Activity Analysis, Summarization, and Visualization for Indoor Human Activity Monitoring

Zhongna Zhou, Xi Chen, Yu-Chia Chung, Zhihai He, *Senior Member, IEEE*, Tony X. Han, and James M. Keller, *Fellow, IEEE*

***Abstract*—In this work, we study how continuous video monitoring and intelligent video processing can be used in eldercare to assist the independent living of elders and to improve the efficiency of eldercare practice. More specifically, we develop an automated activity analysis and summarization for eldercare video monitoring. At the object level, we construct an advanced silhouette extraction, human detection and tracking algorithm for indoor environments. At the feature level, we develop an adaptive learning method to estimate the physical location and moving speed of a person from a single camera view without calibration. At the action level, we explore hierarchical decision tree and dimension reduction methods for human action recognition. We extract important ADL (activities of daily living) statistics for automated functional assessment. To test and evaluate the proposed algorithms and methods, we deploy the camera system in a real living environment for about a month and have collected more than 200 hours (in excess of 600 G bytes) of activity monitoring videos. Our extensive tests over these massive video datasets demonstrate that the proposed automated activity analysis system is very efficient.**

***Index Terms*—Action recognition, activity analysis, eldercare, video summarization.**

I. Introduction

VIDEO-BASED activity monitoring, coupled with intelligent video processing, has found many important applications in security monitoring, battlefield surveillance, environmental tracking, and health monitoring [4]. In this work, we study persistent video activity monitoring in an indoor environment for eldercare. We develop automated human activity analysis, summarization, and visualization methods to improve the efficiency of eldercare practice and to assist the independent living of elders.

Our society is increasingly aging, and elderly people desire to live as independently as possible [1]. But independent lifestyles often come with risks. With age-associated functional declines in mobility, cognition, and the senses, older adults (especially those in the ages of 80's and 90's) are exposed to various risks, and even harmful situations [1], [2]. A critical element in eldercare to maintain their independence of living is persistent activity monitoring and early identification of abnormal activities that indicate changing conditions, health crisis, or emergency situations. During the past decades, many *smart home* technologies have been developed. A variety of sensors, such as gait monitors, motion sensors, and radio-frequency devices, have been designed to monitor activities of elderly persons at home and to assist their independent living [1], [2]. In this work, we explore another important approach: *video-based activity monitoring and functional assessment for eldercare*. The video data, coupled with intelligent computer vision and learning algorithms [4], provides a rich and unique set of information that cannot be obtained from other types of sensors. For example, gait monitors and motion sensors often fail to distinguish between a person and an object falling. However, with video-based activity monitoring, this can be determined using human motion tracking [4].

A. Major Issues in Video-Based Activity Monitoring for Eldercare

Video-based activity monitoring for eldercare needs to deal with the following important issues. First, activity monitoring should be unobtrusive and privacy-protecting. For unobtrusiveness, the size and total number of cameras need to be as small as possible so that they can be easily and almost invisibly embedded in the living environment. In this work, we deploy a single small fisheye camera in the living room, as shown in Fig. 1. For privacy protection, we develop a fast and efficient silhouette extraction algorithm to extract and block-out the person in the video frame. In Section III, we will explain this in more detail. Second, in continuous activity monitoring for eldercare, as well as in many other video monitoring applications, such as aerial video surveillance of battlefields, security monitoring, and law enforcement [18], the data generated by the video monitoring network is voluminous. It is imperative to develop automated algorithms to extract important information from videos, aggregate and summarize these massive videos into compact activity records in a hierarchical database.

To address these issues, we propose to develop automated video processing algorithms 1) to detect and track human in a home-living environment and 2) to extract important ADL statistics and functional assessment data from videos.

B. Major Contributions of This Work

The major contribution of this work lies in the construction of a prototype system for real-time processing and automated analysis of indoor activity monitoring videos. Specific contributions are summarized as follows:

Manuscript received March 06, 2008; revised July 09, 2008. First published September 26, 2008; current version published October 29, 2008. This work was supported in part by National Institute of Health under Grant 5R21AG026412. This paper was recommended by Associate Editor D. Xu.

The authors is with the Department of Electrical and Computer Engineering, University of Missouri, Columbia, MO 65211 USA (e-mail: HeZhi@missouri.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSVT.2008.2005612

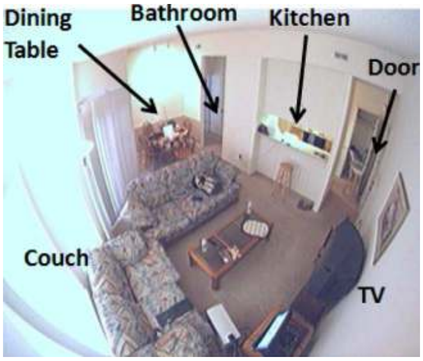


Fig. 1. A fisheye camera installed in a living room of a one-bedroom apartment.

1. We developed an accurate and robust silhouette extraction and human tracking algorithm which is able to effectively remove shadow and handle dynamic background changes in an indoor living environment.
2. We developed an adaptive learning and fuzzy inference system to estimate physical locations and moving speeds of persons from a single camera view without calibration.
3. Using hierarchical decision tree and dimension reduction methods, we developed an adaptive feature selection and human action recognition scheme to extract important activity statistics and functional assessment data from continuous activity monitoring videos.
4. We deployed the camera system in a real living environment for about a month and collected more 200 hours (about 600 G bytes) of video data. Our extensive tests on this massive test video set demonstrate the efficiency of the proposed algorithms.

C. Paper Organization

The rest of the paper is organized as follows. Section II gives an overview of the proposed scheme for automated activity analysis, summarization, and visualization. Section III presents our silhouette extraction and human detection algorithm. Section IV presents our algorithm for physical location and moving speed estimation. The action recognition and automated activity analysis are presented in Section V. Experimental results are presented in each technical section. Section VI concludes our paper.

II. Approach Overview and Data Collection

In this section, we provide an overview of the proposed approach for activity analysis, summarization, and visualization for eldercare video monitoring. We also discuss how we establish a dataset for algorithm testing and performance evaluation.

A. Overview of the Proposed Approach

Fig. 2 shows the proposed framework for automated activity analysis, summarization, and visualization for eldercare video monitoring. At the *object* level, we construct advanced algorithms for silhouette extraction, human detection and tracking in an indoor living environment. During algorithm development, we focus on two major issues: *shadow removal* and *adaptive background update*. At the feature level, we employ an adaptive learning method which is able to estimate physical motion parameters, such as 3-D location and speed, of the person from a single camera view. We also extract other features to describe the human activity in the spatiotemporal domain. At the *action* level, we propose a hierarchical decision tree and dimension reduction approach for recognizing major actions and for collecting important ADL statistics. At the *presentation* level, we summarize human activities over an extended period of time. In the following sections, we will explain each component in more detail.

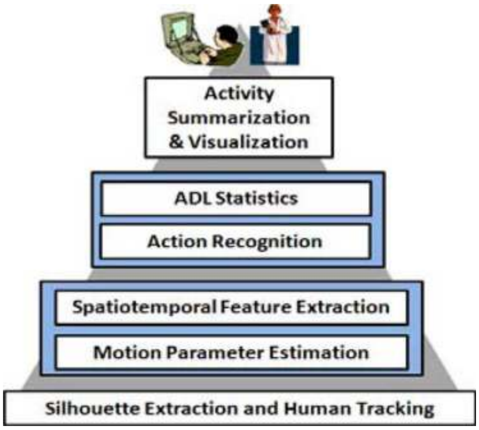

Activity Summarization & Visualization
ADL Statistics
Action Recognition
Spatiotemporal Feature Extraction
Motion Parameter Estimation
Silhouette Extraction and Human Tracking

Fig. 2. Overview of the proposed framework for automated activity analysis, summarization and visualization.

B. Data Collection

To test and evaluate the proposed algorithms and methods for automated activity analysis and summarization in eldercare, we need to build a large dataset of activity monitoring videos. To accomplish this, we rented a furnished one-bedroom apartment near our campus and asked 5 graduate students to live in the apartment, each for approximately 8 hours per day (daytime) for 5 days. We deployed a fisheye Unibrain camera which has a viewing angle of 180° in the living room, as shown in Fig. 1. From the camera, we can see the apartment door entry, a kitchen, a door to bathroom, a dining table, two couches, a coffee table, and a TV. We collected more than 200 hours of videos. This provides a sufficiently large amount of data for algorithm testing and performance evaluation. Once the algorithms and technologies in this research project are proven to be effective and robust, as our final objective, we will deploy cameras in 3–4 apartments at TigerPlace, an independent living facility at Columbia MO for aging in place [29].

III. Silhouette Extraction and Human Tracking

Silhouette extraction, namely, segmenting a human body or objects from a background, is the first and enabling step for many high-level vision analysis tasks, such as video surveillance, people tracking and activity recognition [4], [8], [23]. For effective privacy protection and accurate activity analysis, we need the silhouette extraction to be accurate and robust. It also needs to immediately reflect sudden scene changes such as objects being moved. A number of efficient silhouette extraction algorithms have been developed in the literature [24]. For background modeling and subtraction, models based on non-parametric kernel density estimation, mixture of Gaussians, least

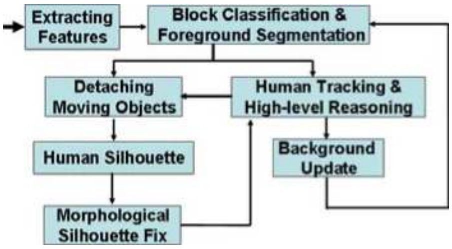

Extracting Features
Block Classification & Foreground Segmentation
Detaching Moving Objects
Human Tracking & High-level Reasoning
Human Silhouette
Background Update
Morphological Silhouette Fix

Fig. 3. The proposed silhouette extraction algorithm.

median of squares (LMeds), eigen-backgrounds, and illumination distribution analysis have been proposed [5], [6], [24], [25]. For human detection and tracking, a combination of shape analysis and tracking has been proposed in [4] and a Maximum a Posteriori Probability (MAP) approach has been introduced in [7]. An excellent review of background subtraction techniques is given in [24].

To design an efficient silhouette extraction algorithm for indoor living environments, the following issues need to be carefully addressed: (1) *time-varying light conditions*, (2) *strong shadow*, and (3) *Dynamic background changes*. Fig. 3 illustrates the proposed scheme for silhouette extraction. We consider silhouette extraction as an adaptive classification problem. We utilize image features which are invariant to changes in lighting conditions. High-level knowledge is fused with low-level feature-based classification results to handle time-varying backgrounds changes.

Extracting features to differentiate foreground objects from background is the first step of silhouette extraction. A basic requirement is that features should be invariant under brightness changes. Further, it should be effective in differentiating shadow from background. In this work, we use two features: *brightness distortion* and *chromaticity distortion*. More specifically, we extract features in the RGB color space [8]. For adaptive background update, we use the past Δ frames for background modeling. At each pixel location i, we compute the average values of its RGB components in the past Δ frames and denote them by vector E_i. We also calculate and standard deviations of the color components at each pixel. Let I_i be the pixel in the current frame. As shown in Fig. 4, we project the vector I_i onto vector E_i. We define brightness distortion α_i as

$$\begin{aligned}\alpha_i &= \arg\min_{\alpha_i} \|I_i - \alpha_i E_i\|^2 \\ &= \frac{\left(\frac{I_R(i)\mu_R(i)}{\sigma_{\mathrm{R}}^2(i)}\right) + \left(\frac{I_G(i)\mu_G(i)}{\sigma_{\mathrm{G}}^2(i)}\right) + \left(\frac{I_B(i)\mu_B(i)}{\sigma_{\mathrm{B}}^2(i)}\right)}{\left(\frac{\mu_R(i)}{\sigma_R(i)}\right)^2 + \left(\frac{\mu_G(i)}{\sigma_G(i)}\right)^2 + \left(\frac{\mu_B(i)}{\sigma_B(i)}\right)^2}\end{aligned} \quad (1)$$

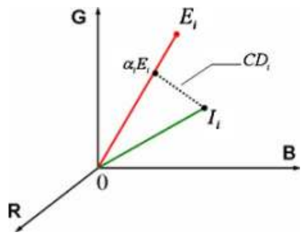


Fig. 4. Brightness and chromaticity distortion.

TABLE I
AVERAGE ERROR RATES ON TEST VIDEO SEQUENCES

Algorithms	Average Error Rate
Algorithm in [7]	12.1%
Our Algorithm	7.3%

and chromaticity distortion as shown in (2) at the bottom of the page, where $[I_R(i), I_G(i), I_B(i)]$ represent the values of red, green and blue components of the $i\mathrm{th}$ pixel in the RGB color space. $[\mu_R(i), \mu_G(i), \mu_B(i)]$ and $[\sigma_R(i), \sigma_G(i), \sigma_B(i)]$ are the mean and standard deviation of these color components. This color model separates the brightness from the chromaticity components as shown in Fig. 4. It has been found that the chromaticity distortion is invariant under brightness changes [7]. Our foreground-background classification is based on the following two observations: 1) *image pixels in the background often have little change in their chromaticity distortion*; and 2) *shadow often causes brightness distortion but little chromaticity distortion*. Based on these two observations, we establish the following decision rules for foreground, background, and shadow detection: 1) if the chromaticity distortion CD_i is large, I_i is a foreground pixel; 2) if the chromaticity distortion is small and the brightness distortion is about 1.0, it is a background pixel; 3) if chromaticity distortion is small and the brightness distortion smaller than 1.0, it is a shadow pixel.

To determine the thresholds for α_i and CD_i, we use the image data in the past Δ frames to compute the distributions of α_i and CD_i. Fig. 5 shows the average brightness distortion of all foreground, background, and shadow pixels in each frame of two video sequences. It also shows the distributions of chromaticity distortion. We model their distributions using Gaussian mixtures. The threshold is determined by the maximum likelihood method.

Fig. 6(c) shows two examples of segmentation and shadow removal using the proposed feature extraction and classification scheme. Here, we use $\Delta = 10$ frames. Fig. 6(b) shows the results obtained with the method in [7]. It can be seen that our method is very effective in detecting and removing shadows. To systematically evaluate the performance of our silhouette extraction algorithm, we selected 5 clips from our test video

$$CD_i = \|I_i - \alpha_i E_i\| = \sqrt{\left(\frac{I_R(i) - \alpha_i \mu_R(i)}{\sigma_R(i)}\right)^2 + \left(\frac{I_G(i) - \alpha_i \mu_G(i)}{\sigma_G(i)}\right)^2 + \left(\frac{I_B(i) - \alpha_i \mu_B(i)}{\sigma_B(i)}\right)^2} \quad (2)$$

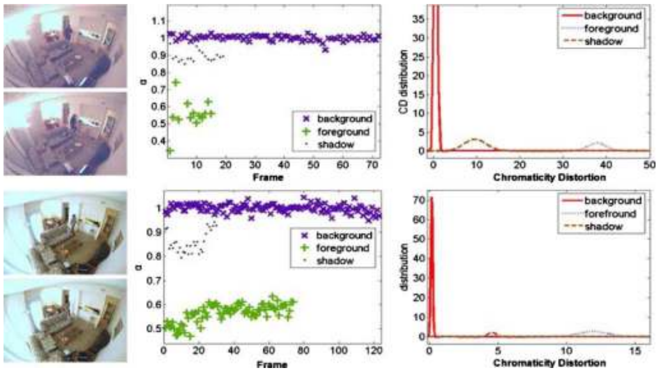


Fig. 5. The values of α_i and CD_i for two video sequences.

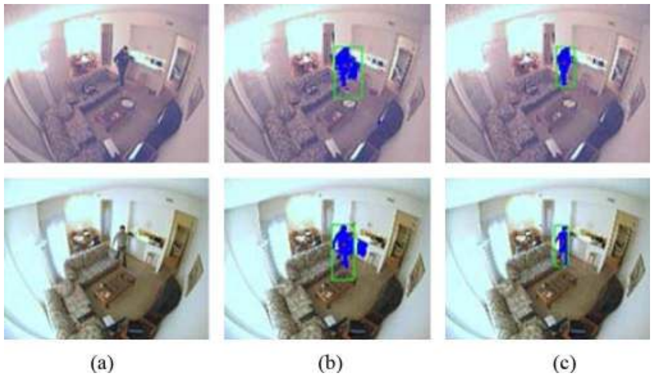


Fig. 6. Shadow removal. (a) Original frames. (b) Silhouette extraction using algorithm in [14]. (c) Silhouette extraction with the proposed scheme.

datasets (with a total of 150 video frames) and used manual segmentation results as ground truth. We calculate the error rate $e = N_d/N_T$ as a performance metric [23]. Here, N_d is the total number of pixels in the silhouette that are different from the ground-truth and N_T is the total number of pixels in the ground-truth silhouette. Table I compares the performance of our silhouette extraction algorithm with that in [7]. It can be seen that our algorithm achieves a significantly smaller error rate than the algorithm in [7].

A. Adaptive Background Update

In silhouette extraction within a dynamic video scene, we need to continuously update the background model by incorporating background changes. A commonly used method to update background is that, if an object or image area remains stationary for a certain period of time, it is considered to be background. Here, we use the past Δ frames to update the background model. For accurate silhouette extraction, we want Δ to be small so that the background can be quickly updated. However, when Δ is small, the human body could be easily updated as background if the person does not move for a while, for example, sitting still on a chair for a few minutes. To solve this problem, we propose to utilize high-level knowledge about human motion as a guideline to perform adaptive update of the background model.

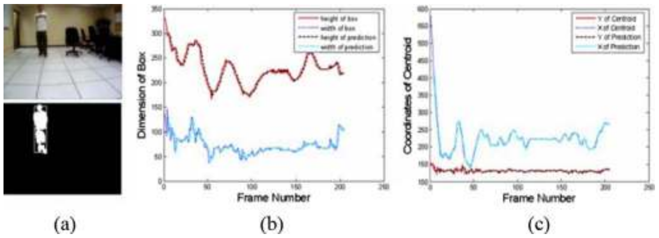


Fig. 7. Human tracking results. (a) A sample frame of the original video. (b) Estimated height and width of the bounding box compared with their ground truth. (c) Estimated center position (X, Y) of the bounding box compared with their ground truth.

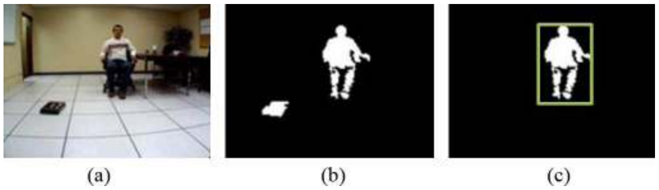


Fig. 8. (a) The original frame. (b) With non-adaptive background update. (c) With adaptive background update.

Many sophisticated human tracking algorithms have been developed in the literature [4], [7]. However, they often have high computational complexity. Here, to achieve low-complexity, we use a simple block-based motion estimation which has been extensively used in video coding [22]. More specifically, suppose that we have obtained the silhouette for frame n. We find a bounding box for the silhouette such that 95% of foreground pixels are included. For each image block within the bounding box in the current frame n, we find its best match in the next frame $n+1$ using SAD (sum of absolute difference) [22] as a distance metric. To speed up the motion estimation process, we use a fast algorithm called diamond search [22]. Once the motion vectors of all blocks are obtained after block-based motion estimation, we take their average to predict the human body position (or the center of its bounding box) in the next frame $n+1$. Those image blocks which contain the human body should be updated very slowly so that the human body won't be absorbed into the background. Those blocks outside the predicted body region can be updated much faster to make sure that new objects are quickly absorbed into the background. After background update and silhouette extraction, we update the dimension, height and width, of the bounding box in frame $n+1$.

Fig. 7 shows the human tracking performance. In Fig. 7(b), we plot the estimated height and width of the bounding box of human silhouette in each frame. We also plot their ground truth values obtained from manual segmentation. Fig. 7(c) shows the results for center position of the bounding box. It can be seen that the tracking and estimation are fairly accurate. Fig. 8 shows one example in which the person drops a book on the floor and sits on the chair for 5 minutes. We set Δ to be 30 frames. With the proposed adaptive background update, the book is quickly updated into the background while the person remains in the foreground silhouette.

IV. Location and Speed Estimation Using Statistical Learning

A key element in automated activity analysis for eldercare is the estimation of the physical location and moving speed of a person. Physical locations of the person in the room provide important contextual information for action recognition. If we know the person is at the dining table, he might be eating or

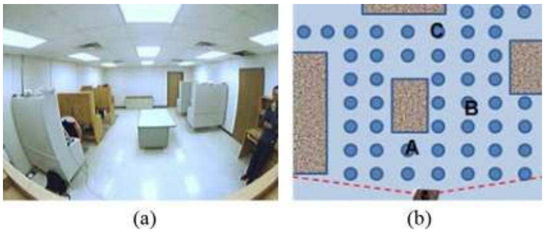


Fig. 9. (a) A fisheye camera deployed in our laboratory. (b) Labeled samples for person localization.

working. If the person in at the stove, he might be cooking. Moving speed is also an important feature for recognizing human actions, such as standing, walking, running, etc [5]. Furthermore, it is a critical variable in assessing physical functions, activity levels, and energy expenditure of elderly people at home [2].

A. Location and Speed Estimation Using Statistical Learning

In this work, we develop a statistical learning approach for estimating physical location and moving speed from a single fisheye camera without calibration. The major difficulties are severe nonlinear lens distortion and object occlusion. Our basic idea in statistical learning for location and speed estimation is: *when a person appears at the same physical location, his silhouette should look similar, even with lens distortion and object occlusion*. For example, if a person stands behind a desk and is partially occluded, his silhouettes should always look the same if the desk is not moved. After the camera is deployed, at the learning stage, we let the person walk all over the room, capture a sequence of video frames, and extract the corresponding silhouettes. For each silhouette, we extract four feature variables: its center $[x, y]$, width W and height H in the video frame. Based on these four feature variables, we use vector quantization methods, such as the MAX-Lloyd algorithm [21], to quantize this sequence of silhouettes into a number of prototypes. The total number of prototypes depends on the required estimation accuracy. According to our experience, 10–20 prototypes are sufficient for a typical room in a home. Each silhouette prototype corresponds to a physical location, as illustrated in Fig. 9(b). For each silhouette prototype, we choose one representative silhouette, visually determine where the person stands and manually measure its physical coordinates.

Once the prototypes are labeled with their physical coordinates, we use a fuzzy inference method to estimate the physical location of new video frames. We employ the adaptive fuzzy inference system (ANFIS), which is provided in MATLAB Fuzzy Logic Toolbox, to obtain the membership functions and decision rules for each physical location, as shown in Fig. 10. Note that, at each location, we have a membership function for each of the four feature variables $\{x, y, W, H\}$. For example, if the center position is $(x, y) = (72, 80)$ and the dimensions are $(W, H) = (30, 70)$, the person is mostly likely at location A. The proposed fuzzy inference method for location estimation works as follows: for the current video frame, we extract the silhouette and four feature variables $\{x, y, W, H\}$, which are then evaluated by the inference system. The outputs (or firing) of these membership functions are aggregated by the fuzzy inference system into a single number indicating the confidence of the person being at the corresponding location. The algorithm then chooses the location with the highest confidence as the estimation result.

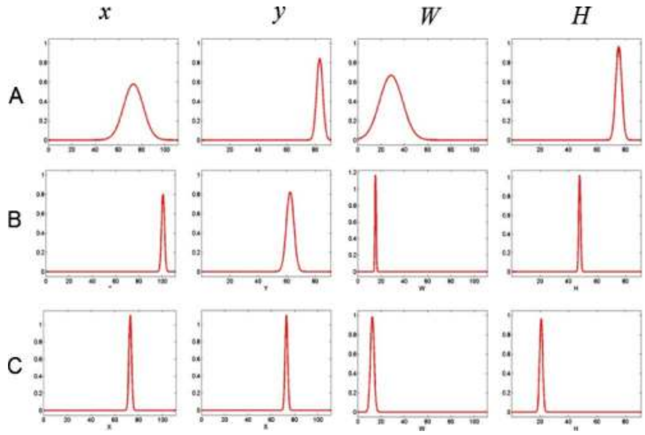


Fig. 10. Membership functions of $\{x, y, W, H\}$ at three locations A, B, and C.

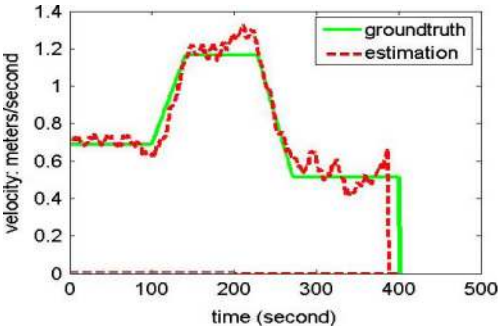


Fig. 11. Speed estimation results in comparison with ground truth.

Once the physical location is estimated, we can compute the average speed. More specifically, at frame n, we compute the displacement between the physical locations at frame $N - L$ and frame $N + L$. We then divide the displacement by frame time difference to obtain the average speed. Here, L is an estimation window. In this work, we set L to be 15 frames.

B. Location and Speed Estimation Results

To evaluate the performance of the proposed algorithm for physical location and moving speed estimation, we collected test videos in our lab, as shown in Fig. 9 and manually measured the ground truth. Fig. 11 shows the moving speed estimation results in comparison with the ground truth. It can be seen that the estimation error is very small, mostly less than 5%. Fig. 12 shows the actual moving trajectory of the person and the estimation result on the physical layout of the room. Notice that when the person is in the near field of view, the estimate is very accurate. However, when he or she is in the far field of view, the estimate is less accurate. Fig. 13 shows the results for another test video (with a new test subject). We then applied the silhouette extraction and moving speed estimation algorithm to the apartment video dataset explained in Section II-B.

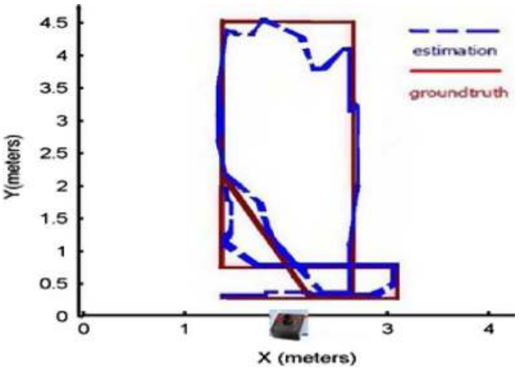


Fig. 12. Estimated moving trajectory in the room in comparison with ground truth.

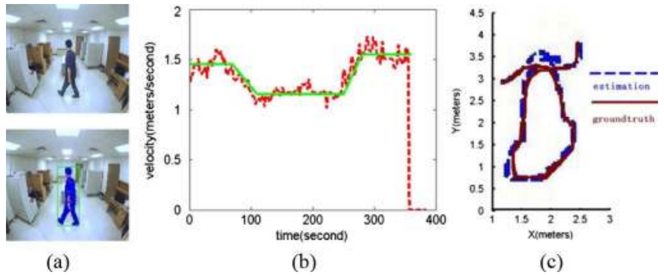


Fig. 13. Location and speed estimation for another test video sequence.

V. Human Action Recognition and Activity Analysis in an Indoor Environment

Human action recognition is a critical component in automated activity analysis and video scene understanding. Due to the non-rigidness of human body, the wide variety of possible human actions, as well as the deep structure of human activities, human action recognition remains a challenging task [13]. Most research work has been focused on feature selection, action modeling and classification. Spatiotemporal features for human activity analysis have been explored in [11], [12]. High-level features, such as spatiotemporal shape, skeleton, or body-part position, have been investigated in [13], [14]. Feature selection has been studied in [15]. Adaboost, and SVMs (Support Vector Machines) [12], [16] have been widely used for activity recognition and classification.

In this work, we aim to develop a low-complexity, efficient, and robust scheme for action recognition. Our objective is to convert the input activity monitoring video stream into meaningful information, such as ADL (Activity of Daily Living) statistics [2], for functional assessment. To extract this type of information, we must first identify the actions being performed in the input videos, including walking, sitting on the couch, standing up, sitting at the dining table, preparing meals in the kitchen, visiting the bathroom, going outdoors, etc.

We observe that, in an indoor living environment, the activities performed by a person often exhibit a hierarchical structure which repeats on a regular (daily and/or weekly) basis. To capture the deep structure of human activities, we propose to use Hierarchical Action Decision Tree (HADT) to classify the human actions. Similar to the classical decision tree, the HADT classifies human actions in a hierarchical manner using multiple-level features, such as location, speed, appearance, and primitive visual features. By confining the features used in the HADT to the available shape, spatial, temporal, appearance, and primitive visual features, we bound its complexity. Therefore, the over-fitting problem is alleviated. Besides, our hierarchical scheme can stop at any level to provide action classifications with different granularities, depending on the application requirements.

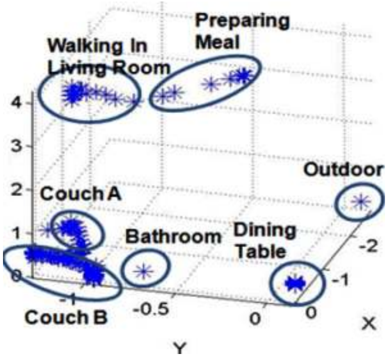


Fig. 14. Recognizing major activities of daily living using location and speed.

TABLE II
The Confusion Matrix for Human Action Recognition

	Being Outdoor	Bathroom Visit	Walk to Kitchen	Dining Table	Preparing Meals	Sitting on Couch	In front of TV
Being Outdoor	100%	0	0	0	0	0	0
Bathroom Visit	0	100%	0	0	0	0	0
Walk to kitchen	0	0.1%	99.9 %	0	0	0	0
Dining Table	0	0	0.13 %	99.87 %	0	0	0
Preparing Meals	0	0	0	0	100%	0	0
Sitting on Couch	0	0	0.04 %	0	0	99.96%	0
In front of TV	0	0	0.03 %	0	0	0	99.97 %

A. Level-1 Action Recognition Based on Physical Location and Moving Speed

At Level-1, we classify human actions based on physical location and speed. The physical location in the room provides important contextual information for action recognition. Based on these two features, using K-means clustering, we are able to identify major activities of daily living, such as walking, visiting the bathroom, sitting on the couch, preparing meals, or sitting at the dining table, as illustrated in Fig. 14. We then use the K-nearest neighbor (KNN) method to determine the human action in the input video segment.

The following results are based upon extensive evaluations on the video datasets described in Section II-B. Table II shows the confusion matrix of our Level-1 human action recognition scheme. It can be seen that the recognition is very accurate. Fig. 15 displays the sequence of actions performed by two persons over a one-hour period. It can be seen that person B is much more active than A. Fig. 16 shows the action statistics of four

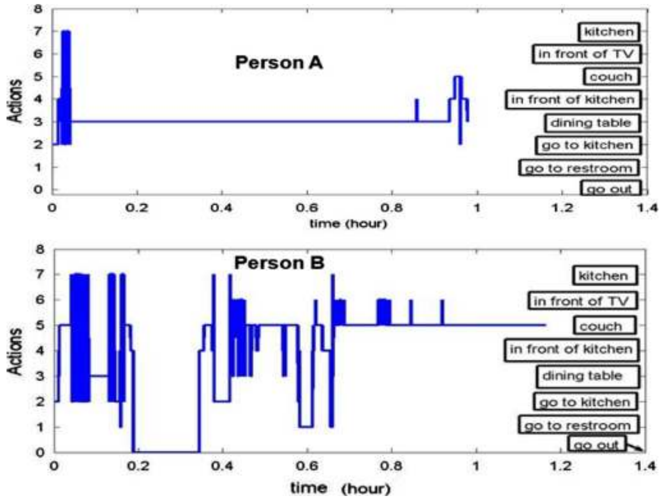


Fig. 15. Action sequences of two persons over a one-hour period.

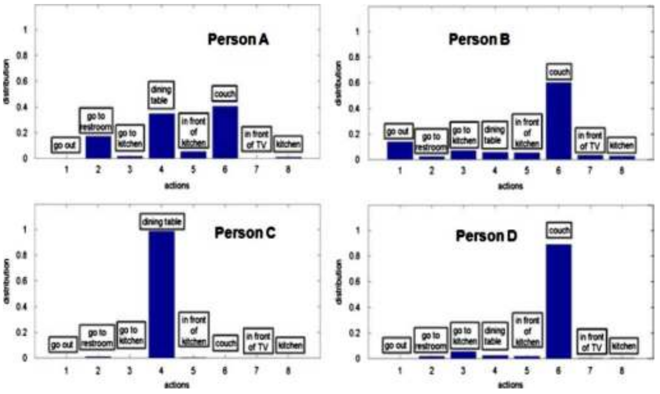


Fig. 16. Distribution of major actions of four persons with a 5-day period.

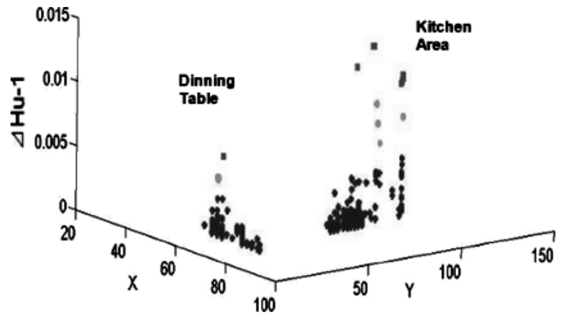


Fig. 17. Values of Hu moment when the person is sitting at the dining table and standing at the kitchen area.

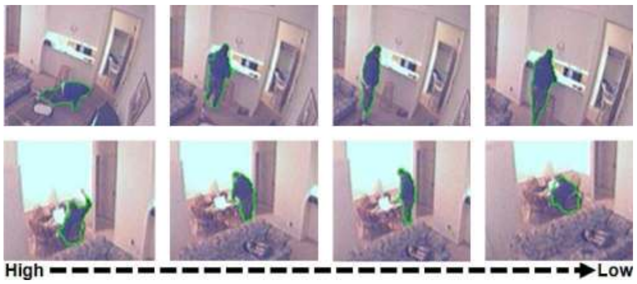


Fig. 18. Sample video frames of different levels of body motion when the person is standing in the kitchen area (top row) and sitting at the dining table (bottom row).

people over the entire data collection period (5 days). We can see that persons A and B are more active than the other two.

B. Level-2 Action Classification Based on Shape Features

At the second level, we recognize and classify more detailed actions. We observe that, at the same location, e.g., sitting at the dining table or standing in the kitchen, the person could perform different actions over times, such as eating, washing, cooking, or just sitting still. Because of the limited camera resolution and picture quality, we are not able to track the motion of body parts for detailed action recognition. Instead, as a part of our objective in functional assessment, we are interested in how active the person is. To differentiate different levels of body motion, we extract body shape information and analyze its variation over time. More specifically, using the silhouette extraction result as an initial step, we find the smoothed boundary of the human body using a snake model [27]. The Hu Moment Invariants (HMI) [26] are used to characterize the body shape in each video frame. Let $\boldsymbol{\varphi}_n$ be the HMI of frame n. Certainly, when body parts are moving, the body shape and its HMI values will change over time. Therefore, the temporal variation of the HMI values can be used to measure the level of body motion. We denote by this temporal variation by $\sigma_{Hu}[n]$, which is the variance of $\{\boldsymbol{\varphi}_{n-L}, \boldsymbol{\varphi}_{n-L+1}, \ldots, \boldsymbol{\varphi}_{n+L}\}$, where L is a window size. Four our experiments, we set $L = 30$ frames. Fig. 17 shows the value of $\sigma_{Hu}[n]$ for two video segments when the person in sitting at the dining table and standing in the kitchen area. Fig. 18 shows samples video frames ordered by the value of $\sigma_{Hu}[n]$. We can see that when the person falls on to the ground (in the top row) and is yawning and stretching his arms (in the bottom row), $\sigma_{Hu}[n]$ takes a larger value.

C. Level-3 Action Recognition Using Primitive Visual Features and Manifold Learning

Using body shape information and its temporal variation, we are able to differentiate different levels of body motion when the person remains at the same physical location. Level-3 recognition operates on top of Level-2 classification. More specifically, at Level-3, we are not interested in those segments where $\sigma_{Hu}[n]$ is small and the person is not moving much. Instead, we are interested in those video segments where the person's body parts are moving consistently while the person is sitting or standing at the same location. For example, the person is cooking, exercising, eating, or making a phone call. In these cases, the value of $\sigma_{Hu}[n]$ is consistently large. Thus, the Level-2 action classification serves as a filter for Level-3 action recognition. This helps reduce the overall computational complexity significantly since a vast majority of video segments do not have consistent body part motion and will be skipped by Level-3 action recognition.

Action recognition at this level of detail is a challenging task because these actions often require very sophistication features and feature selection methods. To address this issue, we propose to use primitive visual features and explore dimension reduction methods to discover the need features for action recognition.

As illustrated in Fig. 19, we partition the video frame into blocks (e.g.,8×8 blocks), denoted by $\{B_K | 1 \leq k \leq K\}$.

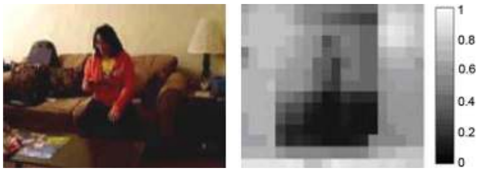


Fig. 19. Primitive visual features for Level-3 action recognition.

Using block-based motion search, we find the motion vector of each block. Let m_k be the magnitude of its motion vector. Define $\boldsymbol{X}_n = \{m_1, m_2, \ldots, m_K\}$, the set of primitive visual features that we extract from frame n. Note that this is a high-dimensional vector. For example, for a 640 × 480 video frame, the vector has 4800 dimensions. These primitive features are able to characterize low-level motion of body parts.

We then use locally linear embedding (LLE) to map these primitive features from a high-dimensional space into a low-dimension one to discover a small set of composite features for action recognition. LLE, proposed by Roweis and Saul [20], is able to preserve neighborhood relationships and nonlinear local structure among data points much more efficiently than other dimension reduction techniques, such as principal component analysis (PCA) and multi-dimensional scaling (MDS). In a recent work [28], LLE has been successfully used for gait and face recognition.

The proposed human action recognition method works as follows: first, we select a set of training video segments, each of which is manually labeled by the corresponding action. We extract the primitive features from these video segments. Using LLE, we map these features from the high-dimensional space into a low-dimensional space. In this low-dimensional space, a video segment becomes a trajectory. Using correlation-based matching and KNN classification, we classify and recognize human actions. In the following, we will explain this procedure in more detail.

Let $\{I_n | 1 \leq n \leq N\}$ be the training video sequence. At frame I_n, we extract the primitive feature vector $\boldsymbol{X}_n$ to describe the human activity in the spatiotemporal domain. In LLE, each feature vector is considered as a data point in the high-dimensional space. It is approximated by a weighted summation of its neighbors:

$$\boldsymbol{X}_n \cong \sum_k w_{nk} \boldsymbol{X}_k \tag{3}$$

where the weights W_{nk} summarize the contribution of the k-th data point $\boldsymbol{X}_k$ to the reconstruction of $\boldsymbol{X}_n$. In other words, these weights describe the local dependency between data points or their local structure. It should be noted that $\boldsymbol{X}_k$ should be within the neighborhood of $\boldsymbol{X}_n$ and $\sum_k W_n k = 1$. Here, we choose the 10 nearest data points as the neighbors of $\boldsymbol{X}_n$. To compute the weights W_{nk}, LLE minimizes the following cost function:

$$\boldsymbol{E}(\{W_{nk}\}) = \sum_n |\boldsymbol{X}_n - \sum_k W_{nk} \boldsymbol{X}_k|^2 \tag{4}$$

which is the total squared distances between all the data points and their reconstructions. This problem can be solved using a least mean squared error (LMSE) approach [10]. Once the local structures of data points are captured by weights W_{nk}, we are ready to map these data points into a low-dimensional space while preserving their local structure. Suppose the data point $\boldsymbol{X}_n$ is mapped to $\boldsymbol{Y}_n$ in a low-dimensional space with dimension d. In preserving the local structure, we need to determine $\boldsymbol{Y}_n$ which minimizes the following embedding cost function:

$$\boldsymbol{\Phi}(\boldsymbol{Y}) = \sum_n |\boldsymbol{Y}_n - \sum_k W_{nk} \boldsymbol{Y}_k|^2. \tag{5}$$

Here, W_{nk} are known from the previous step. This cost function can be minimized by solving a sparse $N \times N$ eigenvector problem [10]. One major advantage of LLE is that the procedure is intuitive, simple to implement, unsupervised, and does not involve local minima in its optimization.

Fig. 20 shows three test video sequences where the person is cooking in the kitchen, exercising in the living room, and brushing his teeth. Fig. 22(a)–(c) shows the primitive feature vectors mapped into a 3-D space. Each video frame corresponds to a point and the whole sequence is represented by a trajectory $\{\boldsymbol{Y}_n\}$ in this low-dimensional space. Fig. 21 shows a test video sequence performed by another person. Its trajectory in the low-dimensional space is shown in Fig. 22(d). We can see that it is very close to that in Fig. 22(a). We use the correlation

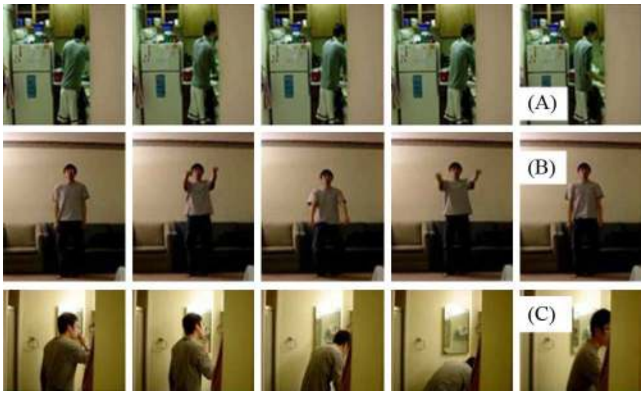


Fig. 20. Training video sequences for LLE-based action recognition.

Fig. 21. A test video sequence for LLE-based action recognition.

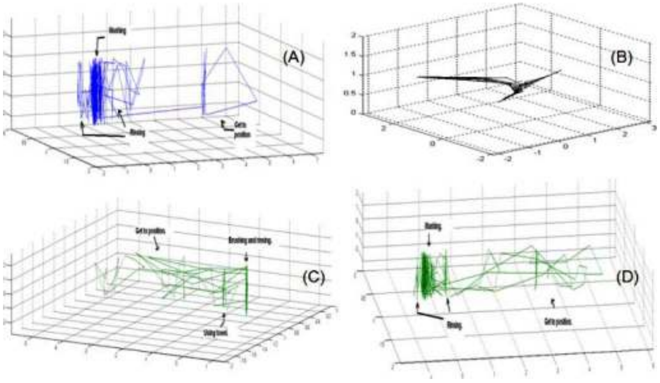


Fig. 22. Primitive features mapped into a low-dimensional space.

TABLE III
CORRELATION-BASED ACTION RECOGNITION RESULTS

	Actions (Labeled Training Videos)		
Test Videos	Cooking	Brushing Teeth	Exercise
Cooking	0.92	0.57	0.34
Brushing Teeth	0.57	0.84	0.40
Exercise	0.34	0.40	0.97

coefficient of $\{\boldsymbol{Y}_n\}$ between two video sequences as a distance metric. Table III contains the average correlation coefficients between the test video sequences and the training data. The action with the highest correlation with the test video is used as the recognition output. It can be seen that we can accurately recognize these three actions.

VI. CONCLUSION

In this paper, we proposed an automated activity analysis and summarization scheme for video-based eldercare monitoring. We developed a silhouette extraction algorithm which is able to efficiently handle shadows and dynamic background changes. This is followed by a fast and reliable physical location and moving speed estimation algorithm using single-view camera videos without camera calibration. We then developed a manifold learning and dimension reduction scheme for human action recognition to extract important activity statistics for functional assessment. Our extensive simulation results demonstrated that the proposed algorithms and methods are very efficient and accurate.

In our next step of research, we plan to deploy camera in 3–4 apartment at TigerPlace and evaluate the proposed algorithms and methods in a real eldercare setting. Our current system focuses on activity analysis of a single person. As our future work, we need to address the issue of multiple persons. We shall also develop advanced data mining algorithms to explore more activity patterns in the massive activity data and link them with medical records for automated functional assessment and early identification of potential health problems. Finally, we need to conduct a series of focus group sessions with eldercare professionals to evaluate the effectiveness of the proposed activity analysis and summarization system.

REFERENCES

[1] M. A. Cohen and J. Miller, *The Use of Nursing Home and Assisted Living Facilities Among Privately Insured and Non-Privately Insured Disabled Elders*. Washington, DC: Government Printing Office, 2000.

[2] M. Alwan, S. Kell, S. Dalal, B. Turner, D. Mack, and R. Felder, “In-home monitoring system and objective ADL assessment: Validation study,” presented at the Int. Conf. Independence, Aging Disability, Washington, DC, Dec. 2003.

[3] S. Intille, “Designing a home of the future,” *IEEE Pervasive Computing*, pp. 80–86, Apr.-Jun. 2002.

[4] I. Haritaoglu, D. Harwood, and L. S. Davis, “W4: Real-time surveillance of people and their activities,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 809–830, Aug. 2000.

[5] C. Stauffer and W. E. L. Grimson, “Learning patterns of activity using real-time tracking,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 747–757, Aug. 2000.

[6] L. Wang, T. Tan, H. Ning, and W. Hu, “Silhouette analysis-based gait recognition for human identification,” *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 1505–1518, Dec. 2003.

[7] C. Wren, A. Azarbayejani, T. Darrell, and A. P. Pentland, “Pfinder: Real-time tracking of the human body,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 780–785, Jul. 1997.

[8] T. Horprasert, D. Harwood, and L. Davis, “A statistical approach for real-time robust background subtraction and shadow detection,” presented at the IEEE ICCV’99 FRAME_RATE WORKSHOP, Kerkyra, Greece, Sep. 1999.

[9] L. Rabiner, “A tutorial on hidden Markov models and selected applications in speech recognition,” *Proc. IEEE*, vol. 77, pp. 257–286, Feb. 1989.

[10] S. Roweis and L. Saul, “Nonlinear dimensionality reduction by locally linear embedding,” *Science*, vol. 290, no. 5500, pp. 2323–2326, Dec. 2000.

[11] E. Shechtman and M. Irani, “Space-time behavior based correlation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, San Diego, CA, Jun. 2005, pp. 405–412.

[12] C. Schüldt, I. Laptev, and B. Caputo, “Recognizing human actions: A local SVM approach,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Cambridge, U.K., Aug. 2004, pp. 32–36.

[13] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, “Actions as space-time shapes,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2005, vol. 2, pp. 1395–1402.

[14] Y. A. Ivanov and A. F. Bobick, “Recognition of visual activities and interactions by stochastic parsing,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 852–872, 2000.

[15] P. C. Ribeiro and J. Santos-Victor, “Human activity recognition from video: Modeling, feature selection and classification architecture,” in *Proc. Int. Workshop on Human Activity Recognit. Modeling*, 2005, pp. 61–70.

[16] P. Viola and M. Jones, “Rapid object detection using a boosted cascade of simple features,” in *IEEE Conf. Comput. Vis. Pattern Recognit.*, 2001, vol. 1, pp. 511–518.

[17] C. Stauffer and E. Grimson, “Learning patterns of activity using real-time tracking,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 747–757, Aug. 2000.

[18] “Special issue on video communications, processing, and understanding for third generation surveillance systems,” *Proc. IEEE*, vol. 89, pp. 1355–1539, Oct. 2001.

[19] G. Medioni, I. Cohen, F. B. Brémond, S. Hongeng, and R. Nevatia, “Event detection and analysis from video streams,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 8, pp. 873–889, 2001.

[20] M. J. Rantz, R. T. Porter, D. Cheshier, D. Otto, C. H. Survey, III, R. A. Johnson, M. Skubic, H. Tyrer, Z. He, G. Demiris, J. Lee, G. L. Alexander, and G. Taylor, “TigerPlace, a state-academic-private project to revolutionize traditional long term care,” *J. Housing for the Elderly*, Oct. 2007.

[21] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*. Boston, MA: Kluwer, 1991.

[22] S. Zhu and K.-K. Ma, “A new diamond search algorithm for fast block matching motion estimation,” *IEEE Trans. Image Process.*, vol. 9, pp. 287–290, Feb. 2000.

[23] D. Xu, J. Liu, X. Li, Z. Liu, and X. Tang, “Insignificant shadow detection for video segmentation,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, pp. 1058–1064, 2005.

[24] M. Piccardi, “Background subtraction techniques: A review,” presented at the IEEE SMC 2004 Int. Conf. Syst., Man Cybern., The Hague, The Netherlands, Oct. 2004.

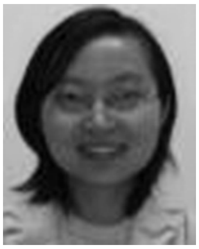
[25] R. Cucchiara, C. Granan, M. Piccardi, and A. Prati, “Detecting moving objects, ghosts, and shadows in video streams,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 10, pp. 1337–1342, Oct. 2003.

[26] M. Hu, “Visual pattern recognition by moment invariants,” *IRE Trans. Inf. Theory*, vol. 8, no. 2, pp. 179–187, 1962.

[27] C. Xu and J. L. Prince, “Snakes, shapes, and gradient vector flow,” *IEEE Trans. Image Process.*, vol. 7, pp. 359–369, Mar. 1998.

[28] X. Li, S. Lin, S. Yan, and D. Xu, “Discriminant locally linear embedding with high order tensor data,” *IEEE Trans. Syst., Man, Cybern.*, vol. 38, no. 2, pp. 342–352, Apr. 2008.

[29] TigerPlace. [Online]. Available: http://www.tigerplace.net/



Zhongna Zhou received the B.S. and M.S. degrees from the University of Science and Technology of China in 2003 and 2007, respectively. She is currently pursuing the Ph.D. degree at the Department of Electrical and Computer Engineering, University of Missouri, Columbia.

Her research interests include pattern recognition and human activity analysis.

Xi Chen received the B.S. degree in electrical engineering from Harbin Institute of Technology (HIT), Harbin, China, in 1998, the M.E. degree from Aerospace Institution of China, China, in 2001, and the M.S. degree in electrical and computer engineering from the New Jersey institute of Technology, Newark, in 2004. He is currently working towards the Ph.D. degree in the Department of Electrical and Computer Engineering, University of Missouri, Columbia.

His main research interests include video compression/coding, video surveillance and intelligent video processing. Currently, he is working on the optimized power-rate-distortion model for video encoder.

Yu-Chia Chung received the B.S. degree in electrical engineering from Yuan-Ze University, Chung-Li, Taiwan, in 2001. He is currently working toward the Ph.D. degree in electrical engineering at University of Missouri-Columbia (UMC).

His research interests include video coding and networking, multimedia compression, computer vision and computational intelligence. Currently, he is working with the Video Processing and Communication Laboratory and Center for Geospatial Intelligence at UMC and active in interdisciplinary research of video processing and computational intelligence.

Zhihai He (S'98–M'01–SM'06) received the B.S. degree from Beijing Normal University, Beijing, China, and the M.S. degree from Institute of Computational Mathematics, Chinese Academy of Sciences, Beijing, China, in 1994 and 1997, respectively, both in mathematics, and the Ph.D. degree in electrical engineering from University of California, Santa Barbara, in 2001.

In 2001, he joined Sarnoff Corporation, Princeton, NJ, as a Member of Technical Staff. In 2003, he joined the Department of Electrical and Computer Engineering, University of Missouri, Columbia, as an Assistant Professor. His current research interests include image/video processing and compression, network transmission, wireless communication, computer vision analysis, sensor networks, and embedded system design.

Dr. He received the 2002 IEEE Transactions on Circuits and Systems for Video Technology Best Paper Award and the SPIE VCIP Young Investigator Award in 2004. Currently, he serves as an Associate Editor for IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY (IEEE TCSVT) and *Journal of Visual Communication and Image Representation*. He is also a guest-editor for IEEE TCSVT Special Issue on Video Surveillance. He is a member of the Visual Signal Processing and Communication Technical Committee of the IEEE Circuits and Systems Society, and serves as Technical Program Committee member or session chair of a number of international conferences.

Tony X. Han received the B.S. degree with honors in Electrical Engineering Department and Special Gifted Class from Jiaotong University, Beijing, China, in 1998, the M.S. degree in electrical and computer engineering from the University of Rhode Island, Providence, in 2002, and the Ph.D. degree in electrical and computer engineering from the University of Illinois at Urbana-Champaign, in 2007.

He then joined the Department of Electrical and Computer Engineering at the University of Missouri, Columbia, in August 2007. Currently, he is an Assistant Professor of the Department of Electrical and Computer Engineering.

Dr. Han received the CSE Fellowship from the University of Illinois in 2005.

James M. Keller (F'00) received the Ph.D. degree in mathematics in 1978.

He holds the University of Missouri Curators' Professorship in the Electrical and Computer Engineering and Computer Science Departments on the Columbia campus. He is also the R. L. Tatum Professor in the College of Engineering. His research interests center on computational intelligence: fuzzy set theory and fuzzy logic, neural networks, and evolutionary computation with a focus on problems in computer vision, pattern recognition, and information fusion including bioinformatics, spatial reasoning in robotics, geospatial intelligence, sensor and information analysis in technology for eldercare, and landmine detection. His industrial and government funding sources include the Electronics and Space Corporation, Union Electric, Geo-Centers, National Science Foundation, the Administration on Aging, NASA/JSC, the Air Force Office of Scientific Research, the Army Research Office, the Office of Naval Research, the National Geospatial Intelligence Agency, and the Army Night Vision and Electronic Sensors Directorate. He has coauthored over 250 technical publications.

Prof. Keller is a Fellow of the IEEE for whom he has presented live and video tutorials on fuzzy logic in computer vision, is an International Fuzzy Systems Association (IFSA) Fellow, a national lecturer for the Association for Computing Machinery (ACM), an IEEE Computational Intelligence Society Distinguished Lecturer, and a past President of the North American Fuzzy Information Processing Society (NAFIPS). He received the 2007 Fuzzy Systems Pioneer Award from the IEEE Computational Intelligence Society. He finished a full six-year term as Editor-in-Chief of the IEEE TRANSACTIONS ON FUZZY SYSTEMS, is an Associate Editor of the *International Journal of Approximate Reasoning*, and is on the editorial board of *Pattern Analysis and Applications*, *Fuzzy Sets and Systems*, *International Journal of Fuzzy Systems*, and the *Journal of Intelligent and Fuzzy Systems*. He is currently the Vice President for Publications of the IEEE Computational Intelligence Society. He was the conference chair of the 1991 NAFIPS Workshop, program co-chair of the 1996 NAFIPS meeting, program co-chair of the 1997 IEEE International Conference on Neural Networks, and the program chair of the 1998 IEEE International Conference on Fuzzy Systems. He was the general chair for the 2003 IEEE International Conference on Fuzzy Systems.