Multimodal Detection of Human Interaction Events in a Nursing Home Environment

Datong Chen, Robert Malkin, Jie Yang
School of Computer Science, Carnegie Mellon University
{datong, rgmalkin, yang+}@cs.cmu.edu

ABSTRACT
In this paper, we propose a multimodal system for detecting human activity and interaction patterns in a nursing home. Activities of groups of people are firstly treated as interaction patterns between any pair of partners and are then further broken into individual activities and behavior events using a multi-level context hierarchy graph. The graph is implemented using a dynamic Bayesian network to statistically model the multi-level concepts. We have developed a coarse-to-fine prototype system to illustrate the proposed concept. Experimental results have demonstrated the feasibility of the proposed approaches. The objective of this research is to automatically create concise and comprehensive reports of activities and behaviors of patients to support physicians and caregivers in a nursing facility.

Categories and Subject Descriptors
I.4.8 [**Scene analysis**]: motion, color, shape, tracking, stereo

General Terms
Algorithms

Keywords
Multimodal, human interaction, group activity, medical care, stochastic modeling

1. INTRODUCTION
Automatic detection of human activities is a prerequisite for many applications, such as surveillance, pervasive computing, and medical monitoring. In this research, we are interested in automatically detecting human activity and interaction events from video and audio for geriatric care applications within skilled-care facilities. In many such institutions, physicians might visit their patients for only a short period of time once a week. Assessment of a patient's progress is based mainly on staff reports. These reports may be incomplete or even biased, due to schedule shift and the fact that each staff person must care for many patients. This may result in insufficient observation for monitoring either progressive change or brief and infrequent occurrences of aberrant activity for diagnosing some diseases. For example, dementia is very common among residents in nursing facilities. An obvious characteristic of dementia is a sustained decline in cognitive function and memory. Studies indicate that elderly patients with dementia may exhibit measurable agitated behaviors that increase confusion, delusion, and other psychiatric disturbances [1][2]. Long-term observation and care become increasingly important for the elderly with dementia in nursing homes [3]. Although no widely accepted measure exists for dementia in care environments [4], quantitative measures of daily activities of these patients can be very useful for dementia assessments.

The long-term goal of this research is to create a system that can automatically extract and classify important antecedents of psychosocial and health outcomes. One such indicator is the frequency, duration and type of interactions of the patients with one another and their caregivers. Care providers may then interpret and assess changes in these behaviors through the recorded visual/audio compilation of activities and interactions of patients' daily lives. This paper describes an essential sub-system of our research which is able to automatically process surveillance video/audio signals recorded in a nursing home, extract salient features, analyze scenes, and detect important events that may contain elementary information to obtain concise but limited semantic descriptions of the signal contents. Based on these events and their associated temporal information, our system can also automatically generate summaries and comprehensive reports of patients' activities and behaviors to support the diagnoses made by physicians and care providers in the nursing facility.

Human activity, especially interaction with others, is generally considered a positive and necessary part of our daily life. Naturally, the level of an interaction a person has can depend on a wide range of factors, such as health, personal preference, and aptitude for interaction. Physical disability is not necessarily socially disabling. As we have observed from our recorded data, many of the most severely disabled patients had daily group activities.

Group activity is mutual or reciprocal action that involves two or more people and produces various characteristic visual/audio patterns. To simplify the problem, in this paper, we analyze a group activity by investigating interactions which occurred between all pairs of people in the group and focus on detecting interactions using multimodal techniques.

Figure 1. Examples of social interaction in video

ICMI'04, October 13-15, 2004 State College, Pennsylvania, USA.

Two major sensors, video cameras and microphones or microphone arrays, are often used for detecting human activities. Vision is an attractive modality to use for this task, as humans typically consider it to be the dominant sense and are more comfortable identifying visual than auditory events. Figure 1 illustrates several examples of interaction patterns in a hallway of a nursing facility. However, vision has some disadvantages. Video images can be expensive to capture, process, and store. Vision systems are usually sensitive to lighting, aspect, and sensor motion conditions. On the other hand, interactions also consist of audio patterns of greeting and conversations. Audio signals are easier to capture, process, and store than video signals, and audio systems are robust to lighting, aspect, and sensor motion. Audio events are, therefore, important compensation resources to video for distinguishing an interaction from independent activities that appear in the video scene by chance. In this research, we use both visual and auditory information for detecting human activities.

Previous methods attempt to categorize human activities according to predefined criteria. Due to the evolving nature of an interaction, huge numbers of categories must be defined corresponding to variations existing in activities among multiple people. A practical method decomposes human activities into a sequence of behaviors to obtain a flexible description. In order to consistently describe human interactions, we break human activities into semantic units and use a dynamic Bayesian network to model the temporal and semantic relationships among these units. By integrating acquired content description data, we construct a hierarchical video/audio content structures with group merging and clustering. This multi-level hierarchy consists of different entities, features and events, based on observations from 10 days of video records in a corridor of a nursing home. At the bottom level of the hierarchy, predefined entities and attributes (features) are detected and tracked using multimodal technologies. High-level events and features are further detected using a dynamic Bayesian network. We discuss experimental results that use the proposed system to detect interaction patterns from recorded video/audio channels.

2. RELATED WORK

A group activity or an interaction consists of both individual human activity and relations between multiple people. Therefore, the work presented in this paper is closely related with audio/visual events detection and human activity analysis, which have been addressed by many researchers in different areas such as multimodal interfaces, multimedia processing, pervasive computing, and computer vision.

2.1 Audio Event or Scene Change Detection

Audio event detection is usually accomplished by tracking changes in some audio feature stream. The simplest feature to use is signal power, as suggested in [28][29].

Auditory scene change detection, or segmentation, has usually been approached for the Broadcast News speech recognition task, using some information-theoretic criterion such as the KL2 metric [32] or the Bayesian Information Criterion [33]. A more general approach based on auditory self-similarity is given in [30][31].

2.2 Vision Based Location Events Detections

A vision-based system can provide location information while overcoming some of the limitations of the above-mentioned systems. Many computer vision algorithms have been developed for not only recovering 3D locations of a person, but also providing detailed appearance information of the person and his/her activities.

Koile et al [5] at MIT proposed a computer vision system to monitor the indoor location of a person and his/her moving trajectory. The living laboratory [6] was designed by Kidd, et. al. for monitoring the actions and activities of the elderly. Aggarwal, et. al. [7] has reviewed different methods for human motion tracking and recognition. Various schemes, single or multiple camera schemes, and 2D and 3D approaches have been broadly discussed in this review.

2.3 Activity Event Detection and Analysis

Previous human activity detection research focused on analyzing individual human behaviors and actions. Apart from the work introduced in the last paragraph, paper [10] proposed a system that combines sound and vision to track multiple people. Tan et al [12] fuse static and dynamic body biometrics for gait recognition. Clarkson and Pentland [28][29] describe a system which uses ambulatory audio and low-resolution video, fused into a single feature stream, to detect and recognize the activities of a single human with wearable sensors. This system clusters the fused signal by building Hidden Markov Models in an unsupervised fashion. This system is able to identify a variety of environmental conditions and events.

Badler [20] also proposed a hierarchical framework based on a set of motion verbs. A motion verb is actually a human behavior, which is modeled using state machines on the basis of rules predefined on static images. The system can be extended theoretically for resolving complex events existing in human activities. However, the system was only tested in an artificial environment. Other rule-based methods [8] have also shown their merits in action analysis. Rule-based systems may have difficulties in defining precise rules for every behavior because some behaviors may consist of fuzzy concepts.

Statistical approaches, from template models, linear models, to graphic models, have been used in human activity analysis. Chomat and Crowley proposed a probabilistic method for recognizing activities from local spatio-temporal appearance [15]. Yacoob and Black [14] used linear models to track cyclic human motion. The model consists of the eigen vectors extracted using principal component analysis from the observations. So far, this methodology is limited to modeling different repeated patterns of human motion.

Various graphical models have been used for modeling human behaviors. Intille and Bobick [11] interpret actions (agents) using Bayesian networks among multiple agents. Bayesian networks can combine uncertain temporal information and compute the likelihood for the trajectory of a set of objects to be a multi-agent action. Dynamic mechanisms existing among group actions were omitted in this work. Jebara and Pentland [9] employed conditional Expectation Maximization to model and predict actions. Their system could synthesize a reaction based on the predicted action. Hidden Markov models [17], layered hidden Markov models [19][22], or other variation of Markov model [18] have been used for recognizing actions and activities, and illustrated their advantages in modeling temporal relationships between visual-audio events. However, large amounts of training

data are usually required to obtain good models of various actions in the spatiotemporal domain [16]. Ivanov [13] proposed a stochastic, context-free grammar to interpret an activity by recursively searching for a complete tree in a non-deterministic probabilistic expansion of context-free grammar. Similar to Kojima's work, this graphical model can generate a natural description of activities based on the detected events. Although this model has great potential advantages to be extended for analyzing interactions, no published work has been found so far.

3. HIERARCHICAL REPRESENTATION OF ACTIVITY AND INTERACTION EVENTS

In this research, we propose to use context hierarchies to characterize interesting events in a nursing home. Human activities and interactions events usually contain certain relationships to each other. An interaction event may consist of many individual activities. To represent this context, we developed a four level hierarchy by observing video/audio records of a hallway in a nursing home for 10 days. Each record was captured at a resolution of 640 x 480 and stored in mpeg-2 format (30 frames/second) and two audio channels. After viewing 80 hours of video (8 hours for each day), we have defined a four-level context hierarchy for representing daily activities of patients, staff, and visitors. From bottom to top, the four levels are conceptual element (*CE*), individual person activity event (*IE*), group activity feature (*GF*), and group event (GE), which are illustrated in Figure 2.

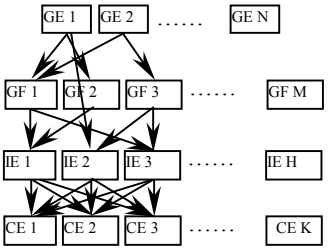


Figure 2. Context hierarchy of nursing home hierarchy

The conceptual elements consist of entities that are objects of interest to us, and some attributes of entities. The entities of a nursing home are walking or standing human beings and patients using wheelchairs. The attributes are features for measuring motions and visual appearances of an entity. We use five visual features: location, moving direction, speed, color, and shape, as explained in Table 1. We will discuss the detail implementation of entity detection and feature extraction in section 4.

Table 1. Attributes of entities in a nursing home

Attributes	Definition
Location (E)	Describing the physical location of the entity "E".
Moving direction	Describing the moving direction of the entity "E".
Speed (E)	Describing the moving speed of the entity "E".
Color (E)	The entity "E" has skin color.
Shape (E)	Shape information of the entity "E"

An individual person activity event (IE) is defined as a combination of a person entity and a sequence of attributes. For example, the IE "Walking (*A*)" indicates person *A* with a sequence of changing locations. Table 2 has listed some IEs in a hallway of a nursing home. Other IEs for different locations in a nursing home, such as dining room, can be defined using different knowledge sources.

Group activity features (GFs) are combinations of IEs that involve two individual person entities as listed in Table 3. GFs are features of relative motions of two IEs. These features that measure relative distance or walking directions between two people, for example, the "distance (*A, B*)" measures the distance between person *A* and person *B*.

Table 2. Individual human activity events (IEs).

Individual people activity events	Definition
Walking(person)	A person is walking.
Sitting(person)	A person is sitting.
Standing (person)	A person is standing.

Table 3. List of group activity features and events (GEs)

Group activity features	Definition
Distance (A, B)	Distance between A and B.
Relative direction (A, B)	Relative moving direction of A & B.
Relative speed (A, B)	A and B are walking together.

A group interaction event (GE) is a segment of a story (a meaningful sequence of video/audio) of human activities consisting of a group of individual activity events and group activity features. For example, a story of a typical conversation in the hallway can be mainly partitioned into three segments:

1. Person A and person B approach to each other;
2. A and B are talking.
3. A and B walk out of the hallway together or separately.

Theoretically, if the observation time and the number of people involved are not limited, the number of possible interactions can be quite large. In this paper we only interested with five events as listed in [WHERE?]

Table 4 Group interaction events

Group interaction events features	Definition
Approaching (person A, person B)	A and B is approaching to each other.
Leaving (person A, person B)	A is leaving B.
Close (person A, person B)	A and B are very close to each other.
Standing conversation	More than one people are standing and talking..
Walking assistance	People are walking together.

4. IMPLEMENTATION OF THE INTERACTION DETECTION SYSTEM

The system is proposed to detect interaction patterns from coarse to fine, which consists of two steps: coarse interaction event detection and fine interaction event detection as shown in Figure 3.

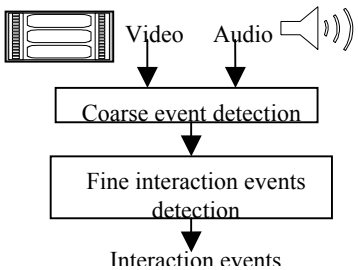


Figure 3. The Architecture of the interaction event detection system

4.1 Coarse Interaction Event Detection

We detect candidate interactions among a group of people by fusing audio and video channels. A camera network installed in the nursing home records both video and audio. To speed-up the process, we first quickly extract video and audio shots that may contain interactions.

4.1.1 Video events detection

For the video channel, we use a background subtraction algorithm to detect frames that contain human activities. To speed up this detection process, only video from one camera in the network is used. The background of a frame is obtained by the adaptive background method [21]. We employ a threshold to extract pixels that have high differences between the current frame and its background. To remove noise, we group extracted pixels into regions and only keep those regions that contain more than 15 pixels. We consider the frame *f* to contain a visual interaction event $V_f=1$ if any of the following rules is satisfied; otherwise $V_f=0$:

1. There are two or more regions in the frame.
2. There is region that does not touch the bottom the frame, whose width to height ratio is more than 0.7.

We choose these thresholds to detect as many interactions as possible without inducing excess false alarms.

The output of the detection is reported every second. For a second of NTSC video, we output the percentage of visual cues in its 30 frames as:

$$C_v = \frac{1}{30}\sum_{f=1}^{30} v_f$$

4.1.2 Audio event detection

To detect events using the audio stream, we use a very simple power-based method like the one proposed by Clarkson and Pentland in [28][29]. This method adaptively normalizes signal power to zero mean and unity variance using a finite-length window; segments where the normalized power exceeds some threshold are designated "events." [28] and [29] describe an ambulatory system which could be exposed to arbitrary acoustic environments; adaptive normalization allows such a system to compensate for unusually loud or quiet environments and still detect events reliably. Our task differs from this one in that we have a *stationary* system where changes in power level really do indicate events and not just changes of venue. As such, instead of adaptive normalization, we use global normalization. That is, a single mean and variance is calculated for each two-hour recording and the globally-normalized power is threshold to detect events a_f.

In this implementation, we extracted 16kHz, 16-bit mono audio from the audio-video stream, and used analysis windows 200ms in length with a 50% overlap. This window length results in a frame rate of 10 frames per second, which is more than adequate to detect events using the power-based approach. After signal power is calculated and normalized, it is passed through a simple 3-frame averaging filter for smoothing. We then apply the power threshold; any segment which exceeds the threshold is designated an event. We also stipulate a minimum event time of 1 second in order to filter out isolated auditory transients. The confidence of audio event per second is defined as:

$$C_a = \frac{1}{10}\sum_{f=1}^{10} a_f$$

4.1.3 Fusing video and audio events detection

We linearly combine the video event confidence and audio event confidence together for final event detection:

$$C_d = \alpha C_v + (1-\alpha)C_a$$

We consider a one-second frame to contain an interaction if its confidence C_d is higher than 0.5.

4.2 Fine Interaction Event Detection

The fine detection of interactions is based on the multi-level context hierarchy we proposed in section 3. The hierarchy of interactions in a nursing home is mapped onto a dynamic Bayesian network (DBN), which represents not only the interactions and the event hierarchy by its states and arcs, but also the evolution of the interactions over time by temporal arcs defined between the interactions. To illustrate the concept, a part of temporal structure of the DBN is depicted in Figure 4. Formally, the DBN *B=(S, M)* is a directed acyclic graph that consists of a state set $S = GE \cup GF \cup IE = \{s_1,...,s_n\}$, which represents events and interactions, a set of directed arches that specifies parents of each state *s*: *Parents(s)*, and a parameter set *M*, which includes the probabilities for any input video sequence $O = (o^1,...,o^k)$: the event data likelihoods $P_M(o^t \mid s_i)$ and the hierarchy of relationships $P_M(s_i \mid Parent(s_i))$. The joint distribution of the DBN is defined as:

$$P(s_1,...,s_n) = \prod_i P(s_i \mid Parent(s_i)) \qquad (1)$$

The graph is built by defining the parents of each state according to the relationships defined in the hierarchy. The temporal arcs are also added into the graph using daily knowledge. In the rest of this section, we will discuss the implementation of the parameter *M* in detail. This template graph can be pre-trained offline. For analyzing a video/audio record, we dynamically build a graph the same as the template graph for each pair of entities that extracted from the scene. After obtaining the data likelihoods for the GEs in all the graphs, we consider the GE that has the highest likelihood to be the current analysis result.

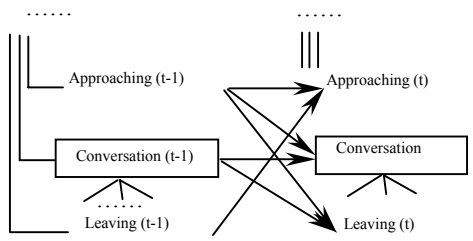


Figure 4. An illustration of a partial DBN with temporal arcs

4.2.1 Entity Detection and Attributes Extraction

We manually labeled the position of all the doors and entrances of the hallway. An entity that appears close to one of these doors and entrances for the first time is initialized and tracked in the hallway. We consider a region extracted in the pre-segmentation step as an entity if it contains skin color pixels in the top 30% of the whole region. The skin color is modeled as a Gaussian mixture [23]. The location and moving direction features can be extracted directly from the tracking results. The appearance features, color and shape, are extracted from key-frames.

4.2.1.1 3D tracking

Since occlusions happen very often in the narrow hallway, we use a particle filtering base multiple camera framework to track human movement. This framework uses one or more cameras to cover the target area. The location of a person in 3D space is obtained by integrating tracking confidence in the images from the cameras. Instead of using a traditional stereo algorithm, this 3D location recovery task uses a new tracking algorithm, which can robustly compensate tracking cues from different numbers of cameras.

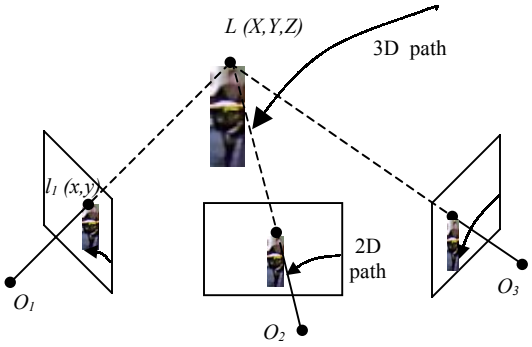


Figure 5 3D tracking with a camera network

A camera network consists multiple cameras covering the interesting areas in the nursing home as illustrated in Figure 5. A simple pin-hole model is used for all the cameras. We can calibrate the cameras off-line because we don't move them once they are calibrated. After calibrating the intrinsic and extrinsic parameters, we can map a spatial point *L(X,Y,Z)* in 3D world coordinates to its corresponding point $l_i(x,y)$ in the image plane of each camera *i* which can be defined by the following equation:

$$\begin{pmatrix} x \\ y \\ f_i \end{pmatrix} = \frac{f_i}{Z} \begin{pmatrix} X \\ Y \\ Z \end{pmatrix}$$

where f_i is the focus length of the camera *i*.

The spatial points can be silhouettes. We use both the head (highest point) and feet (lowest point) in this research. Using particle filters, we are able to track a silhouette in 3D world coordinates using the tracked features from all the cameras.

The idea of particle filters was first developed in the statistical literature, and recently this methodology, namely sequential Monte Carlo filtering [21][25] or CONDENSATION, has shown to be a successful approach in several applications of computer vision [26][27]. A particle filter is a particle approximation of a Bayes filter, which addresses the problem of estimating the posterior probability $p(L_t|O_{1:t})$ of a dynamic state given a sequence of observations, where L_t denotes the state *L* (3D position in the world coordination) at time t and $O_{1:t}$ denote the observed images sequence from all the cameras from time 1 to time t. Assuming independence of observations conditioned on the states and a first order Markov model for the sequence of states, we obtain the following recursive equation for the posterior:

$$p(L_t|O_{1:t}) = \alpha p(O_t|L_t) \int_{L_{t-1}} p(L_t|L_{t-1})\, p(L_{t-1}|O_{1:t-1}) dL_{t-1} \quad (2)$$

where α is a normalization constant and the transition probability $p(L_t \mid L_{t-1})$ is assumed to be a Gaussian distribution. The data likelihood is obtained by first mapping the 3D position *L(X, Y, Z)* of a silhouette to the current images from cameras and then computing the average tracking confidences $C(l_i)$ at these 2D positions l_i:

$$p(O \mid L) = \frac{1}{N} \sum_{i=1}^{N} \frac{C(l_i)}{|L_i|}, \quad C(l_i) > C \qquad (3)$$

Here, $|L_i|$ is the distance from the optical center of the camera *i* to the point *L*. The threshold *C* is a constant for removing tracking errors. If a mapped 2D point is out of the image, the corresponding tracking confidence is set to 0. *N* is the number of cameras that contain tracking results with high enough confidences.

In practice, a head silhouette has less chance to be occluded than a foot silhouette. However, the 3D location of a head silhouette can only be recovered if it is tracked in the frames from at least two cameras. Therefore, for tracking a head silhouette, *N* must be greater than 1. One the other hand, although a foot silhouette is often occluded, it can indicate the 3D location of a person using only one camera. This is very important in the case that a person is only visible in only one camera.

Following the idea of a particle filter, the posterior $p(L_t|O_{1:t})$ is approximated by a set of weighted samples of locations *L*. The weight of a location is defined as its data likelihood. The initial weighted sample set contains only one state L_0, which is obtained by performing a full search around the 3D position near the entrance where the person is initialized. Then, for each frame 100

new samples are generated and their confidences are computed. To keep the size of the weighted sample set, among these 100 new samples, the first 50 samples with the highest confidences are then treated as the new weighted sample set for the next frame. The final current tracked position is set to be the value of the sample (3D location) with the highest confidence.

One advantage of this tracking framework is that it can reduce tracking errors with multiple cameras. Figure 6 illustrates the compensation of tracking results of two persons using this multiple cameras framework in simulation sequences. The results of tracking using individual cameras and the proposed multiple cameras framework is shown on a time axis. A vertical bar at time *t* indicates that the person is tracked at time *t*, otherwise the person is not tracked. We can see that the proposed method obtained no blank (loss of tracking) here. Tracking results from the 10 minute long sequences are shown in Figure 7. The proposed tracking framework reduces tracking errors by 58% on average, which can significantly prevent tracking errors from occlusions.

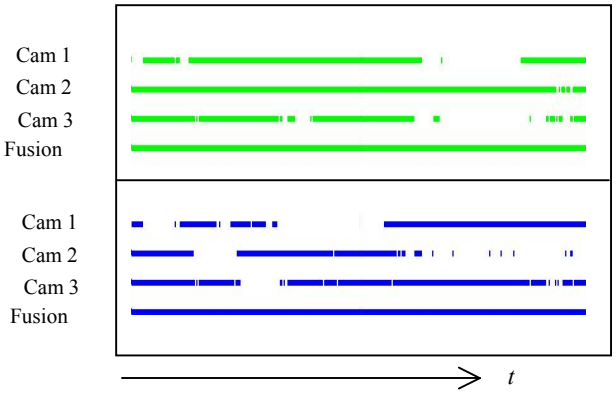


Figure 6. An illustration of tracking results using the proposed framework. A color mark at time t indicates that the person is tracked at time t by the corresponding camera or combination of cameras, otherwise the person is lost to tracking.

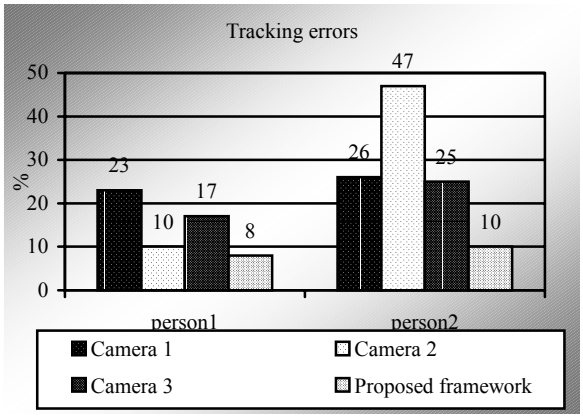


Figure 7. Tracking errors in 10 minute simulation video

4.2.1.2 Color and shape features

Color features are mostly used to distinguish different entities in the tracking process. We use 8-bin histograms in RGB color space as features for each entity.

Shape information is represented by partitions with Manhattan distances. In this method, each extracted region that contains people or facilities is divided into 9 sub-regions, as shown in the figure 8. The density of each sub-region is calculated and threshold to equal '1' if it is greater than 50% and '0' otherwise. Finally, a shape feature vector of a region is a 10 dimensional vector: 9 city block features and the width/height ratio of the region.

All the attributes (features) are extracted every second. The "location" is represented by (*X, Y*) of the tracked 3D spatial point *L(X, Y, Z)* at the beginning of each second. Speed and moving direction are computed every second. Color and shape features are also extracted from the first frame of each second. Therefore, the input of the event detection level is uniform attribute (feature) vectors per second.

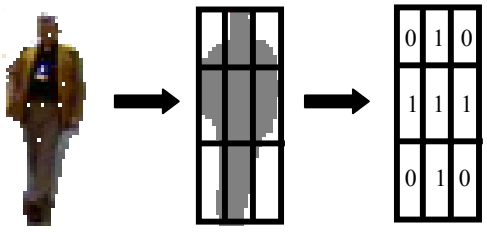


Figure 8 Shape feature.

4.2.2 IE Detection

Each IE is modeled individually using Gaussian mixture models (GMMs). We train each IE separately using the standard EM algorithm [24]. In order to train good models using limited training data, we perform feature selection using χ^2 for each event for reducing the feature space. The parameters of GMMs are optimized using 2-fold cross validation.

4.2.2.1 GF extraction and GE detection

After detecting IEs, we can build graphs for each pair of the IEs. Group activity features are also extracted for each pair of entities based on the extracted features. The data likelihood of GEs in each graph can then be computed using equation (1). For every second, we output the GE which has the highest likelihood as the result.

5. EXPERIMENTAL RESULTS

To evaluate the coarse event detection, we labeled 10 hours of video/audio records. Using only video detection, we extract 33.3% of the whole video as candidate interaction shots, which is listed in Table 5. In order to not miss any interactions, we only filter out the one-second-long video segments with zero confidence.

Table 5 Total event time from video

	Total Event Time (second)	**Event Time as % of Total Signal**
No activity	13711	38.1%
Individual	6700	18.6%
Interaction events	15589	33.3%

Using only audio detection with varying thresholds, we obtain the results listed in Table 6. The table shows the total event time and percentage of the recordings for three thresholds.

Table 6 Total event time from audio per threshold.

Threshold	Total Event Time (second)	Event Time as % of Total Signal
1.1	6705	18.6%
1.6	5582	15.5%
2.1	4327	12.0%

By fusing the audio (threshold 1.6) and video results, we extracted in sum 9435 seconds from the whole 10 hour record. In this way, 85 our of 91 interactions in the ground truth are covered by the candidate shots, which obtain reasonable recall and precision in terms of event time as listed in Table 7. The audio has a lower recall due to the presence of silent interactions such as walking assistance of a wheelchair-bound patient. The audio precision is actually higher in general than is reported here. The hallway environment is a poor representative of audio precision, as many events that are audible in the hallway are off-camera and not in the ground-truth labels; thus audio event detection generates many false alarms. Even so, our results show that by fusing audio and video results, we can achieve more than 90% recall and 20% precision. We project even better precision when we test our fused system over the entire set of nursing home environments.

Table 7 Coarse detection results

	Recall	Precision	Process speed
Video	98%	13%	real time
Audio	71%	28%	10% real time
Multimodal	92%	21.0%	

More interaction sequences are needed to train and evaluate the fine interaction event detection. We have selected 160 short video sequences of interactions from 80 hours of hallway video at a nursing home (8 hours each day for 10 days). The average length of these video sequences consists of 400 frames. To avoid interpreting very complex activities, most of the sequences contain interactions only involving two persons. We manually labeled the ground truth of these video sequences.

Figures 9-10 illustrate speed features and "distance" of four typical video sequences in our database. Each of the four videos contains an interaction of two persons. Video (1) shows person *A* meeting person *B* coming from the entrance located at another side of the hallway. They hug each other and then stand and talk to each other for a while. Finally, person *B* accompanies person *A,* walking towards the entrance. Using the concepts defined in our hierarchy, video (1) can be simply interpreted as: "approaching, standing conversation and walking assistance". Concisely, we can interpret the video (2-4) as: (2) standing conversation and walking assistance; (3) approaching, standing conversation and leaving; (4) approaching, close, and leaving. In Figure 9 and 10, different scales are used for the *Y* axis in order to show the results in as much detail as possible. We can observe there are some errors in the figures in that the "speed" and "distance" of video (2) are shifted to a different scale. The errors are caused by precision of the tracking algorithm and the calibration of the camera network. Fortunately, the errors can be controlled within a small range.

We use 80 videos in the database as the training set and use the remaining 80 videos as the test set. Table 8 lists the number of interactions in the training set and test set. Only 2 interactions are listed here because approaching, close and leaving are events that related with interactions.

6. CONCLUSIONS

This paper described a system for detecting human interaction events in a nursing home using multimodal technology. Human interaction is one of the most complex human activities in a nursing home and can provide potentially important information regarding long-term care patients. We have demonstrated that the proposed hierarchical system can automatically detect interaction events of groups of people in the hallway of a nursing home. The experiments show that fusing video and audio signals can improve the coarse events detection compared to using only video or audio channels. Future efforts will involve video/audio fusion in fine detection and identification of human interactions as well as coarse event detection.

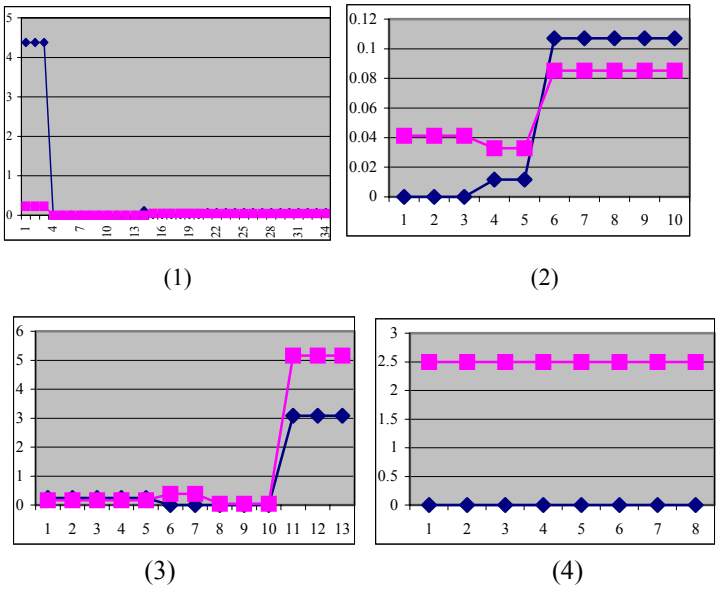
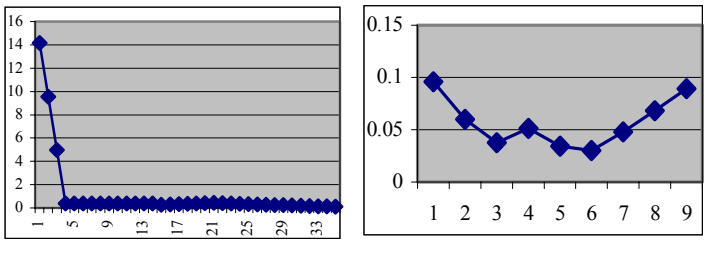
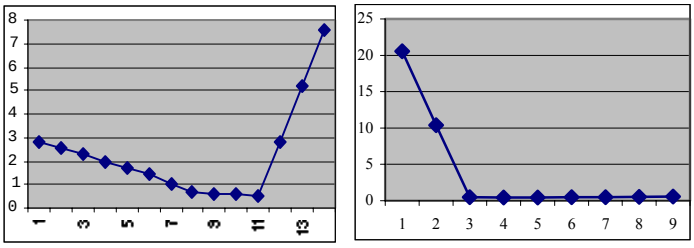


Figure 9. Speed features of some video samples: (1) approaching, stand conversation, walking assistance; (see original key frames in the 1st row of Fig. 1) (2) stand conversation, walking assistance; (see Fig. 1, 2nd row, left) (3) approaching, stand conversation, leaving; (see Fig. 1, 2nd row, middle) (4) approaching and leaving (see Fig. 1, 2nd row, right).





(3) (4)

Figure 10. "distance" extracted from four video sequences described in in Figure 9.

Table 8. Activity detection results

Interactions	Training set	Test set	Recall	False alarms
No interaction	21	15	93%	4
Stand conversation	32	34	88%	9
Walking assistance	40	44	86%	6

7. ACKNOWLEDGEMENT

This research is partially supported by the National Science Foundation (USA) through project CareMedia No. 0205219, and the European Commission within the project CHIL (http://chil.server.de) under contract No. 506909.

8. REFERENCES

[1] J. Nelson, "The influence of environmental factors in incidents of disruptive behavior", Journal of Gerontological Nursing 21(5):19-24, 1995.

[2] P. D. Sloane, C. M. Mitchell, K. Long and M. Lynn, "TESS 2+ Instrument B: Unit observation checklist – physical environment: A report on the psychometric properties of individual items, and initial recommendations on scaling", University of North Carolina 1995.

[3] F. J. Eppig and J. A. Poisal, "Mental health of medicare beneficiaries: 1995", Health Care Financing Review, 15, pages: 207-210, 1995.

[4] F. Carp, "Assessing the environment", Annul review of gerhierarchy and geriatrics, 14, pages: 302-314, 1994.

[5] Kimberle Koile, Konrad Tollmar, David Demirdjian, Howard E. Shrobe, Trevor Darrell, "Activity Zones for Context-Aware Computing", Ubicomp 2003, pp. 90-106, 2003.

[6] Cory D. Kidd, Robert Orr, Gregory D. Abowd, Christopher G. Atkeson, Irfan A. Essa, Blair MacIntyre, Elizabeth Mynatt, and Thad E. Starner and Wendy Newstetter, "The Aware Home: A Living Laboratory for Ubiquitous Computing Research". Proc. of CoBuild '99, pp.191-198, 1999.

[7] J. K. Aggarwal, Q. Cai, "Human Motion Analysis: A Review," Computer Vision and Image Understanding, Vol. 73, pp. 428-440, 1999.

[8] D. Ayers, M. Shah, "Monitoring Human Behavior from Video Taken in an Office Environment," Image and Vision Computing, Vol. 19, pp. 833-846, 2001.

[9] T. Jebara, A. Pentland, "Action Reaction Learning: Analysis and Synthesis of Human Behavior," IEEE Workshop on the Interpretation of Visual Motion, 1998.

[10] Neal Checka, Kevin Wilson, Michael Siracusa, Trevor Darrell, "Multiple Person and Speaker Activity Tracking with a Particle Filter," ICASSP, 2004

[11] S. Intille and A. Bobick, "Recognizing planned, multi-person action", Computer Vision and Image Understanding, vol. 81, no. 3, pp. 414-445, March 2001

[12] L. Wang; H. Ning; T. Tan and W. Hu, "Fusion of static and dynamic body biometrics for gait recognition," IEEE Trans. Circuits and Systems for Video Technology, 14 (2), pp. 149 – 158, 2004.

[13] Y. A. Ivanov, A. F. Bobick, "Recognition of Visual Activities and Interactions by Stochastic Parsing," IEEE Trans. PAMI, Vol. 22, pp. 852-872, 2000.

[14] Y. Yacoob, M. J. Black, "Parameterized Modeling and Recognition of Activities," ICCV, pp. 232-247, 1998.

[15] O. Chomat and J.L. Crowley, "Probabilistic recognition of activity using local appearance", in ICVPR, pp. 104-109 June 1999.

[16] A. D. Wilson, A. F. Bobick, "Realtime Online Adaptive Gesture Recognition," International Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems, pp. 111-117, 1999.

[17] D. J. Moore, "I. A. Essa, M. H. Hayes, "Exploiting Human Actions and Object Context for Recognition Tasks," Proc. of ICCV, Vol. 1, pp. 80-86, 1999.

[18] Milind Napahade, Ashutosh Garg and T. S. Huang, "Duration Dependent Input Output Markov Models for Audio-Visual Event Detection," ICME, 2001.

[19] N. Oliver, A. Garg, E. Horvitz, "Layered Representation for Learning and Inferring Office Activity from Multiple Sensory Channels," Fourth IEEE Conference on Multimodal Interfaces, pp. 3-8, 2002.

[20] N. Badler, "Temporal Scene Analysis: Conceptual Description of Object Movements," University of Toronto TR No. 80, 1975.

[21] C. Stauffer and W.E.L Grimson. "Adaptive background mixture models for real-time tracking", Proc. of CVPR,, 1999.

[22] I. McCowan, D. Gatica-Perez, S. Bengio, G. Lathoud, M. Barnard and Dong Zhang, "Automatic Analysis of Multi-modal Group Actions in Meetings," IEEE Trans. on PAMI, 2004.

[23] J. Yang, W. Lu, and A. Waibel, "Skin-color modeling and adaptation". In Proc. of ACCV, vol. II, pp. 687-694, 1998.

[24] H. Hartley, "Maximum likelihood estimation from incomplete data". Bio-metrics, 14:174–194, 1958.

[25] A. Doucet, N. de Freitas, and N. Gordon. Sequential Monte Carlo Methods in Practice. Springer-Verlag, 2001.

[26] K. Nummiaro, E. Koller-Meier, and L. Van Gool. Object tracking with an adaptive color-based particle filter. In Proc. Symposium for Pattern Recognition of the DAGM, Sep. 2000.

[27] P. Perez, A. Blake, and M. Gangnet. Jetstream: Probabilistic contour extraction with particles. Proc. of ICCV, pages 424–531, Vancouver, July 2001.

[28] B. Clarkson and A. Pentland. Framing Through Peripheral Perception. Proc. of ICIP, Vancouver, September 2000.

[29] B. Clarkson and A. Pentland. Unsupervised Clustering of Ambulatory Audio and Video. Proc. of the ICASSP, Phoenix, 1998.

[30] J. Foote. Visualizing Music and Audio using Self-Similarity. Proc. of ACM Multimedia, Orlando, October 1999.

[31] J. Foote. Automatic Audio Segmentation using a Measure of Audio Novelty, Proc. ICME, July 2000.

[32] M. Siegler, U. Jain, B. Raj, R. Stern. Automatic Segmentation, Classification, and Clustering of Broadcast News Audio, Proc. of the 9th DARPA Spoken Language Systems Technology Workshop, New York, 1997.

[33] S. Chen and P.S. Gopalakrishnan. Speaker, Environment, and Channel Change Detection and Clustering via the Bayesian Information Criterion, Proc. of the DARPA Broadcast News Transcription and Understanding Workshop, Virginia, 1998.