking," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '01)*, pp. 543–550, December 2001.

[7] I. Haritaoglu, D. Harwood, and L. S. Davis, "Ghost: a human body part labeling system using silhouettes," in *Proceedings of the International Conference on Pattern Recognition (ICPR '98)*, pp. 77–82, 1998.

[8] T. B. Moeslund and E. Granum, "A survey of computer vision-based human motion capture," *Computer Vision and Image Understanding*, vol. 81, no. 3, pp. 231–268, 2001.

[9] C.-W. Chu, O. C. Jenkins, and M. J. Matarić, "Markerless kinematic model and motion capture from volume sequences," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '03)*, pp. 475–482, June 2003.

[10] G. J. Brostow, I. Essa, D. Steedly, and V. Kwatra, "Novel skeletal representation for articulated creatures," in *Proceedings of the 8th European Conference On Computer Vision (ECCV '04)*, vol. 3023 of *Lecture Notes in Computer Science*, pp. 66–79, May 2004.

[11] R. Ishiyama, H. Ikeda, and S. Sakamoto, "A compact model of human postures extracting common motion from individual samples," in *Proceedings of the 18th International Conference on Pattern Recognition (ICPR '06)*, vol. 1, pp. 187–190, August 2006.

[12] A. Sundaresan and R. Chellappa, "Segmentation and probabilistic registration of articulated body models," in *Proceedings of the 18th International Conference on Pattern Recognition (ICPR '06)*, vol. 2, pp. 92–95, August 2006.

[13] C.-C. Chen, J.-W. Hsieh, Y.-T. Hsu, and C.-Y. Huang, "Segmentation of human body parts using deformable triangulation," in *Proceedings of the 18th International Conference on Pattern Recognition (ICPR '06)*, vol. 1, pp. 355–358, August 2006.

[14] R. Navaratnam, A. Thayananthan, P. H. S. Torr, and R. Cipolla, "Hierarchical part-based human body post estimation," in *Proceedings of British Machine Vision Conference (BMVC '05)*, pp. 479–488, 2005.

[15] C. Bregler, "Learning and recognizing human dynamics in video sequences," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '97)*, pp. 568–574, June 1997.

[16] Y. Yacoob and M. J. Black, "Parameterized modeling and recognition of activities," *Computer Vision and Image Understanding*, vol. 73, no. 2, pp. 232–247, 1999.

[17] N. M. Oliver, B. Rosario, and A. P. Pentland, "A Bayesian computer vision system for modeling human interactions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 831–843, 2000.

[18] M. Brand and V. Kettnaker, "Discovery and segmentation of activities in video," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 844–851, 2000.

[19] S. S. Intille and A. F. Bobick, "Recognizing planned, multiperson action," *Computer Vision and Image Understanding*, vol. 81, no. 3, pp. 414–445, 2001.

[20] H. Buxton, "Learning and understanding dynamic scene activity: a review," *Image and Vision Computing*, vol. 21, no. 1, pp. 125–136, 2003.

[21] C. P. Town, "Ontology-driven Bayesian networks for dynamic scene understanding," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '04)*, vol. 7, p. 116, 2004.

[22] A. F. Bobick and A. D. Wilson, "A state-based approach to the representation and recognition of gesture," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 12, pp. 1325–1337, 1997.

[23] Q. Dong, Y. Wu, and Z. Hu, "Gesture segmentation from a video sequence using greedy similarity measure," in *Proceedings of the 18th International Conference on Pattern Recognition (ICPR '06)*, vol. 1, pp. 331–334, August 2006.

[24] N. V. Boulgouris, K. N. Plataniotis, and D. Hatzinakos, "Gait recognition using linear time normalization," *Pattern Recognition*, vol. 39, no. 5, pp. 969–979, 2006.

[25] N. V. Boulgouris and Z. X. Chi, "Human gait recognition based on matching of body components," *Pattern Recognition*, vol. 40, no. 6, pp. 1763–1770, 2007.

[26] M. Shah, O. Javed, and K. Shafique, "Automated visual surveillance in realistic scenarios," *IEEE Multimedia*, vol. 14, no. 1, pp. 30–39, 2007.

[27] R. Cutler and M. Turk, "View-based interpretation of real-time optical flow for gesture recognition," in *Proceedings of the International Conference Automatic Face and Gesture Recognition*, pp. 416–421, 1998.

[28] M. J. Black, "Explaining optical flow events with parameterized spatio-temporal models," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '99)*, pp. 326–332, June 1999.

[29] Y. Ke, R. Sukthankar, and M. Hebert, "Efficient visual event detection using volumetric features," in *Proceedings of the 10th IEEE International Conference on Computer Vision (ICCV '05)*, pp. 166–173, October 2005.

[30] T. Ogata, W. Christmas, J. Kittler, and S. Ishikawa, "Improving human activity detection by combining multi-dimensional motion descriptors with boosting," in *Proceedings of the 18th International Conference on Pattern Recognition (ICPR '06)*, pp. 295–298, August 2006.

[31] A. A. Efros, A. C. Berg, G. Mori, and J. Malik, "Recognizing action at a distance," in *Proceedings of the 9th IEEE International Conference on Computer Vision*, pp. 726–733, October 2003.

[32] N. Johnson and D. Hogg, "Learning the distribution of object trajectories for event recognition," *Image and Vision Computing*, vol. 14, no. 8, pp. 609–615, 1996.

[33] A. Madabhushi and J. K. Aggarwal, "A Bayesian approach to human activity recognition," in *Proceedings of IEEE International Workshop on Visual Surveillance (VS '99)*, pp. 25–32, 1999.

[34] J. Owens and A. Hunter, "Application of the self-organizing map to trajectory classification," in *Proceedings of IEEE International Workshop on Visual Surveillance (VS '00)*, pp. 77–83, 2000.

[35] T. W. Liao, "Clustering of time series data—a survey," *Pattern Recognition*, vol. 38, no. 11, pp. 1857–1874, 2005.

[36] C. Piciarelli and G. L. Foresti, "On-line trajectory clustering for anomalous events detection," *Pattern Recognition Letters*, vol. 27, no. 15, pp. 1835–1842, 2006.

[37] C. Stauffer and W. E. L. Grimson, "Learning patterns of activity using real-time tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 747–757, 2000.

[38] H. Murase and R. Sakai, "Moving object recognition in eigenspace representation: gait analysis and lip reading," *Pattern Recognition Letters*, vol. 17, no. 2, pp. 155–162, 1996.

[39] M. J. Black and A. D. Jepson, "EigenTracking: robust matching and tracking of articulated objects using a view-based representation," *International Journal of Computer Vision*, vol. 26, no. 1, pp. 63–84, 1998.

[40] M. M. Rahman and S. Ishikawa, "Recognizing human behaviors employing global eigenspace," in *Proceedings of the International Conference on Pattern Recognition (ICPR '02)*, 2002.

[41] J. Wei, "Video content classification based on 3-D eigen analysis," *IEEE Transactions on Image Processing*, vol. 14, no. 5, pp. 662–673, 2005.

[42] M. M. Rahman and S. Ishikawa, "Human motion recognition using an eigenspace," *Pattern Recognition Letters*, vol. 26, no. 6, pp. 687–697, 2005.

[43] J. W. Davis and A. F. Bobick, "Representation and recognition of human movement using temporal templates," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '97)*, pp. 928–934, June 1997.

[44] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 3, pp. 257–267, 2001.

[45] G. R. Bradski and J. W. Davis, "Motion segmentation and pose recognition with motion history gradients," *Machine Vision and Applications*, vol. 13, no. 3, pp. 174–1843, 2002.

[46] S.-F. Wong and R. Cipolla, "Continuous gesture recognition using a sparse Bayesian classifier," in *Proceedings of the 18th International Conference on Pattern Recognition (ICPR '06)*, vol. 1, pp. 1084–1087, China, August 2006.

[47] J. Davis and A. Bobick, "Virtual PAT: a virtual personal aerobics trainer," in *Proceedings of Perceptual User Interfaces*, pp. 13–18, 1998.

[48] C. Shan, Y. Wei, X. Qiu, and T. Tan, "Gesture recognition using temporal template based trajectories," in *Proceedings of the 17th International Conference on Pattern Recognition (ICPR '07)*, vol. 3, pp. 954–957, August 2004.

[49] N. Vaswani, A. R. Chowdhury, and R. Chellappa, "Activity recognition using the dynamics of the configuration of interacting objects," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '03)*, vol. 2, pp. 633–640, June 2003.

[50] W. Hu, D. Xie, T. Tan, and S. Maybank, "Learning activity patterns using fuzzy self-organizing neural network," *IEEE Transactions on Systems, Man, and Cybernetics B*, vol. 34, no. 3, pp. 1618–1626, 2004.

[51] D. J. Fleet, M. J. Black, Y. Yacoob, and A. D. Jepson, "Design and use of linear models for image motion analysis," *International Journal of Computer Vision*, vol. 36, no. 3, pp. 171–193, 2000.

[52] E. L. Andrade, R. B. Fisher, and S. Blunsden, "Detection of emergency events in crowded scenes," in *Proceedings of the Institution of Engineering and Technology Conference on Crime and Security*, vol. 2, pp. 528–532, London, UK, 2006.

[53] A. Adam, E. Rivlin, I. Shimshoni, and D. Reinitz, "Robust real-time unusual event detection using multiple fixed-location monitors," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 3, pp. 555–560, 2008.

[54] H. Zhong, J. Shi, and M. Visontai, "Detecting unusual activity in video," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '04)*, vol. 2, pp. 819–826, July 2004.

[55] N. Rea, R. Dahyot, and A. Kokaram, "Semantic event detection in sports through motion understanding," in *Proceedings of the 3rd International Conference on Image and Video Retrieval (CIVR '04)*, vol. 3115 of *Lecture Notes in Computer Science*, pp. 88–97, July 2004.

[56] "BEHAVE Interactions Test Case Scenarios," University of Edinburgh, 2007, http://groups.inf.ed.ac.uk/vision/BEHAVEDATA/INTERACTIONS/.

[57] C. Stauffer and W. E. L. Grimson, "Adaptive background mixture models for real-time tracking," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '99)*, vol. 2, pp. 246–252, June 1999.

[58] P. KaewTrakulPong and R. Bowden, "An improved adaptive background mixture model for real-time tracking with shadow detection," in *Proceedings of the 2nd European Workshop on Advanced Video Based Surveillance Systems*, pp. 149–158, Kingston, UK, 2001.

[59] A. Elgammal, R. Duraiswami, D. Harwood, and L. S. Davis, "Background and foreground modeling using nonparametric kernel density estimation for visual surveillance," *Proceedings of the IEEE*, vol. 90, no. 7, pp. 1151–1162, 2002.

[60] M.-K. Hu, "Visual pattern recognition by moment invariants," *IEEE Transactions on Information Theory*, vol. 8, no. 2, pp. 179–187, 1962.

[61] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*, Kluwer Academic Publishers, Norwell, Mass, USA, 1981.