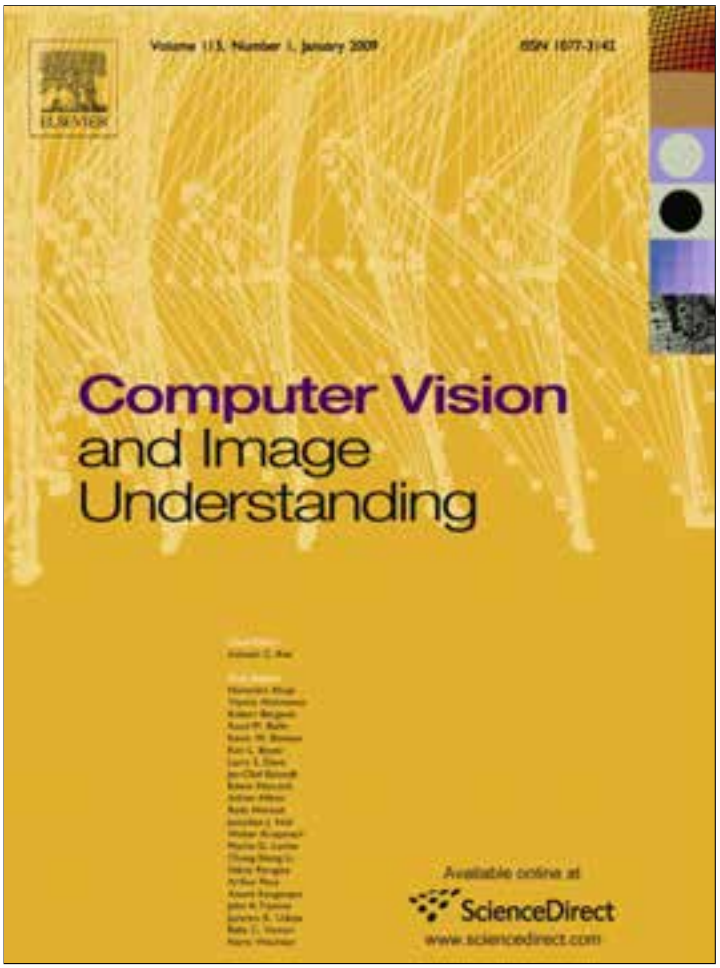

Computer Vision and Image Understanding 113 (2009) 80–89

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

Linguistic summarization of video for fall detection using voxel person and fuzzy logic

Derek Anderson [a,*], Robert H. Luke [a,1], James M. Keller [a,1], Marjorie Skubic [a,1], Marilyn Rantz [b,2], Myra Aud [b,2]

[a] Department of Electrical and Computer Engineering, University of Missouri, 349 Engineering Building West, Columbia, MO 65211-2300, USA
[b] Sinclair School of Nursing, University of Missouri, Columbia, MO 65211, USA

ARTICLE INFO

Article history:
Received 7 November 2007
Accepted 11 July 2008
Available online 29 July 2008

Keywords:
Linguistic summarization
Activity analysis
Fuzzy logic
Fall detection
Eldercare
Voxel person

ABSTRACT

In this paper, we present a method for recognizing human activity from linguistic summarizations of temporal fuzzy inference curves representing the states of a three-dimensional object called voxel person. A hierarchy of fuzzy logic is used, where the output from each level is summarized and fed into the next level. We present a two level model for fall detection. The first level infers the states of the person at each image. The second level operates on linguistic summarizations of voxel person's states and inference regarding activity is performed. The rules used for fall detection were designed under the supervision of nurses to ensure that they reflect the manner in which elders perform these activities. The proposed framework is extremely flexible. Rules can be modified, added, or removed, allowing for per-resident customization based on knowledge about their cognitive and physical ability.

1. Introduction

Falls are a severe problem among the elderly. Many elders fall and sustain an injury or remain on the floor for long durations until someone discovers them, further compounding the severity. Our goal is the continuous monitoring of human activity for the assessment of the "well-being" of a resident and the detection of abnormal or dangerous events, such as falls. It is important that both theories and realistic technologies be developed for recognizing elderly activity, and that they do so in a non-invasive fashion. Video sensors are a rich source of information that can be used to monitor a scene, but privacy is always a concern. To preserve privacy, segmentation of the human from an image results in a silhouette, a binary map that distinguishes the individual from the background. The raw video is not stored and only silhouettes are used to track the individual's activity.

Silhouette extraction, namely, segmenting the human from an image with the camera at a fixed location, is the first stage in video-based activity analysis. The standard approach involves constructing a background model and regions in subsequent images with significantly different characteristics are classified as foreground [1–10]. Stauffer and Grimson introduced an adaptive method for background modeling and subtraction that utilizes a mixture of Gaussians per pixel with a real-time, online approximation to the model update [1]. Oliver et al. carry out foreground segmentation in eigenspace, where the background is modeled as an eigenbackground, however, no model update method was proposed [2]. These two well known approaches focus on adaptation and background modeling at a relatively low level of computer vision. They do not present robust features for change detection or medium to high level computer vision algorithms for region identification and tracking. The Wallflower algorithm addresses a wider range of extreme real-world conditions for complex and dynamic environments through pixel-level probabilistic background prediction with a Wiener filter, region-level processing, and heuristics for global sudden change detection and model correction [3]. We proposed an adaptive system that uses higher level computer vision for background modeling and reliable change detection through fusing new texture and color histogram-based descriptors and a modified hue, saturation, and value (HSV) space for shadow removal [4].

While change detection is full of technical challenges, even after the human is segmented from the background a larger problem arises regarding the higher level processing of this information for recognizing activity and detecting deviations from patterns of normal activity. The most widely accepted approaches to modeling human activity include; graphical models [11], dynamic Bayesian networks [12], also known as dynamic graphical models, and more specifically, hidden Markov models (HMMs) [2,13] and its variants

* Corresponding author. Fax: +1 573 882 0397.
E-mail addresses: dtaxtd@mizzou.edu (D. Anderson), rhl3db@mizzou.edu (R.H. Luke), kellerj@missouri.edu (J.M. Keller), skubicm@missouri.edu (M. Skubic), rantzm@health.missouri.edu (M. Rantz), audm@health.missouri.edu (M. Aud).
[1] Fax: +1 573 882 0397.
[2] Fax: +1 573 884 4544.

1077-3142/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.cviu.2008.07.006

(hierarchical HMMs [14], entropic-HMMs [15], coupled-HMMs [2,16], etc).

In the area of short-term activity recognition, we used HMMs for fall detection [13]. Our preliminary results indicate that a single camera, geometric features calculated from silhouettes, and HMMs can be used to detect some types of falls under a constrained set of view dependent assumptions about how and where activities are performed in the environment. However, while HMMs can be used to identify a maximum likely model, from one of K known models, they are not presently sufficient for rejecting unknown activity. Thome and Miguet used hierarchical HMMs (HHMM) for video-based fall detection [14]. The interesting aspect of that research is the feature employed in the model. They use image rectification to derive relationships between the three-dimensional angle corresponding to the individual's major orientation and the principal axis of an ellipse fit to the human in a two-dimensional image. The HHMM is hand designed and operates on an observation sequence of rectified angles.

Martin et al. presented a soft-computing approach to monitoring the "well-being" of elders over long time periods using non-video sensors such as passive infrared, toggle switches, vibration, temperature, and pressure sensors [17]. They outlined data analysis methods based on fuzzy reasoning, statistics, association analysis, and trend analysis. Procedures for interpreting firings from relatively simple sensors into fuzzy summaries were presented. These summaries assist in characterizing resident's trends and aid in answering queries about deviations from these patterns, such as "has the occupant's sleep pattern changed significantly in the past few months".

In [18], we presented a method for constructing a three-dimensional representation of the human from silhouettes acquired from multiple cameras monitoring the same scene. Fuzzy logic is used to determine the membership degree of the person to a pre-determined number of states at each image. In this paper, a method is presented for generating a significantly smaller number of rich linguistic summaries of the human's state over time, in comparison to the large number of state decisions made at each image, and a procedure is introduced for inferring activity from features calculated from linguistic summarizations. Summarization and activity inference makes fall detection possible, something that was not accomplished in our earlier work. The next section is an overview of voxel person construction and state reasoning. Some material is reported again here because it is necessary for understanding the approach taken in this paper.

2. Fuzzy logic for voxel person state classification

Our approach to monitoring human activity is based on fuzzy set theory. Fuzzy set theory, introduced by Lotfi A. Zadeh in 1965 [19], is an extension of classical set theory. The memberships of elements in a set are allowed to vary in their degree, instead of being restricted to two values, as in classical set theory. A fuzzy set is defined over a particular domain, and it is characterized by a membership function that maps elements from the domain to a real valued number, $\mu_A : A \rightarrow [0,1]$. The fuzzy sets used in this paper are trapezoidal membership functions, which are characterized according to four ordered numbers, {a,b,c,d}. The membership of the element x in the fuzzy set A is

$$\mu_A(x) = \text{maximum}\left(\text{minimum}\left(\frac{(x-a)}{(b-a)}, 1, \frac{(d-x)}{(d-c)}\right), 0\right).$$

The way in which fuzzy set theory includes and models uncertainty has led to extremely valuable applications in mathematics and engineering [20–23]. One of the more well known branches of fuzzy set theory is fuzzy logic, introduced by Zadeh in 1973 [20]. Fuzzy logic is a powerful framework for performing automated reasoning. An inference engine operates on rules that are structured in an IF–THEN format. The IF part of the rule is called the antecedent, while the THEN part of the rule is called the consequent. Rules are constructed from linguistic variables. These variables take on the fuzzy values or fuzzy terms that are represented as words and modeled as fuzzy subsets of an appropriate domain. An example is the fuzzy linguistic variable "height of voxel person's centroid", which can assume the terms low, medium, and high, all defined as membership functions over an appropriate numerical domain. In this work, we use the standard Mamdani-Assilion fuzzy inference system [20,24].

What is needed in the area of human activity analysis is not another non-interpretable likelihood value that is useful for classifying one of K known models or the ad hoc training of garbage models [25] for reducing false alarms, but a confidence value that can be understood and reliably used to reject a wide range of unknown activities. The core representation and computing basis in our work is significantly different from most. We believe that fuzzy set theory and fuzzy logic are necessary in order to address the inherent uncertainty related to modeling and inferring human activity. Linguistic variables are used to describe features extracted from a three-dimensional representation of the human. A separate set of linguistic variables are used for representing the human's state and activity. Fuzzy logic is used for inferring the state and activity. The system's output are membership values that reside in [0,1], fuzzy sets (terms) have been defined and assist in the interpretation of these values, and fuzzy logic is the inference mechanism. A fuzzy approach also has the advantage that the rules and linguistic variables are understandable and simplify addition, removal, and modification of the system's knowledge.

Multiple cameras that jointly view the same environment are crucial for the reliable recognition of activity. Different viewpoints assist in coping with issues like occlusion and makes the construction of three-dimensional objects possible. After silhouettes are individually extracted from each camera in a scene, a three-dimensional representation of the human is constructed in voxel space, which we call voxel person. Like pixels in a two-dimensional image, a voxel (volume element) is an element resulting from a discretization of three-dimensional space. A voxel is defined here as a non-overlapping cube. The set of voxels belonging to voxel person at time t are $V_t = \{\bar{v}_{t,1}, \bar{v}_{t,2}, \ldots, \bar{v}_{t,P}\}$, where the center of the jth voxel at time t is $\bar{v}_{t,j} = \langle x_j, y_j, z_j \rangle^{\mathrm{T}}$. The capture time for each camera is recorded and the silhouettes, one from each camera, that are the closest in time are used to build V_t. The construction of voxel person from a single camera is the planar extension of the silhouette along the direction of the camera viewing angle. Voxels in the monitored space that are intersected by this planar extension are identified. The projection procedure involves using the camera's intrinsic parameters to estimate pixel rays, and these rays are tested for intersection with voxels [18]. Voxel person, according to camera i ($1 \leqslant i \leqslant C$) at time t is V_t^i, whose cardinality, $|V_t^i|$, is P_i. The planar extensions of voxel person from multiple cameras, $\{V_t^1, \ldots, V_t^C\}$, are combined using an operation, such as intersection, $V_t = \Lambda_{i=1}^{c} V_t^i$, to assemble a more accurate object representation. An illustration of voxel person construction from two cameras is shown in Fig. 1. In [18], further processing is performed on voxel person to remove additional shadows and reflections given a priori knowledge about the three-dimensional environment. Voxel person's volume is analyzed to detect error time intervals, and an efficient method for dynamically increasing the resolution (detail) of voxel person is discussed.

For each image, the goal is the calculation of the membership degree of voxel person to a set of pre-determined states. This state information is used to infer activity. An activity is characterized according to state duration, frequency of state visitation, and state

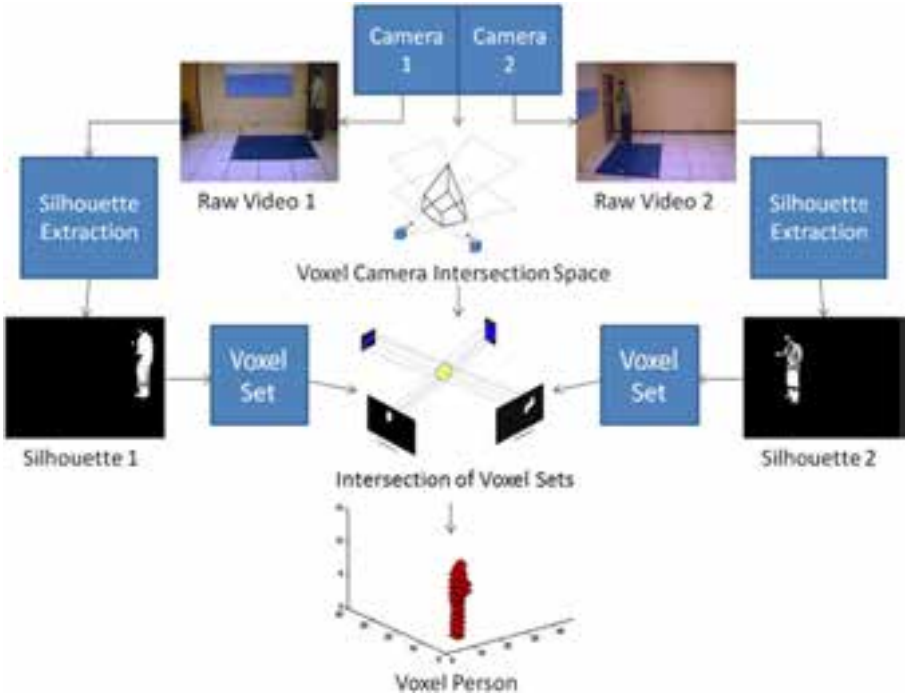

Fig. 1. Voxel person construction. Cameras capture the raw video from different viewpoints, silhouette extraction is performed for each camera, voxel sets are calculated from the silhouettes for each camera, and the voxel sets are intersected to calculate voxel person.

transition behavior. The collection of states, where state i is denoted by S_i, that we have identified for fall recognition include

- **Upright** (S_1): This state is generally characterized by voxel person having a large height, its centroid being at a medium height, and a high similarity of the ground plane normal with voxel person's primary orientation. Activities that involve this state are, for example, standing, walking, and meal preparation.
- **On-the-ground** (S_2): This state is generally characterized by voxel person having a low height, a low centroid, and a low similarity of the ground plane normal with voxel person's primary orientation. Example activities include a fall and stretching on the ground.
- **In-between** (S_3): This state is generally characterized by voxel person having a medium height, medium centroid, and a non-identifiable primary orientation or high similarity of the primary orientation with the ground plane normal. Some example activities are crouching, tying shoes, reaching down to pick up an item, sitting in a chair, and even trying to get back up to a standing stance after falling down.

It is important to note that being on the ground, a state, does not imply a fall, an activity. The method presented in this paper is required for reasoning about activity, such as falls. None of the features above sufficiently identify voxel person's state the majority of the time. **On-the-ground** can include variations of the three features depending on how he or she fell and our interpretation of the state. In addition, each state is difficult to classify from the features alone, which is further complicated by noise resulting from the segmentation process. Each feature can be used to help determine a degree to which voxel person is in a particular state. Descriptions such as a large, medium, or a low amount of each feature characterize the states. There is no crisp point where the features change between states. These factors lead us to use fuzzy inference to classify voxel person's present membership in each state, and ultimately, to recognize human activity.

In [18] we showed how robust statistical features can be extracted from voxel person for the goal of inferring the state by fuzzy logic. Voxel person features include the: centroid, eigen-based height (more robust to noise than calculating the maximum observed voxel height value), major orientation of the body, and similarity of the major orientation with the ground plane normal (a rough indication of if the individual is standing upright or lying on the ground). There are 24 rules for each of the three output states, whose values were empirically determined under the guidance of nurses. We note, however, that these values, and even the set of rules, can be learned from training data.

3. Temporal linguistic summarization of video

The result of reasoning about the state of voxel person at time is three membership values corresponding to the confidence of being **upright**, **in-between**, and **on-the-ground**, $\vec{\mu}_t = \langle \mu_{t,1}, \mu_{t,2}, \mu_{t,3} \rangle$. Decisions regarding activity can be made at each image from the state memberships, but the result is too much information. The objective is to take seconds, minutes, hours, and even days of resident activity to produce succinct linguistic summarizations, such as "the resident was preparing lunch in the kitchen for a moderate amount of time" or "the resident has fallen in the living room and is down for a long time". This is a situation in which less information is more useful. Reporting activity for every frame results in information overload. Linguistic summarization is designed to increase the understanding of the system output, and produce a reduced set of salient descriptions that characterizes a time interval. The linguistic summarizations help in informing nurses, residents, residents' families, and other approved individuals about the general welfare of the resident, and they are the input for the automatic detection of cognitive or functional decline or abnormal event detection.

State summarizations are produced through the temporal processing of the fuzzy inference results regarding voxel person's state. The sequence $D = \{\vec{\mu}_1, \ldots, \vec{\mu}_N\}$ has N elements, e.g. state

decisions for N images. An example plot of the membership outputs over time is illustrated in Fig. 2. The camera's capture rate was three frames per second and the scenario in Fig. 2 represents a person falling and not getting back up.

In the case of three states, voxel person can be color coded to illustrate the state memberships of the resident. The membership consequent values, all within [0,1], determine the amount of red, blue, and green in voxel person. Fig. 3 is a sequence that shows the color coding of voxel person for four frames of the sequence shown in Fig. 2. Movies illustrating voxel person fall detection are available for download at http://cirl.missouri.edu/fallrecognition. These movies include the raw video feed, the silhouettes, color coded voxel person, the accompanying fuzzy rule base outputs, linguistic summarizations, and fall confidences.

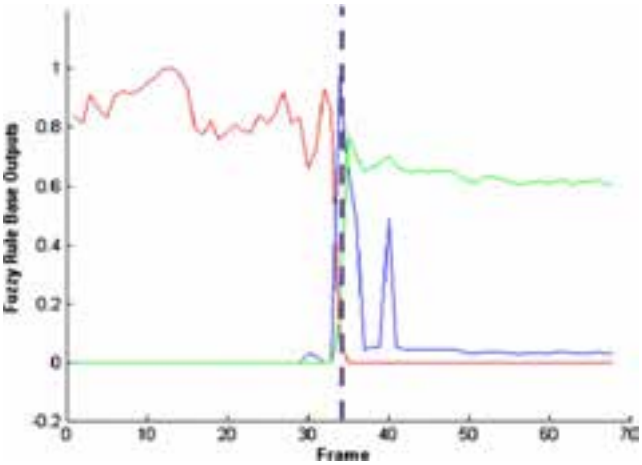

Fig. 2. Fuzzy inference outputs plotted for a voxel person fall. The x-axis is time, measured in frames, and the y-axis is the fuzzy inference outputs. The red curve is **upright**, the blue curve is **in-between**, and the green curve is **on-the-ground**. The frame rate was three per second, so the above plot is approximately 23 s of activity.

Before D is summarized, elements from D in which the maximum membership value, $\mu_{t,\max} = \text{maximum}_k(\mu_{t,k})$, is not clearly distinguishable from the other memberships are removed. The new sequence, D', has cardinality $N' = \mid D' \mid$, where $N' \leqslant N$. A parameter, $\tau_1 \in [0,1]$ for indeterminate maximum state identification, is used along with a parameter, $\tau_2 \in [0,1]$, for removing elements where $\mu_{t,\max}$ is below some acceptable membership value (e.g. the confidence is too low). We experimentally determined τ_1 to be 0.1 and τ_2 to be 0.5. An element $\bar{\mu}_t$ is removed if

$$((\mu_{t,\max} - \text{maximum}_{j \neq \text{argmax}_k(\mu_{t,k})}(\mu_{t,j})) < \tau_1) \text{ and } (\mu_{t,\max} > \tau_2).$$

Removed elements are not used in summarization. Because N' does not always equal N, each $\bar{\mu}'_t$, an element in D', has its original position in D recorded, $I = \{i_1, \ldots, i_{N'}\}$. A maximum state index sequence, $\omega = \{s_1, \ldots, s_{N'}\}$, is also constructed, where $s_t = \text{argmax}_k(\mu'_{t,k})$.

Linguistic summarization of the state membership values is the generation of meaningful human understandable information of the form

X_c is S_i in $P_k$for T_j.

The object of interest, voxel person, is denoted as $X_c (1 \leqslant c \leqslant C$, where C is the number of objects being tracked). Here, only a single resident is tracked, hence $C = 1$, but it is possible to detect and track multiple disjoint voxel objects either through labeling regions in the silhouettes, a computer vision classification task, or through the

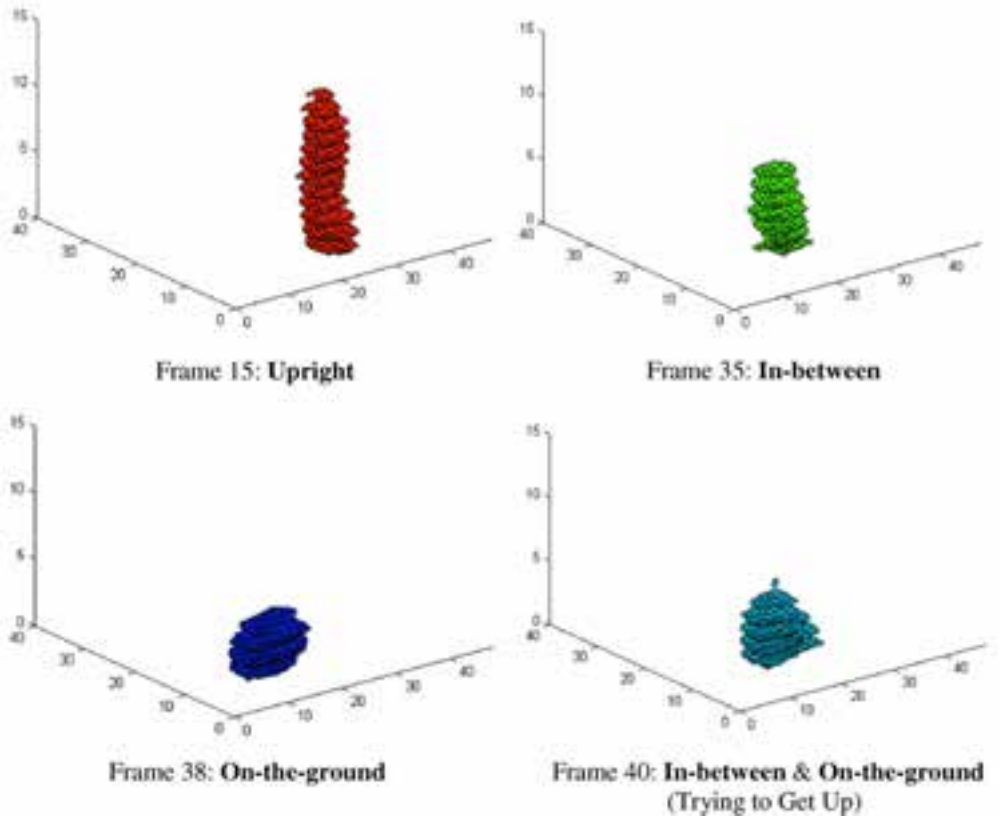

Fig. 3. Frame 15: **upright**, frame 35: **in-between**, frame 38: **on-the-ground**, frame 40: **in-between** & **On-the-ground** (trying to get up). Color-coding of voxel person according to the membership output values. Voxel persons color is a mixture of the fuzzy rule system outputs. The **upright** state determines the amount of red, **in-between** is green, and **on-the-ground** is blue.

analysis of connected and disjoint voxel regions. Voxel person's ith state is S_i. Important world segments, $P_k (1 \leqslant k \leqslant K$, where K is the number of segments), are recorded. The scene is manually partitioned into K non-overlapping segments. Example locations might include the living room, kitchen, and other areas that provide a context for subsequent activity analysis. The duration of each linguistic summarization is $T_j (1 \leqslant j \leqslant J$, where J is the number of fuzzy sets defined over the time domain). The quantity X_c is crisp, while S_i and T_j are fuzzy sets. The apartment location, P_k, can be crisp or fuzzy. We use a crisp P_k in this paper. An example linguistic summarization of this form is "voxel person is **on-the-ground** in the living room for a moderate amount of time".

The filtered sequence D' contains G summarizations, which are found by partitioning the maximum state index sequence ω. Indices where $s_i \neq s_{i+1}$, for $1 \leqslant i \leqslant (N' - 1)$, are recorded, $U = \{u_1, \ldots, u_{G-1}\}$ when $G > 1$ and $U = \emptyset$ when $G = 1$. Indices 1 and N' are added to U, hence U' is denoted by $\{1, U, N'\}$, where $| U' | = G + 1$, and the gth summary $(1 \leqslant g \leqslant G)$, Sum_g, is the sequence from u'_g to u'_{g+1}, where $u'_g \in U'$. Since the goal is the recognition of elderly activity, specifically falls, we generate summaries representing sufficient time periods of consecutive state occupancy. Our goal is not the recognition of high frequency activity occurring in a fraction of a second. These periods are possibly due to incorrect silhouette segmentation, inaccuracies in fuzzy inference, or high frequency activity that is not related to fall detection. Elders do not generally perform extremely quick activities, such as being on the ground for only one second. Any Sum_g, where $| Sum_g | < \tau_3$ is removed. The parameter τ_3 was determined to be F^*2, where F is the number of frames captured per second, $F = 3$ in our system, thus any summary that is less than 2 s in duration is removed. Also, any Sum_g whose sequence of image indices is not consecutive, $(i_{u'_g+k} + 1) \neq i_{u'_g+(k+1)}$, where $0 \leqslant k \leqslant (|Sum_g| - 2)$, is removed. This removes segments that were broken up by a brief time interval where there was another state with a larger membership value or segments that were broken up by a brief time interval of too low a state membership confidence. The result of these two filters is a new sequence that has G' summarizations, $\{Sum'_1, \ldots, Sum'_{G'}\}$. The first stage of automated linguistic summarization for the sequence shown in Fig. 2 is

Derek is **upright** in the laboratory for 11 s
Derek is **on-the-ground** in the laboratory for 11.3 s.

The individual's name, Derek, is included in the linguistic summarization. This personalizes the summarizations, increasing readability for an end user or health care individual interested in analyzing the activity of the resident. The location is determined by looking at voxel person's (x, y)position and finding which scene segment he or she is presently in.

The final step in summarization involves the generation of temporal linguistic descriptions. We use a single linguistic variable over the time domain that has the following terms, specified in seconds, with corresponding trapezoidal membership functions: brief = [−1 1 1 2], short = [1 5 10 15], moderate = [10 120 480 720], and long = [480 900 86400 86400]. These fuzzy sets were determined by nurses to make sure that they reflect older adults, the target group for this system. The full precision time value is not discarded by the system, it is just not included as part of the linguistic summary. The nurses fuzzy set 'long' occupies the majority of the time domain. This can result in problems as it relates to the sampling rate for a domain in inference. Long is redefined to be [480 900 900 901], the minimum of the time value and 900 is calculated, and a large sample rate is used during inference. The summarization above has its time components converted into the appropriate fuzzy terms to linguistically report the time durations. Each term and its respective membership degree can be reported. However, for the sake of display, only the term with the maximum confidence is presently being reported. The final summarization for the fuzzy rule base output shown in Fig. 2 is therefore

Derek is **upright** in the laboratory for a moderate amount of time
Derek is **on-the-ground** in the laboratory for a moderate amount of time.

Thus, we demonstrated a method for generating linguistic summarizations of the form X_c is S_i in P_k for T_j for fall detection. A generalization of this form is X_c is S in P for T, where S, P, and T are now linguistic variables, not a single fuzzy set. If desired, the membership degrees with respect to each fuzzy set for each linguistic variable can be utilized by an activity recognition system. In the next section, we present a higher level reasoning module that processes information from linguistic summarizations of the form X_c is S_t in P_k for T_j.

4. Fuzzy logic for fall detection

In this section, we describe a hierarchical system of fuzzy inference for activity reasoning and report our baseline system for fall detection. The variables and rules in our expert system are not automatically learned from the video data; they are manually determined by engineers and nurses at the University of Missouri. Linguistic variables were identified by the engineers and validated by the nurses. Nurses also validated the existing rules constructed by the engineers and created new rules based on their direct knowledge and experience with falls of older adults.

As already discussed, a hierarchy of fuzzy logic systems is used to recognize human activity. The first level was described above, which involves acquiring the confidences in states. The next level of fuzzy logic performs activity recognition from features computed from linguistic summaries. This second layer uses domain expert knowledge about activities to produce a confidence in the occurrence of an activity, which we were unable to reliably produce using HMMs. Rules allow for the recognition of common performances of an activity, as well as the ability to model special cases, which are extremely difficult to do with an HMM. This flexible framework also allows for rules to be added, deleted, or modified to fit each particular resident based on knowledge about their typical daily activities, physical status, cognitive status, and age. In addition, our approach is not restricted with respect to the amount of time that it can include to evaluate activity, which is not the case for the typical standard first-order HMM with Markov and i.i.d. assumptions. Many linguistic summarizations can be used, giving rise to variable amounts of time, when evaluating rules. This makes it possible to enforce longer-term specific performances of activities. Fig. 4 is our activity recognition framework.

Linguistic summarization results in a reduction of information in a format that is understandable by a human. However, while this information is useful to a human observer, our goal is the automation of reasoning about activity. To achieve this goal, features are extracted from linguistic summarizations, which in return are used by a fuzzy inference system to recognize activity. An advantage in generating linguistic summarizations is that they are the trigger to look for a fall once an **on-the-ground** is observed. They are also a reduced expression of the original higher sampled video information that typically results in long observation sequences as inputs to HMMs. The check for a fall is repeated at a user specified interval rate during an **on-the-ground** summary (here, every 5 s), until a fall is detected or the resident makes it out of the **on-the-ground** state.

There is a single linguistic output variable that reflects the confidence that a fall has occurred. This output variable is comprised of the terms low = [−0.5 −0.2 0.2 0.5], medium = [0.1 0.5 0.5 0.9], and high = [0.5 0.8 1.2 1.5]. The inputs to the second level in the

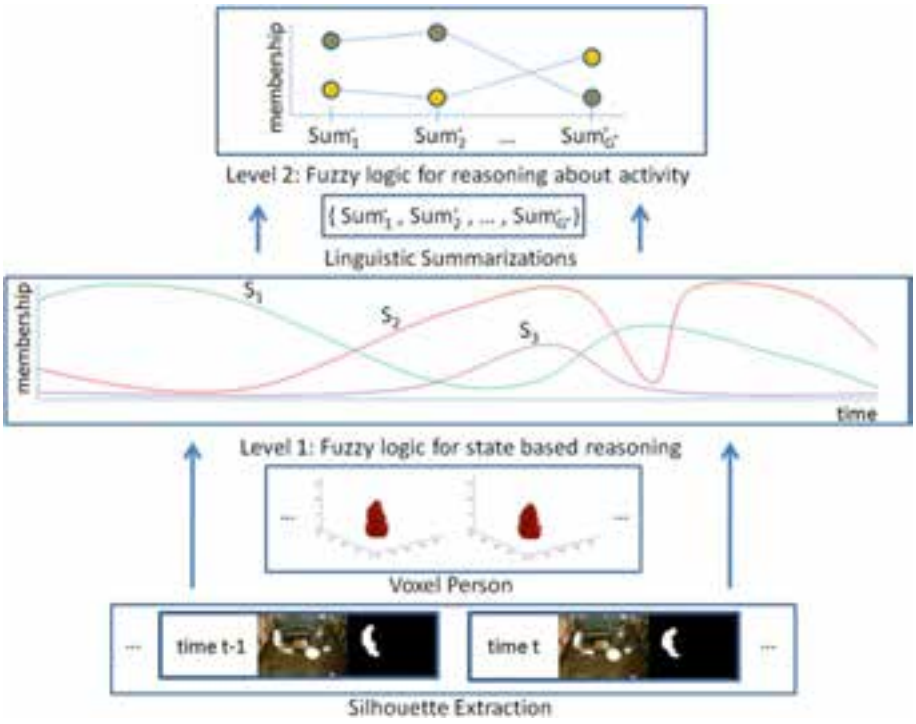

Fig. 4. Activity recognition framework, which utilizes a hierarchy of fuzzy logic based on voxel person representation. The first level is reasoning about the state of the individual. Linguistic summarizations are produced and fuzzy logic is used again to reason about human activity.

fuzzy logic hierarchy, for a single summary, include: the average state membership, time duration, confidence in a quick change in voxel person's speed before the start of the summary, voxel person's average speed, and the confidence in recent oscillating behavior between the **on-the-ground** and **in-between** states.

The first feature is the average membership for the state with the maximum membership during Sum'_g. The average membership for state i, S_i, in Sum'_g is

$$\pi_{i,Sum'_g} = \left(\frac{1}{|Sum'_g|}\right) \sum_{j=0}^{|Sum'_g|-1} \mu_{t_{\text{start}}+j,i},$$

where t_{start} is the first index in Sum'_g. The fuzzy sets for the average membership are low = [−0.5 −0.2 0.2 0.5], medium = [0.1 0.5 0.5 0.9], and high = [0.5 0.8 1.2 1.5].

The detection of a large recent change in voxel person's speed involves the analysis of voxel persons centroid, $\vec{c}_t = (1/P)\sum_{j=1}^{P} \vec{v}_{t,j}$. The motion vector between time t and $t+1$ is $\vec{m}_{t\to t+1} = \vec{c}_{t+1} - \vec{c}_t$. The magnitude of the motion vector, $\|\vec{m}_{t\to t+1}\|$, is an indicator of the person's speed. A window of size W of magnitudes of motion vectors is analyzed before an **on-the-ground** summary, $\{\|\vec{m}_{t\to W\to t-W+1}\|, \ldots, \|\vec{m}_{t-1\to t}\|\}$. The parameter W was experimentally determined to be 40 (approximately 13 s). Elements in this window are first smoothed with a mean filter of size 5, $\|\vec{m}'_{t\to t+1}\| = (1/5)\sum_{j=-2}^{2}\|\vec{m}_{t+j\to t+j+1}\|$. The derivative is then calculated using forward finite difference, $\nabla\|\vec{m}'_{t\to t+1}\| = \|\vec{m}'_{t+1\to t+2}\| - \|\vec{m}'_{t\to t+1}\|$. The detection of a quick change involves identifying a relatively large change in the sequence $\nabla\|\vec{m}'_{t\to t+1}\|$ in the second half of the window. The maximum $\nabla\|\vec{m}'_{t\to t+1}\|$ from the first half of the window, $sd_{\text{half1}} = \text{maximum}(\{\text{maximum}_i(\nabla\|\vec{m}'_{t-W+i\to t-W+i+1}\|), 0\})$, where $i = \{0, \ldots, \lfloor W/2\rfloor - 1\}$, and the maximum in the second half of the window, $sd_{\text{half2}} = \text{maximum}(\{\text{maximum}_i(\nabla\|\vec{m}'_{t-W+i\to t-W+i+1}\|), 0\})$, where $i = \{\lfloor W/2\rfloor, \ldots, W-1\}$, are calculated. The feature is $(sd_{\text{half2}}/sd_{\text{half1}})$ and the corresponding fuzzy sets are low = [0 1 1 1.3], medium = [0.8 1.2 1.2 1.6], and high = [1.5 2 100 102]. Fig. 5 illustrates this procedure for the sequence shown in Fig. 2.

Voxel person's average speed during Sum'_g is

$$\varphi_{Sum'_g} = \left(\frac{1}{(|Sum'_g| - 1)}\right) \sum_{j=0}^{(|Sum'_g|-2)} \|\vec{m}_{t_{start}+j\to t_{start}+j+1}\|,$$

where t_{start} is the index of the first motion vector magnitude in Sum'_g. The fuzzy sets for **on-the-ground** motion are low = [−0.2 0 0 0.2] and high = [0.1 0.4 100 102].

The measure of recent oscillating behavior between **on-the-ground** and **in-between**, which is useful for detecting if a resident has fallen and is trying to make it back up, involves searching the sequence of summarizations backwards from the current **on-the-ground** summary until the end of a moderate amount of time, where moderate is the fuzzy set determined by the nurses. The fourth trapezoid value, d, the set end point, for moderate is used to determine when to terminate the search. The number of times that the system changed between **on-the-ground** and **in-between** is counted. If an **upright** state is encountered, the counting stops. The fuzzy sets for recent oscillating behavior between the **on-the-ground** and **in-between** states are low = [−2 0 2 4], medium = [1 3 5 7], and high = [4 6 8 10]. The minimum of the recent oscillating state behavior count and 8 is calculated. The nurses indicated that anything over 8 is high, so taking the minimum helps avoid any consequence domain sampling issues that could arise if the set 'high' was extended outwards with respect to the points c and d. The following rules, reported in Table 1, are currently used to detect a fall.

5. Experiments and results

The reliable recognition of different ways in which elders fall is the subject of analysis in this section. Successfully recognized fall types are highlighted and false alarms are discussed. All data was

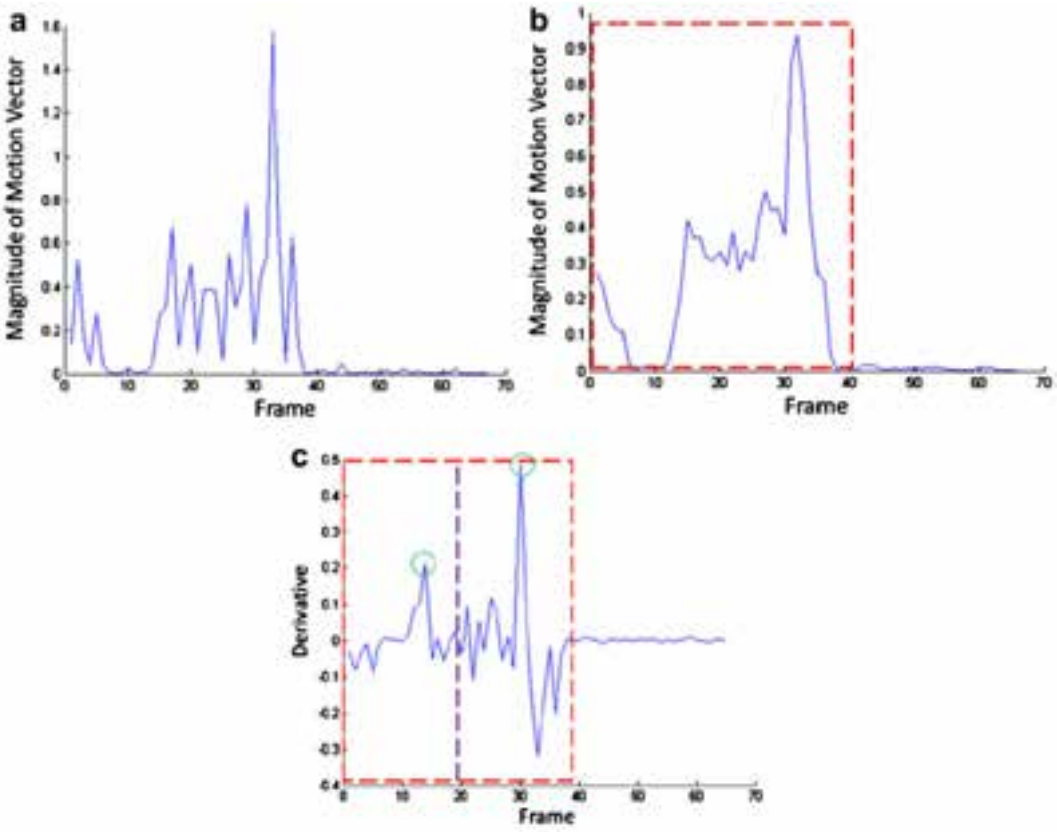

Fig. 5. Detection of a large recent change in voxel person's speed. (a) Motion vector magnitudes are computed, (b) a fixed size window, placed directly before the start of the summarization, is smoothed with a mean filter, and (c) the maximum of the derivative of the filtered motion vector magnitudes is found in the first and second halves of the window. The feature is the ratio of the two maximum values.

Table 1
Fuzzy rules for activity analysis

Rule		On the ground	Time duration	Change in speed	Motion	Oscillating		Fall
1	If	High	Long				Then	High
2	If	High	Moderate	High			Then	High
3	If	High	Moderate		High		Then	High
4	If	High	Moderate		Low		Then	High
5	If	High	Moderate			High	Then	High
6	If	High	Short	High			Then	Medium
7	If	High	Short			High	Then	Medium
8	If	High	Short			Medium	Then	Medium
9	If	Medium	Moderate	High	Low		Then	High
10	If	Medium	Moderate		Low		Then	High
11	If	Medium	Short	High			Then	Medium
12	If	Medium	Short			High	Then	High
13	If	Medium	Short			Medium	Then	Medium

captured in the Computational Intelligence Laboratory at the University of Missouri. We do not have any elderly fall data and cannot acquire any because of the age of the individuals and the risk of injury. Because of this, fall data is captured in our lab using students as subjects. As mentioned above, movies illustrating the following sequences and our processing of them can be found at http://cirl.missouri.edu/fallrecognition.

A total of 19 fall sequences were studied. The majority of sequences presented in this paper are the same sequences used in our voxel person construction and fuzzy logic for state classification paper [18]. These two bodies of work are interconnected, so processing the same data advances the previous paper and answers any questions regarding the exact information acquired and processed below. The camera capture rate was three frames per second and a total of 6713 frames, approximately 37 min, were captured. The subjects walked around the room, stood still, knelt down, fell, sometimes went slow to the ground and stretched, and sat on the couch and the chair (example images are shown in Fig. 6). Kneeling, lying on the couch, stretching on the ground, and sitting on the chair with feet on the coffee table were included to show some common activities that might appear as a fall. A variety of falls were performed including forward, backwards, and to the side. Fall scenarios also included falls that lasted for only a couple of seconds after which the person got back up, falls where the person stayed down on the ground but attempted to get back up, and falls where the person simulated a severe injury and laid on the ground motionless.

We begin by presenting approximately 11 min of video analysis (2042 frames) in Fig. 7. The fuzzy rule base outputs are shown, fall time points are manually identified to determine the success of our

Fig. 6. Example images and their corresponding silhouettes from the fall data set. Lying on the couch and sitting on the chair with feet up activities, which could be misinterpreted as a fall, are not recognized as a fall in our system, an advantage of rule-based reasoning and knowledge about three-dimensional voxel person.

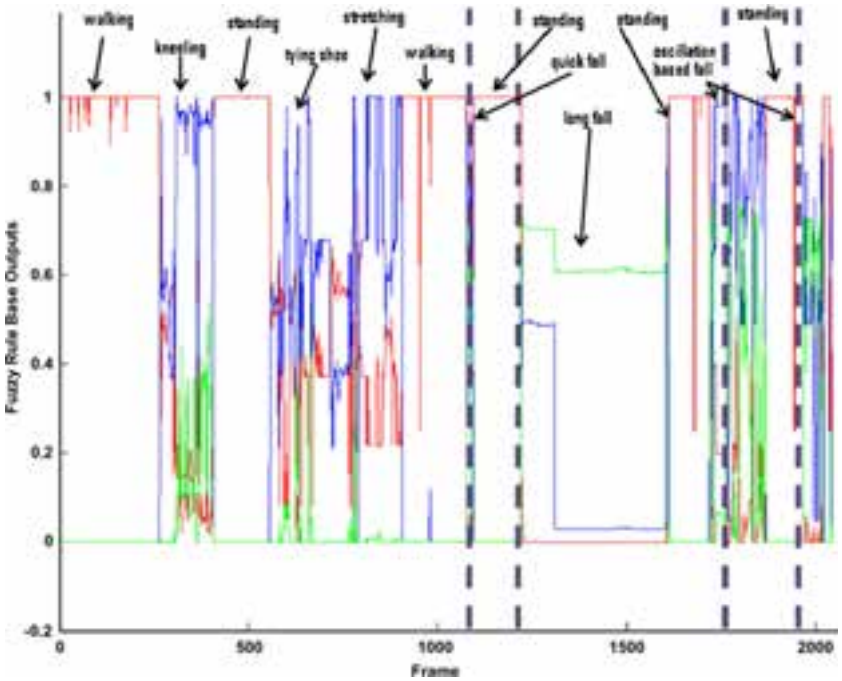

Fig. 7. Approximately 11 min of video analysis, 2042 frames total. A total of 4 falls occurred and 38 linguistic summarizations were produced. The **upright** membership is shown in red, **in-between** membership is shown in blue, and **on-the-ground** is shown in green. Dashed vertical purple lines are the manually inserted moments where a fall occurred.

automated reasoning, and fall confidences are reported. A total of 38 summaries were produced, too many to display in the paper; however, a drastic reduction from 2042 individual frame-by-frame decisions. Instead, we present the fall confidences associated with each **on-the-ground** linguistic summarization.

Fall confidence during each of the **on-the-ground** linguistic summarizations:

On-the-ground 1 (Fall 1): confidence is 0.50
On-the-ground 2 (Fall 2): confidence is 1.00
On-the-ground 3 (Fall 3): confidence is 0.50
On-the-ground 4 (Fall 3): confidence is 0.50
On-the-ground 5 (Fall 3): confidence is 0.67
On-the-ground 6 (Fall 3): confidence is 0.81
On-the-ground 7 (Fall 4): confidence is 0.50
On-the-ground 8 (Fall 4): confidence is 0.50.

The first fall in Fig. 7 is where the person fell for a short amount of time and then was able to make it to an **upright** state. Nurses have indicated that they do not want this to generate an alert, but they would like a daily report detailing the number of times that the resident was on the ground during a day, when each occurred, the fall confidences, and a movie of voxel person during that time period, or at least a few frames, to look at later. The storage of voxel person, not the original image, helps in the preservation of privacy.

Generating a fall alert involves identifying a confidence threshold τ_4. An alert is triggered if a fall confidence is greater than τ_4. A significant advantage of using fuzzy logic for the inference of activity is that can be interpreted. This makes it possible for a human to determine a location along the output domain based on a membership degree of specific fuzzy set. We use a τ_4 of 0.7, which is the consequent domain location corresponding to the left most point for an alpha cut of 0.6 in the set 'high' (e.g. high fall confidence). This method is more reliable than attempting to pick a threshold for the likelihood of a model occurring in an HMM, or a ratio of the top two most likely models, which does not necessarily tell

us if the activity was even performed. The first fall shown in Fig. 7 does not result in an alert, but it is listed in the daily activity report. Fig. 8 shows a shorter time duration performance of this type of fall. It is easier to view the behavior of the membership functions for this shorter time period.

The second fall in Fig. 7 is where the person fell for a moderate amount of time, a high sudden change in speed was detected, followed by a low amount of motion during the **on-the-ground** state. This fall is similar to the fall presented in Fig. 2, only longer in time. This fall resulted in a confidence of 1, which would trigger an alert and help would be dispatched.

The third fall in Fig. 7 involves the resident falling and unsuccessfully attempting to get back to an **upright** state. There were four **on-the-ground** summarizations produced for this third fall, and the respective fall confidences were: 0.5, 0.5, 0.67, and 0.81. As the person kept trying to get back up, the number of oscillations increased, which resulted in a greater fall confidence over time. An alert would be triggered for the fourth **on-the-ground** summarization in this case. Fig. 9 is a shorter time duration sequence that illustrates this general "trying to get back up" behavior. However, in this sequence the person never makes it far enough up to switch between the **on-the-ground** and the **in-between** states.

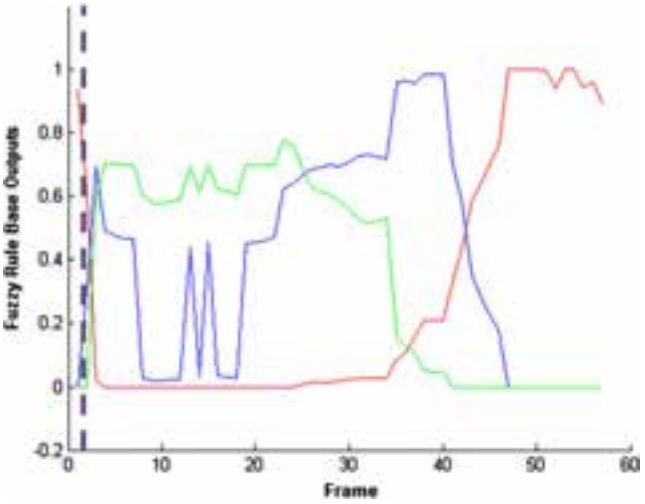

Fig. 8. Fifty-eight frames (approximately 19 s) from a sequence where the person fell and was able to get back up. Red is **upright**, blue is **in-between**, and green is **on-the-ground**.

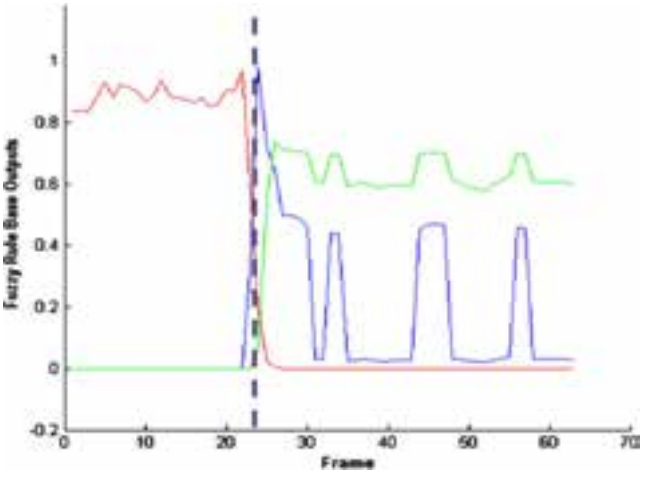

Fig. 9. Sixty-three frames (approximately 21 s) where the person fell and tried to get back up three times. Red is **upright**, blue is **in-between**, and green is **on-the-ground**.

The fourth fall in Fig. 7 is also of this "trying to get back up" type. In this particular case the person was able to make it to an **upright** state in an acceptable amount of time. The result was a confidence of 0.5, which is less than τ_4, so no alert is generated.

Table 2 is the confusion matrix for the classification of falls with respect to the **on-the-ground** summarizations for the 19 fall sequences. All falls in the video sequences were identified by a human and their start and end points were recorded.

The system correctly classified all of the actual falls and of the non-fall activities were incorrectly classified as a fall. This table only measures classification rates given an **on-the-ground** summary. There were no situations in which a fall occurred and an **on-the-ground** summary was not produced. This is an advantage of our approach, the identification of moments in which to evaluate the confidence in a fall. Non-fall activity classification rates would be much better if all linguistic summarizations, not just **on-the-ground**, were included.

The first false alarm is a situation in which the subject was on the ground and exercising for a moderate amount of time (a result of the firing of rule 3). Exercising was not factored into the design of this baseline system. Many of the rules will trigger an alert if an individual is on the ground exercising, performing leg lifts for example. However, if a resident is known to be active, the rules triggering a fall can be removed, or if they exercise at a predetermined time of day the rules can be disabled for that period. Additional pose information and features from voxel person can be extracted and rules can be added to take into account exercise for active seniors. The flexibility of the design of the features and rules allows us to minimize false alarms if we know the routines and physical capabilities of a resident.

The second false alarm is a product of the fuzzy sets and rules, referring to the fact that they were hand designed and not learned from training data. There was one case in which the subject went to the ground, resulting in the detection of a quick change in speed, a high confidence in **on-the-ground** was inferred, but a very low confidence in the moderate set was observed. Only rule 2 had an antecedent strength higher than zero. The result is a very low rule antecedent strength firing, because of the very low confidence in the moderate set, resulting in a low activation of the high confidence fuzzy set. However, the activation of only the high fuzzy set to any degree results in a defuzzified value above the threshold τ_4. The problem is that there was no other rules fired and the rule antecedent strength was not taken into consideration. To address this, we are adopting different consequent membership functions that better take into account the rule antecedent firing strength, referring to where the centroid is calculated with respect to the rule firing strength, such as the non-linear spline-based z-shaped and s-shaped membership functions. A z-shaped membership function is defined as

$$\mu_A(x) = \begin{cases} 1 & x \leqslant a \\ 1 - 2\left(\frac{x-a}{b-a}\right)^2 & a \leqslant x \leqslant \frac{a+b}{2} \\ 2\left(\frac{b-x}{b-a}\right)^2 & \frac{a+b}{2} \leqslant x \leqslant b \\ 0 & x \geqslant b \end{cases},$$

Table 2
Confusion matrix for classification of falls with respect to **on-the-ground** summarizations

		Ground truth for falls	
		Fall activity (14 falls)	Non-fall activity (32 non-falls)
Systems fall decision (fall confidence > τ_4)	Fall activity	$\frac{14}{14} = 1$	$\frac{2}{32} = 0.0625$
	Non-fall activity	$\frac{0}{14} = 0$	$\frac{30}{32} = 0.9375$

while an s-shaped function is just a z-shaped function in the opposite direction. These function names are based on their shapes, hence why the s-shaped function is the z-shaped function going in the opposite direction. The parameters $\{a, b\}$ and the sampling domain interval can be selected to shift the centroid towards 1 for the 'high' set when the rule antecedent firing strength is high, and shift the centroid towards 0.5 when the rule antecedent firing strength is low.

6. Conclusions

We have demonstrated a flexible framework for detecting human activity, in particular, falls, with a focus on older adults. This approach results in human understandable information and confidences regarding activities for the sake of monitoring the "well-being" of a resident. Silhouettes from multiple cameras are used to build a three-dimensional approximation of the human, i.e. voxel person. Features are extracted from voxel person and used along with fuzzy inference to determine the state of the resident. The resulting fuzzy rule base outputs are then temporally processed and used to generate temporal linguistic summarizations. Features from these linguistic summarizations are the input to another fuzzy inference system for reasoning about human activity. Nurse gerontology experts assisted in the design of the rules for fall detection. It is the nurses' experience that helps us relate this work to the ways in which older adults fall.

7. Future work

One extension to this work involves extracting richer linguistic summarizations from the fuzzy rule base outputs. There is a fair amount of information that our present approach discards, only accepting moments in which the maximum membership is clearly distinguishable. These moments might prove to be meaningful for a better assessment of the "well-being" of a resident. Many of the system parameters used in this work are based on empirical observations. It is important that we use training data in the future to determine the fuzzy sets, fuzzy rules, and thresholds. This will require a database of activity captured from the elderly and assistance in interpreting the data by nurses and other caregivers. We have just captured a larger dataset of falls using stunt actors. To make sure that the actors performed falls in a similar fashion to the way that elders fall, nurses coached the stunt actors.

In addition, the **in-between** state proposed in this work is rather broad. It is used in this context for detecting falls, but we plan on showing the extendable nature of this framework by the addition of more rules for state classification and more rules for activity monitoring. The detection of falls is a form of short-term monitoring, but the work presented here is in no way limited to short-term activity recognition. Hours, days, weeks, and even months worth of data will be collected and summarized based on the work presented.

Nurses at the University of Missouri assisted us in determining the rules for fall detection, but there is still much work to be done. We are currently outlining additional common types of falls for the elderly, and the states, fuzzy sets, and rules are being expanded in order to detect these various types of falls. For example, nurses express that many older adults fall after sitting for a period of time, getting up to their feet and being light headed or losing their balance. In addition, we are looking at common causes for false alarms. The rule set will be modified to take these cases into consideration.

Acknowledgments

Derek Anderson and Robert Luke are pre-doctoral biomedical informatics research fellows funded by the National Library of Medicine (T15 LM07089). This work is also supported by the National Science Foundation (ITR award llS-0428420) and the Administration on Aging (90AM3013).

References

[1] C. Stauffer, W.E.L. Grimson, Learning patterns of activity using real-time tracking, in: Proceedings of the IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, 2000, pp. 747–757.
[2] N.M. Oliver, B. Rosario, A.P. Pentland, A Bayesian computer vision system for modeling human interactions, in: Proceedings of the IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, 2000, pp. 831–843.
[3] K. Toyama, J. Krumm, B. Brumitt, B. Meyers, Wallflower: principles and practice of background maintenance, in: Proceedings of the Seventh IEEE International Conference on Computer Vision, vol. 1, 1999, pp. 255–261.
[4] R.H. Luke, D. Anderson, J.M. Keller, M. Skubic, Moving object segmentation from video using fused color and texture features in indoor environments, Journal of Real-Time Image Processing, Submitted for publication.
[5] T. Parag, A. Elgammal, A. Mittal, A framework for feature selection for background subtraction, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, 2006, pp. 1916–1923.
[6] S. McKenna, S. Jabri, Z. Duric, H. Wechsler, Z. Rosenfeld, Tracking groups of people, in: Proceedings of the Computer Vision and Image Understanding, vol. 9, 2000, pp. 42–56.
[7] I. Haritaoglu, D. Harwood, L.S. Davis, W4: real-time surveillance of people and their activities, in: Proceedings of the IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, 2000, pp. 809–830.
[8] N. Ohta, A statistical approach to background suppression for surveillance systems, in: Proceedings of the IEEE International Conference on Computer Vision, vol. 25, 2001, pp. 481–486.
[9] L. Wang, T. Tieniu, H. Ning, W. Hu, Silhouette analysis-based gait recognition for human identification, in: Proceedings of the IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 25, 2003, pp. 1505–1518.
[10] L. Dar-Shyang, Effective Gaussian mixture learning for video background subtraction, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (2005) 827–832.
[11] W.L. Buntine, Operations for learning with graphical models, Journal of Artificial Intelligence Research, vol. 2, 1994, pp. 159–225.
[12] K. Murphy, Dynamic Bayesian networks: representation, inference and learning, PhD thesis, Department of Computer Science, UC Berkeley, 2002.
[13] D. Anderson, J.M. Keller, M. Skubic, X. Chen, H. Zhihai, Recognizing falls from silhouettes, in: Proceedings of the IEEE 2006 International Conference of the Engineering in Medicine and Biology Society, vol. 1, 2006, pp. 6388–6391.
[14] N. Thome, S. Miguet, A HHMM-based approach for robust fall detection, in: Proceedings of the Ninth International Conference on Control, Automation, Robotics and Vision, vol. 5, 2006, pp. 1–8.
[15] M. Brand, V. Kettnaker, Discovery and segmentation of activities in video, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (2000) 844–851.
[16] M. Brand, N. Oliver, A. Pentland, Coupled hidden Markov models for complex action recognition, in: Proceedings of the IEEE Computer Vision and Pattern Recognition, vol. 23, 1997, pp. 994–999.
[17] T. Martin, B. Majeed, L. Beum-Seuk, N. Clarke, Fuzzy ambient intelligence for next generation telecare, in: Proceedings of the IEEE International Conference on Fuzzy Systems, vol. 15, 2006, pp. 894–901.
[18] D. Anderson, R.H. Luke, J.M. Keller, M. Skubic, Modeling human activity from voxel person using fuzzy logic, IEEE Transactions on Fuzzy Systems, accepted for publication.
[19] L. Zadeh, Fuzzy sets, Information Control 8 (1965) 338–353.
[20] L.A. Zadeh, Outline of a new approach to the analysis of complex systems and decision processes, in: Proceedings of the IEEE Transactions on System, Man, and Cybernetics, vol. 3, 1973, pp. 28–44.
[21] J.C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, New York, 1981.
[22] M. Sugeno, Fuzzy measures and fuzzy integrals – a survey, Fuzzy Automata and Decision Processes (1977) 89–102.
[23] J.M. Keller, X. Wang, A fuzzy rule-based approach to scene description involving spatial relationships, Computer Vision and Image Understanding 80 (2000) 21–41.
[24] E.H. Mamdani, S. Assilian, An experiment in linguistic synthesis with a fuzzy logic controller, International Journal of Man–Machine Studies 7 (1975) 1–13.
[25] L.D. Wilcox, M.A. Bush, Training and Search Algorithms for an Interactive Wordspotting System, in: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, vol. 2, John Wiley & Sons, 1991, pp. 12–49.