6th ACM SIGMM Int'l Workshop on Multimedia Information Retrieval, in Proc. of ACM Multimedia 2004, New York, NY, pp. 283-290, October 10-16, 2004

Towards Automatic Analysis of Social Interaction Patterns in a Nursing Home Environment from Video

Datong Chen, Jie Yang, Howard D. Wactlar

Informedia
School of Computer Science
Carnegie Mellon University
{datong, yang+, hdw}@cs.cmu.edu

ABSTRACT

In this paper, we propose an ontology-based approach for analyzing social interaction patterns in a nursing home from video. Social interaction patterns are broken into individual activities and behavior events using a multi-level context hierarchy ontology framework. To take advantage of an ontology in representing how social interactions evolve, we design and refine the ontology based on knowledge gained from 80 hours of video recorded in the public spaces of a nursing home. The ontology is implemented using a dynamic Bayesian network to statistically model the multi-level concepts defined in the ontology. We have developed a prototype system to illustrate the proposed concept. Experiment results have demonstrated feasibility of the proposed approach. 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

H.3 [**Content Analysis**] & I.4.8 [**Scene analysis**]: motion, color, shape, tracking, stereo

General Terms

Algorithms

Keywords

Ontology, social interaction, human activity, medical care, stochastic modeling

1. INTRODUCTION

The explosive growth of information sources has created new challenges for the multimedia community. Besides broadcast news, we have to process multimedia information from many other sources, ranging from sporting events to video surveillance. In this research, we are interested in automatically extracting information 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. The reports may be incomplete or even biased, due to schedule shift and the fact that each staff person has to take care of 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 [16]. Studies indicate that the elderly with dementia may exhibit measurable agitated behaviors that increase confusion, delusion, and other psychiatric disturbances [23][30]. Long-term observation and care become increasingly important for the elderly with dementia in nursing homes [9]. Although no widely accepted measure exists for dementia care environments [5], quantitative measures of daily activities of these patients can be very useful for dementia assessments.

The 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 social interactions of the patients with one another and their caregivers. Caregivers may then interpret and assess changes in these behaviors through the recorded visual/audio compilation of social interactions of their daily lives. The system is able to automatically process surveillance video signals recorded in a nursing home, extracting salient features, analyzing scenes, and finally obtaining a concise but limited semantic description of these videos. The system can also automatically generate summaries and comprehensive reports of activities and behaviors of patients to support the diagnoses of physicians and caregivers in a nursing facility. Interaction with others is generally considered a positive and necessary part of our daily life. Naturally, the level of social interaction a person has can depend on a wide range of factors, such as his/her health, his/her personal preference, and aptitude for social 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 social interactions. In this paper, we focus on analysis of social interaction patterns based on information from individuals' activities.

Figure 1. Examples of interaction patterns in a nursing home

Interaction is mutual or reciprocal action that involves two or more people. Social interaction produces many visual patterns. These patterns are related to many different parameters, such as presence (how many), identity (who), relationship (who to whom) and environment (where). Figure 1 illustrates several examples of social interaction patterns in a hallway of a nursing facility. Many social interaction patterns in the hallway are associated with motion parameters, e.g., a group of people standing, a group of people moving, a group of people merging from different directions, people changing their motion parameters (directions and speed), etc. Therefore, we can analyze social interaction patterns through these motion parameters.

Previous taxonomy-based methods attempt to categorize human activities according to predefined criteria. Due to the evolving nature of a social interaction, huge numbers of categories have to be defined corresponding to variations existing in activities among multiple people. Some methods decompose human activities into a sequence of behaviors to obtain a flexible description. Since no semantically clear conceptual framework has been established amongst the behaviors, many redundant behaviors have to be defined, even though the defined scenarios may not be of interest. In order to consistently describe human social interaction, we propose an ontology-based approach in this paper. To parse and interpret human social interaction, we break human activities into meaningful units under a well-organized framework. We describe human social interaction based on an ontology hierarchy that has real semantic meaning. By integrating acquired content description data, we can construct a hierarchical video content structure with group merging and clustering. This approach can, therefore, provide a more flexible way to classify features into different concepts and to interpret social interaction using concepts at different levels of detail. To demonstrate the proposed approach, we have built an ontology for different entities, events, and social interactions, based on observations from 10 days of video records in a corridor of a nursing home. The ontology consists of multiple levels of context hierarchy. At the bottom level of the ontology hierarchy, predefined entities and attributes are detected and tracked using computer vision technologies. These entities are further classified using a dynamic Bayesian network. We discuss experimental results that use the proposed approach to identify social interaction patterns from recorded video.

The rest of the paper is organized as follows: Section 2 discusses the related work in location awareness and human activity analysis based on taxonomies. Recent ontology efforts are also discussed. Section 3 presents the social ontology we designed for the nursing home. Section 4 describes the implementation of the ontology framework in technical detail. Section 5 focuses on the training and validation of the framework using experiments on video shots. Section 6 concludes with the advantages and limitations of the proposed framework.

2. RELATED WORK

A social interaction consists of both individual human activity and relations between multiple people. Therefore, the work presented in this paper is closely related with location awareness and human activity analysis, which have been addressed by many researchers in different areas such as multimedia processing, pervasive computing, and computer vision.

2.1 Non-Vision Based Location Awareness

A GPS (Global Position System)-based system can compute the location of a radar reflection using the difference in time-of-flight between 3 precisely synchronized satellites [5]. The Active Bat Location system [12] obtains location of a mobile tag using ultrasound sensors mounted on the ceiling of a room. PlusOn time modulated ultra wideband technology [33] provides location measures to centimeter precision. These indoor and outdoor localization systems provide quite precise location information but require a user to wear a special receiver or a tag, which may present operational difficulties at a nursing facility.

Power line network [5] and Ogawa's monitoring system use switches and motion detectors to track human activities indoors. In these systems, the tracking is extended from a person's body to his environment, for example, the water level in the bath. The data provided by switches and motion sensors are reliable and very easy to process. However, they cannot provide detailed information. For example, a motion sensor can only tell that there is a person in the monitored area but cannot tell the exact location.

2.2 Vision Based Location Awareness

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 [18] at MIT proposed a computer vision system to monitor the indoor location of a person and his/her moving trajectory. The living laboratory [17] was designed by Kidd, et. al. for monitoring the actions and activities of the elderly. Aggarwal, et. al. [1] 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 Individual Human Activity Analysis

Earlier human activity recognition research focused on analyzing individual human behaviors and actions. Apart from the work introduced in the last paragraph, Kojima and Tamura [16] proposed an individual human action recognition method using a case framework, which is widely used in natural language processing. Case frames are defined to be action rules organized in a hierarchical structure. Observed actions are interpreted with a sentence based on the action grammar (frames) corresponding to a sequence of motion events. Although this work does not employ statistical models, it has shown that breaking an individual's activity into action segments can provide a natural-like interpretation.

Badler [3] 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 [3][12] 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. Davis and Bobick [5] proposed a template model-based method for tracking human movement. They constructed temporal templates using motion energy and motion history. However, no evidence has been shown that these temporal templates can be efficiently associated with meaningful activities. Yacoob and Black [35] 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 graphic models have been used for modeling human behaviors. Intille and Bobick [14] interpret actions (agents) using Bayesian network among multiple agents. The Bayesian network can combine uncertain temporal information and compute the likelihood for the trajectory of a set of objects to be a multi-agent action. This work proposed that group actions could be "compiled down" into collections of visual features and temporally coordinated (individual) activities. Dynamic mechanisms existing among group actions were omitted in this work. Jebara and Pentland [16] employed a conditional expectation maximization to model and predict the actions. Their system could synthesize a reaction based on the predicted action. Hidden Markov models [22], layered hidden Markov models [26][12], or coupled hidden Markov models [27] have been used for recognizing actions and activities, and illustrated their advantages in modeling temporal relationships between visual-audio events. However, huge training data is usually required to obtain good models of various actions in the spatiotemporal domain [33]. Ivanov [12] 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 graphic model can generate a more 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.

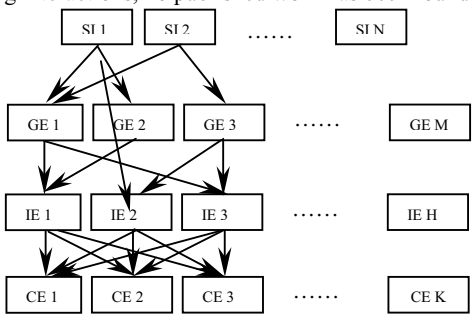

Figure 2. Context hierarchy of nursing home ontology

3. ONTOLOGY FOR REPRESENTING SOCIAL INTERACTION

In this research, we propose to use ontology to characterize social interaction in a nursing home. Ontology is the study of the categories of things that exist or may exist in some domain [30]. Using ontology, a transcription can express relationships about the entities in the domain of interest, which has potential advantages in information searching and retrieval, natural language processing, and automated inferencing [2][9][22][26][30].

The method of generating an ontology can be tracked back to Aristotle [30]. In Aristotle's method, new categories can be defined by specified properties that distinguish different species of the same parent category. Using this methodology, we developed an ontology by observing video records of a hallway in a nursing home for 10 days. The video was captured at a resolution of 640 x 480 and stored in mpeg-2 format (30 frames/second). After viewing 80 hours of video (8 hours for each day), we have come up with 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 and event (*GE*), and social interaction (*SI*), which are illustrated in Figure 2.

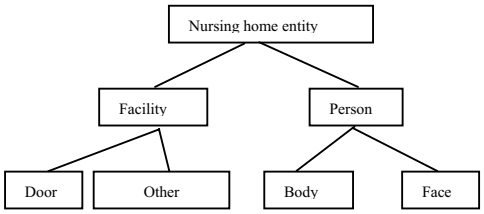

Figure 3. Entities of ontology in a nursing home

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 defined by the ontology in Figure 3. The functions that can distinguish a facility from a person or distinguish the body, hand, and face of a person can be implemented by building corresponding detectors. We will discuss the detail implementation 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 (E)	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.
Front face (person)	Front face has been detected for the person.
Shape (E)	Shape information of the entity "E"

Attributes are very abstract and difficult to categorize. We simply put them into one category as listed in Table 1.

Table 2. Some common individual activity events (IEs) in a nursing home.

Individual people activity events	Definition
Walking(person)	A person is walking.
Sitting(person)	A person is sitting.
Standing (person)	A person is standing.
Being pushed (wheel chair)	A wheel chair is being pushed.
Door used (door)	Some entities are passing the door.
Head turning (person)	Person is turning the head.

Attributes are also extracted by detectors and trackers, which are discussed in Section 4.

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. Some interactions between a person and facility are also defined as events, such as "sitting in a wheel chair" and "door being used". 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.

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

Group activity features and events (GEs)	Definition
Distance (person A, person B)	Distance between A and B, which can be deduced to three categories: approaching, distance close, and leaving.
Relative direction (person A, person B)	Relative moving direction between A and B.
Walking together (person A, person B)	A and B are walking together.
Lower speed (person A, person B)	A reduces his/her speed for B.

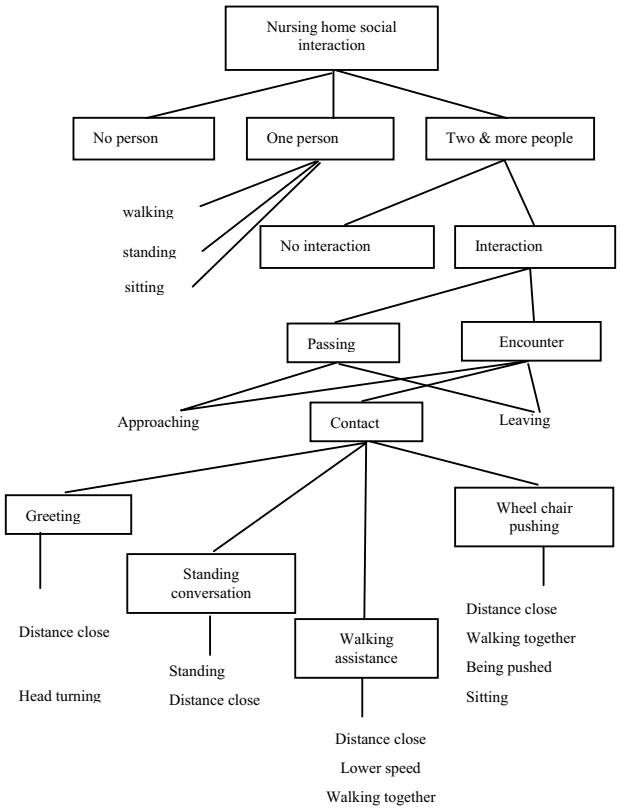

Figure 4. Social interaction ontology in a nursing home

Group activity features and events (GEs) are combinations of IEs that involve two individual person entities as listed in Table 3. We prefer to process GEs at different levels with SIs, because most GEs are only measures of relative motions of two IEs, which should be called features. 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*. Some GEs require that long period temporal information be extracted. For example, in order to extract the event: "lower speed (*A, B*)", we need to check the speed history of a person over "a long period of time".

A social interaction (SI) is a sequence of IEs, GEs or a combination of other social interactions. If the observation time and the number of people involved are not limited, the number of possible interactions is too big to handle. In order to limit the set of the taxonomy of social interactions to a reasonable size, we define the taxonomy implicitly by the ontology shown in Figure 4. Due to the space limitation, the detailed connections from social interactions (the items in rectangle boxes) to other levels are not completely expanded. Based on this ontology, our analysis system interprets activities of a nursing home into sequences of social interactions.

4. IMPLEMENTATION OF THE SOCIAL INTERACTION ANALYSIS SYSTEM

The ontology of social interaction in a nursing home is mapped onto a dynamic Bayesian network (DBN), which represents not only the interaction and event hierarchy by its states and arches, but also the evolution of the interactions by temporal arches defined between the interactions. To illustrate the concept, a part of temporal structure of the DBN is depicted in Figure 5. Formally, the DBN *B=(S, M)* is a directed acyclic graph that consists of a state set $S = SI \cup SE \cup IE = \{s_1,...,s_n\}$, which represents events and interactions, a set of directed arches that specifies parents of each state *s*: *Parent(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 ontology 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 (SI) according to the relationships defined in the ontology. For example, the "interaction" is the father of "passing" and "Encounter". Using directed arches, we also defined two Sis to be fathers of each other. The temporal arches 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.

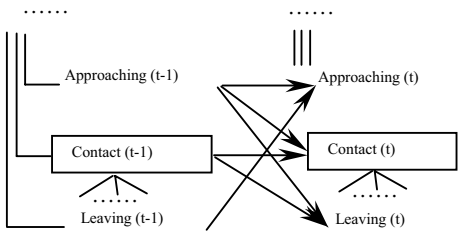

Figure 5. A partial DBN with temporal arches

4.1 Entity Detection

The facility is assumed to be almost stationary. We manually label the positions of all the doors and employ an adaptive background method [32] to register pixels of the ceiling, walls and the floor. A person can be detected initially by subtracting the registered background around the doors. The face is detected using the algorithm proposed in [37] and tracked using a Gaussian mixture skin color model [29] after faces are detected. "Body" is the extracted person region after removing the face if it is detected.

4.2 Entity Tracking & Attributes Extraction

Since occlusions happen very often in the narrow hallway, we use a particle filtering base multiple cameras 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 grabbed 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.

We 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*. 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 [2][3] or CONDENSATION, has shown to be a successful approach in several applications of computer vision [9][26]. 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:

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

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 feet 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 feet silhouette is often occluded, it can recover its 3D location of a person from 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.

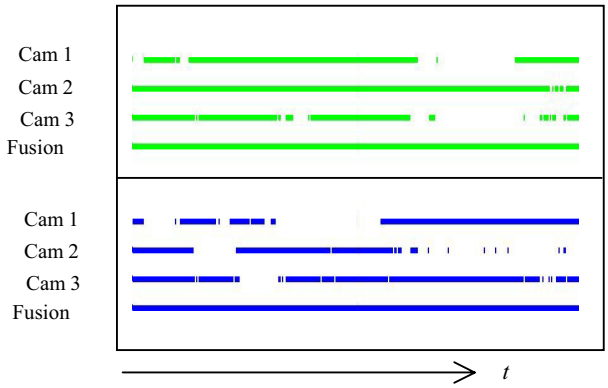

Figure 6. An illustration of people (2) tracking results using the proposed method. A color mark at time *t* indicates that the person is tracked by the corresponding camera or combination of cameras.

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. Therefore, the input of the event detection level is uniform attribute (feature) vectors per second. Shape information is represented by partitions with Manhattan distances. In this method, each extracted region that contains people or facility is divided into 9 sub-regions. Density of each sub-region is calculated and binarized to value '1' if it is greater than 50% and '0' otherwise. Finally, a shape feature vector of a region is a 10 dimensional vector: 9-dimension city blocks and width/height ratio of the region.

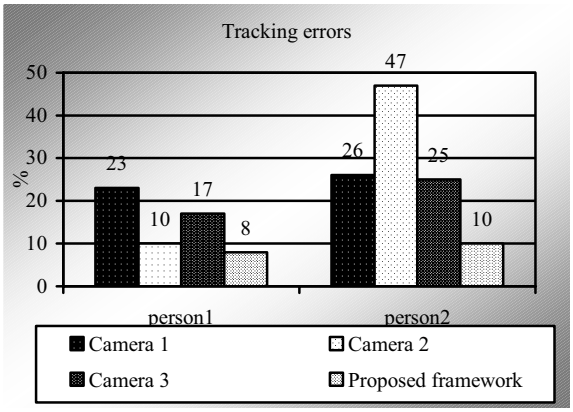

Figure 7. Tracking errors in 10 minute simulation video

4.3 IE and GE Detection

Each IE is modeled individually using Gaussian mixture models (GMMs). The training can be done using the standard EM algorithm [9]. A special case is that the event "being pushed" is considered to be an IE due to the difficulty of segmenting the wheelchair and the person who is pushing the wheelchair. A similar concept is also reused at the higher level but conditioned by other events or interactions. 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.

Some GEs require temporal information and are modeled by hidden Markov models (HMMs) based on individual event detection and raw features, such as "approaching", "leaving" and "lower speed". Others are modeled using GMMs directly based on features. When raw features are used, the input of a social event detector is two feature vector sequences from different persons. Although we use different models at this level, fortunately both HMMs and GMMs are not in conflict with the ontology DBN.

5. EXPERIMENTAL RESULTS

To evaluate the proposed framework, we have selected 160 short video sequences of social 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 is around 400 frames. To avoid interpreting very complex activities, most of the sequences contain social interactions only involving two persons. We manually labeled the ground truth of these video sequences.

Horizontal axis -> frame number; Vertical axis -> speed.

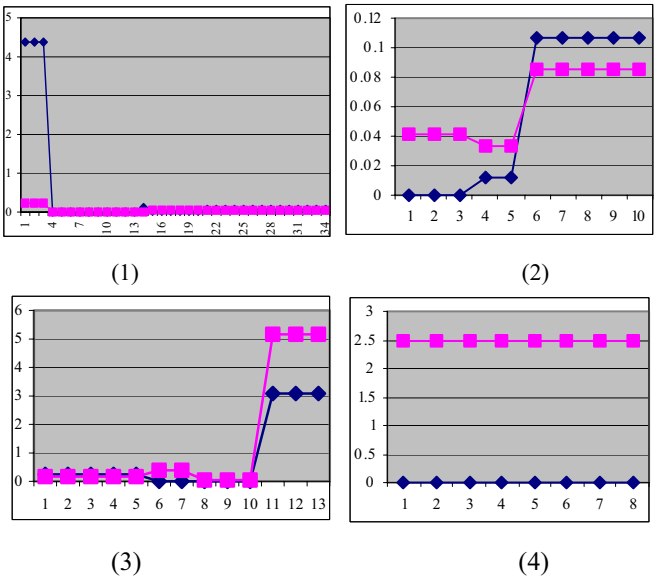

Figure 8. Speed features of some video samples: (1) approaching, greeting, stand conversation, walking assistance; (see original key frames in the 1st row of Fig. 1) (2) stand conversation, wheelchair pushing; (see Fig. 1, 2nd row, left) (3) approaching, stand conversation, leaving; (see Fig. 1, 2nd row, middle) (4) passing (see Fig. 1, 2nd row, right).

Horizontal axis -> frame number; Vertical axis -> distance.

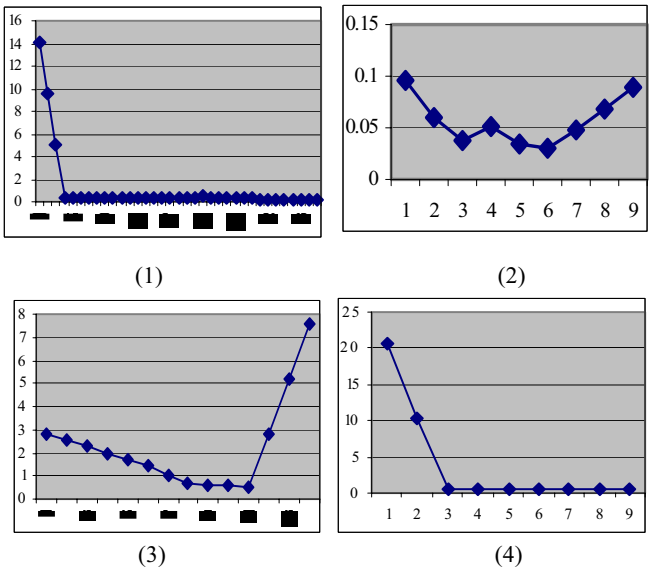

Figure 9. GEs "distance" extracted from the four video sequences described in Figure 8.

Figures 8-10 illustrate speed features, GEs, "distance", and "relative direction" 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 ("greeting") 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 ontology, video (1) can be simply interpreted as: "approaching (*A, B*) -> greeting (*A, B*) -> standing conversation (*A, B*) -> walking assistance (*A, B*)". Concisely, we can omit the subjects (persons) and interpret the video (2-4) as: (2) standing conversation -> wheelchair pushing; (3) approaching, standing conversation, leaving; (4) passing includes approaching -> distance close -> leaving. In Figure 8 and 9, 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.

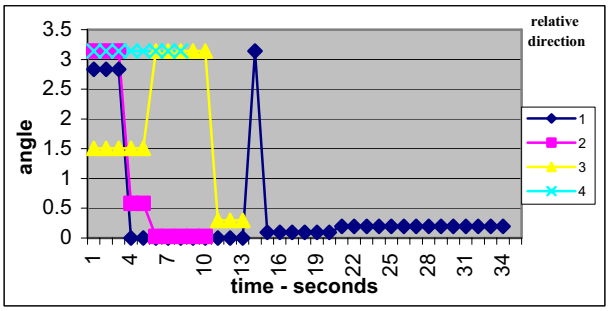

Figure 10. GEs "relative direction" extracted from the four video sequences described in Figure 8

We use 80 videos in the database as the training set and use the remaining 80 videos as the test set. Table 4 lists the number of interactions in the training set and test set. Only 6 interactions are listed here.

Table 4. Social interaction recognition results

Interactions	Training set	Test set	Recognition rate	False alarms
Passing	21	15	93%	4
Standing conversation	25	28	100%	7
Greeting	7	6	33%	2
Walking assistance	35	40	88%	4
Wheelchair pushing	5	4	75%	2
Encounter	59	65	94%	1

There may be more than one interaction in an "encounter"; for example, an "encounter" consisting of a "standing conversation" and a "walking assistance". Therefore the number of "encounters" is less than the sum of all its component interactions. We can see that detailed interactions (for example, "greeting" and "wheelchair pushing") are not well recognized. However, the recognition of higher lever concepts (such as "encounter" and "passing") is more than 93%. For transcribing, we can report a sequence of interactions for each patient using detail concepts if they are available or only report high-level concepts, which still can give an idea about the patient's activities.

Taxonomy categories of complex interaction patterns can be dynamically defined on the transcription produced by the ontology. For example, we are interested with the sequences that a nurse met a patient in the hall way and helped the patient to get some water to drink. We simply looking for the sequences consisting of the transcription "approaching, stand conversation (or greeting), leaving, approaching, stand conversation" using logical referencing power.

6. CONCLUSIONS

This paper proposes an ontology-based approach for analyzing social interactions in a nursing home using statistical methods. Social interaction is one of the most complex human activities in a nursing home and can provide potentially important information regarding long-term care patients. Traditional taxonomy frameworks have difficulty in categorizing social interactions using predefined criteria due to the large number of possible variations of the interaction processes. Ontology has the advantage in representing relationships by interpreting the evolving nature of the interaction processes. We have demonstrated that the proposed ontology-based framework can automatically interpret social interactions of two people in the hallway of a nursing home and provide transcriptions that may be used for describing activities of the elderly in a nursing home.

The evolving nature of interactions can be modeled using statistical approaches. As shown in this paper, a DBN model could be implemented corresponding to the hierarchical structure of the concepts in the ontology, based on event detection from individual to group. We have introduced an individual activity event level so that many existing techniques for analyzing individual human behaviors could be further re-used in social activity analysis. Using a DBN, we can also optimize the result interaction sequences with the maximum likelihood without making any hard decisions at the event detection levels.

At the detection and tracking level, we have focused on improving robustness of people tracking from a camera network. We have presented a method that integrates real world 3D position prediction and multiple camera-based tracking using particle filters. The tracked features from different cameras can enhance robustness of the entire tracking results. The method is also robust to the number (one or several) of un-occluded tracking images from the camera network, which is another advantage for application in an often-occluded environment such as the hallway. Using this method, we have reduced the tracking errors about 58% on average in our experiments. In our current work, we have ignored some parameters of social interaction, such as "head turning," which may lead to errors in interaction recognition, for example, "greeting". More precise detection methods need to be developed in the future.

We have showed that a DBN can easily be mapped from the structure of ontology. However, DBN-based generative models cannot fully take advantage of the categorization capability of ontology. They can model each of the categories individually without considering any discriminative criteria between those categories that belong to the same parent category. Discriminative models have a potential advantage when applied to ontology-based applications. We will work on developing discriminative model based DBNs in the future.

7. ACKNOWLEDGEMENTS

This research is supported by the National Science Foundation (USA) through project CareMedia No. 0205219. The authors would like to thank the colleagues in the Informedia for their valuable discussions and support.

8. REFERENCES

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

[2] K. Ahrens, S. F. Chung, and C. Huang, "Conceptual Metaphors: Ontology-based Representation and Corpora Driven Mapping Principles". In *Proceedings of the ACL Workshop on the Lexicon and Figurative Language*.

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

[4] N. Badler, "Temporal Scene Analysis: Conceptual Description of Object Movements," *University of Toronto Technical Report No. 80*, 1975.

[5] B. Brumitt, J. Krumm, B. Meyers, and S. Shafer, "Ubiquitous computing and the role of geometry". In Special Issue on Smart Spaces and Environments, volume 7-5, pages 41-43. IEEE Personal Communications, October 2000.

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

[7] J. W. Davis, A. F. Bobick, "The Representation and Recognition of Human Movement Using Temporal Templates," *Proc. of CVPR*, pp. 928-934, 1997.

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

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

[10] S. Farrar, W. Lewis, and T. Langendoen, "A Common Ontology for Linguistic Concepts". In *Proceedings of the Knowledge Technologies Conference*, 2002.

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

[12] A. Harter, A. Hopper, P. Steggles, A. Ward, and P. Webster, "The anatomy of a context-aware application". In Proceedings of the 5th Annual ACM/IEEE International Conference on Mobile Computing and Networking, pages 59-68, Seattle, WA, August 1999. ACM Press.

[13] S. Hongeng, R. Nevatia, "Multi-Agent Event Recognition," *International Conference on Computer Vision*, pp. 84-91, 2001.

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

[15] Y. A. Ivanov, A. F. Bobick, "Recognition of Visual Activities and Interactions by Stochastic Parsing," *IEEE Transactions of Pattern Analysis and Machine Intelligence*, Vol. 22, pp. 852-872, 2000.

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

[17] C. D. Kidd, R. Orr, G. D. Abowd, C. G. Atkeson, I. A. Essa, B. Macintyre, E. Mynatt, and T. E. Starner and W. Newstetter, "The Aware Home: A Living Laboratory for Ubiquitous Computing Research". *Proc. of CoBuild '99*, pp.191-198, 1999.

[18] K. Koile, K. Tollmar, D. Demirdjian, H. E. Shrobe, T. Darrell, "Activity Zones for Context-Aware Computing", Ubicomp 2003, pp. 90-106, 2003.

[19] A. Kojima, T. Tamura, "Natural Language Description of Human Activities from Video Images Based on Concept Hierarchy Actions," *International Journal of Computer Vision*, Vol. 50, pp. 171-184, 2001.

[20] R. Lubinski, "Dementia and communication", Philadelphia: B. C. Decker, 1991.

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

[22] 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.

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

[24] I. Niles and A. Pease, "Towards a Standard Upper Ontology". In Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001), Chris Welty and Barry Smith, eds, 2001.

[25] 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.

[26] 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.

[27] N. M. Oliver, B. Rosario, A. P. Pentland, "A Bayesian Computer Vision System for Modeling Human Interactions," *IEEE Transactions of Pattern Analysis and Machine Intelligence* , Vol. 22, pp. 831-843, 2000.

[28] A. Pease and I. Niles, "Practical Semiotics: A Formal Theory". In *Proceedings of the 2002 International Conference on Information and Knowledge Engineering*, Las Vegas, Nevada, June 24-27, 2002.

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

[30] 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.

[31] J. F. Sowa, "Building, Sharing, and Merging Ontologies", web site: http://www.jfsowa.com/ontology/ontoshar.htm.

[32] D. Subrata, K. Shuster, and C. Wu, "Ontologies for Agent-Based Information Retrieval and Sequence Mining". In *Proceedings of the Workshop on Ontologies in Agent Systems*, International Joint Conference on Autonomous Agents and Multi-Agent Systems Bologna, Italy, July 15-19, 2002.

[33] Time Domain Corporation, 7057 Old Madison Pike, Huntsville, AL 35806. PulsON Technology: Time Modulated Ultra Wideband Overview, 2001.

[34] 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.

[35] Y. Yacoob, M. J. Black, "Parameterized Modeling and Recognition of Activities," *International Conference of Computer Vision*, Vol. 73, pp. 232-247, 1998.

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

[37] D. Zhang, S. Z. Li, D. Gatica-Perez, "Real-Time Face Detection Using Boosting Learning in Hierarchical Feature Spaces," 17th International Conference on Pattern Recognition 2004.