Extension of a Soft-Computing Framework for Activity Analysis from Linguistic Summarizations of Video

Derek Anderson, Robert H. Luke, James M. Keller, and Marjorie Skubic

***Abstract*—Video cameras are a relatively low-cost, rich source of information that can be used for "well-being" assessment and abnormal event detection for the goal of allowing elders to live longer and healthier independent lives. We previously reported a soft-computing fall detection system, based on two levels from a hierarchy of fuzzy inference using linguistic summarizations of activity acquired temporally from a three dimensional voxel representation of the human found by back projecting silhouettes acquired from multiple cameras. This framework is extremely flexible and rules can be modified, added, or removed, allowing for per-resident customization based on knowledge about their cognitive and physical ability. In this paper, we show the flexibility of our activity analysis framework by extending it to additional common elderly activities and contextual awareness is added for reasoning based on location or static objects in the apartment.**

I. Introduction

Our ultimate goal is the continuous monitoring of activity for assessment of the "well-being" of a resident, and in particular, the detection of abnormal or dangerous events, such as falls. It is important that technologies are developed that can recognize various activities and do so in a non-invasive fashion. In order to preserve the privacy of the residents being monitored, segmentation of the human from an image results in a silhouette, which is a binary map that distinguishes the individual from the background. These silhouettes are used to track the activity of an individual.

Video-based eldercare monitoring is only one component in a large interdisciplinary collaboration between Engineers, Nurses, and other Health Care individuals at the University of Missouri-Columbia [1,2,3]. This collaboration is unique because the technologies are being deployed at an "aging in place" facility of residential apartments called TigerPlace [4]. Passive monitoring systems are being deployed in the homes of older adults and surveys are being conducted to not only test the effectiveness of the tools and processes, but the realistic integration of technology into residents' lives. Some of the non-video sensors include: binary indications of motion, activity and appliances used in the kitchen, and bed sensors for restlessness analysis. Focus groups indicate that residents are willing to consider silhouette-based images for abnormal event detection such as falls [5].

Martin et al. [6] presented a soft-computing approach to monitoring the "well-being" of elders over long time periods from non-video sensors. Procedures for interpreting firings from these sensors into fuzzy summaries were presented. These summaries assist in characterizing a resident's trends and aid in answering queries about deviations from patterns, such as "has the occupant's sleep pattern changed significantly in the past few months".

We previously conducted preliminary work in the area of fall detection using Hidden Markov Models (HMM) [7]. Preliminary results were encouraging, but the resulting likelihood values are difficult to robustly interpret and use to reject false alarms. Thome and Miguet demonstrated a fall detection procedure that uses Hierarchical Hidden Markov Models (HHMM) [8], which has the same likelihood-based decision-making limitations as our HMM work. They used image rectification to derive approximate 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.

Johnson and Sixsmith [9] used an infrared array technology to acquire a low resolution thermal image of the resident and they track the human using an elliptical-contour gradient-tracking scheme. Falls were detected using a neural network that took the vertical velocity of the person as input. Their fall classification results were poor, only capturing around one-third of all falls. However, no non-fall scenarios resulted in a fall alarm.

Town showed a video surveillance procedure based on augmenting a ground truth ontology with image based object detection and tracking for activity analysis [10]. Annotated training sequences were used to train the structure and parameters of a Bayesian network. High level events were inferred through the use of low level tracking of image blobs, and the syntactic and semantic constraints of the ontology of states, roles, situations and scenarios were used to constrain the training of the network.

The remainder of this paper is organized as follows. Voxel person construction and state based reasoning is presented in section 2. Section 3 shows how linguistic summarizations are produced and reasoning about activity is performed. The fall recognition system is extended in section 4, new features are identified, spatial contextual awareness is added, and new activities are recognized. Experiments and results are reported in section 5.

Manuscript received December 13, 2007. This work was supported in part by the National Science Foundation (ITR award llS-0428420) and the Administration on Aging (90AM3013).

D. Anderson is with the University of Missouri-Columbia, Columbia, MO 65211 USA (phone: 573-882-6387; fax: 573-882-0397; e-mail: dtaxtd@mizzou.edu).

R. H. Luke, J. M. Keller, and M. Skubic are with the University of Missouri-Columbia, Columbia, MO 65211 USA (e-mails: rhl3db@mizzou.edu, kellerj@missouri.edu, skubicm@missouri.edu).

978-1-4244-1819-0/08/$25.00©2008 IEEE

II. FUZZY LOGIC FOR STATE CLASSIFICATION

Our immediate goal is the detection of elderly falls, a relatively short-time activity, but this activity analysis framework is being developed for higher level reasoning about the resident's "well-being" over longer periods, such as days, weeks, and even months. This starts with silhouette segmentation, which is a classification problem. The question is whether a location in the image belongs to a known background or if it belongs to the person. We assume the camera is stationary and we build a model of the background without the person in it. Figure 1 shows our silhouette segmentation procedure.

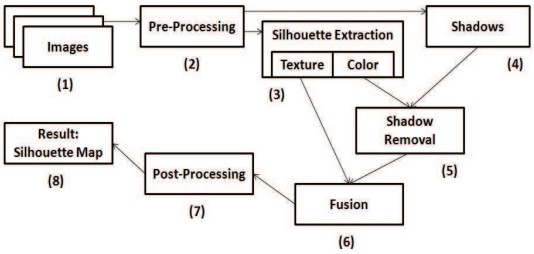

Fig. 1: Silhouette Segmentation Procedure

Images are (1) captured from a video camera, (2) pre-processed, (3) color and texture features are extracted and used for foreground segmentation, (4) shadows are identified, (5) shadows are removed from the color component of silhouette extraction, (6) fusion of texture silhouettes and shadow removed color based silhouettes is performed, and (7) morphology is used to acquire the (8) final silhouette. Details about each specific stage can be found in [11].

In [12] we presented a robust method for the construction of a three dimensional object, specifically a human called voxel person, from the back projection of multiple silhouettes. The environment is first partitioned into discrete regions, typically cubes, called volume elements (voxels). Each camera builds a list of voxels that intersect with its viewing region, and the pixel that the voxel is viewable from is recorded. Corresponding silhouettes, those with the closest time stamps, are acquired from multiple cameras and for each camera a new list is constructed, which is the union of voxels for foreground pixels in the silhouette. The next step is the intersection of these new voxel lists, which results in voxel person. The procedure is summarized in Figure 2.

Figure 3 shows voxel person viewed within our OpenGL-based visualization tool. This interactive environment helps us gain a better understanding about what is happening in voxel space and it can assist health care individuals by showing them the activities in a space that they can move around in and inquire about the activity performed, such as assessing the type and severity of a fall. Raw images are projected onto three dimensional surfaces through projective texturing, however, the silhouette can be used to mask out the individual during projection, or the background model, without the person, can be projected.

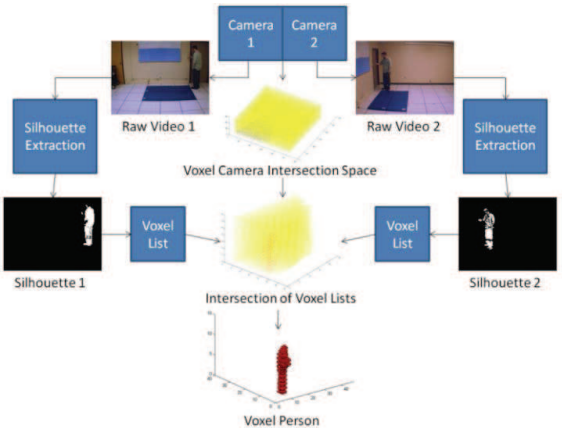

Fig. 2: Voxel person construction. Cameras capture the raw video, silhouette extraction is performed for each camera, lists of voxels that each camera can see are constructed, these voxel lists are intersected to compute voxel person. This procedure, and a subsequent morphological reconstruction process, creates a good approximation to the shape and size of voxel person, even though the silhouette may have segmentation errors, as shown above.

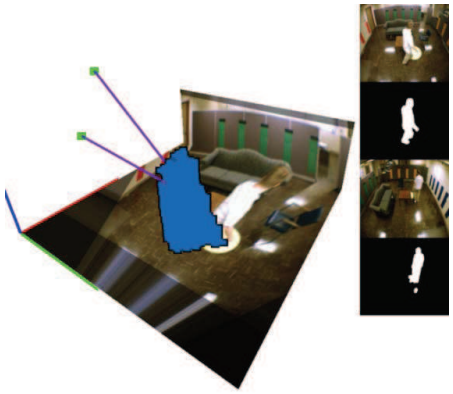

Fig. 3: Raw video feed, silhouettes captured from two cameras, and the reconstruction of the individual in voxel space. Blue pixels are voxel person, purple lines are the camera view vectors, and green indicates the placement of the cameras in the scene. Each voxel dimension is 5 inches and only half of each image dimension was processed (320x240 pixels). The corresponding object is still detailed enough to track the type of complete body activities referenced in this work.

One of the motivating factors for moving from a two dimensional to a three dimensional representation for tracking silhouettes involves the ability to model the environment. Knowledge about the world allows for the identification of voxels that correspond to walls, floor, ceiling, or other static objects or surfaces. These voxels are removed because they do not provide any significant contribution to voxel person's shape. By removing them, we effectively remove many areas upon which shadows are projected and segmentation errors occur due to reflective surfaces and a dynamic environment.

As mentioned above, fall recognition involves two levels of fuzzy logic. Fuzzy set theory, introduced by Lotfi A.

Zadeh in 1965, is an extension of classical set theory [13]. One of the more well known branches of fuzzy set theory is fuzzy logic, introduced by Zadeh in 1973 [14]. 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.

The first stage in monitoring human activity from video involves acquiring confidences in the states of voxel person, a frame by frame decision process [12]. The next stage is linguistically summarizing this information and recognizing activities [15]. This second stage uses domain expert knowledge regarding 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 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. Our rules can evaluate as many linguistic summarizations as necessary, looking as far back in time as desired, making it possible to enforce longer-term specific performances of activities. Figure 4 is our activity recognition framework.

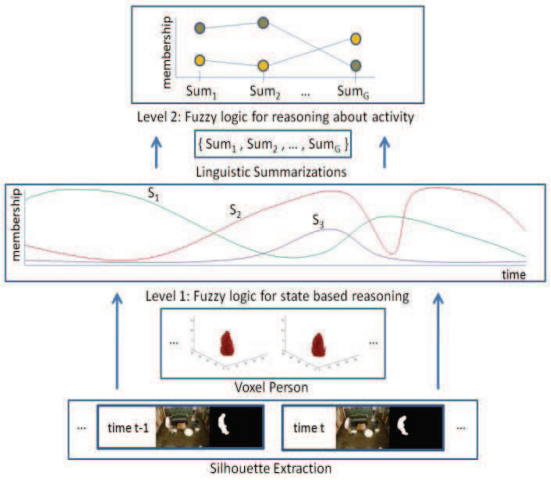

Fig. 4: Activity recognition framework, which utilizes a hierarchy of fuzzy logic based on voxel person. The first level is reasoning about the state of the individual. Linguistic summarizations are produced and fuzzy logic is used again to reason about human activity.

In [12] we presented a fuzzy logic system for determining the membership degrees of voxel person according to a collection of pre-determined states at each moment. Features are extracted from voxel person and are used to determine his or her current state. 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 does not imply a fall. Additional temporal processing is necessary to determine this. 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 low amounts of each feature characterizes the states above. There is no crisp point where the features change between states. These factors are what lead us to use fuzzy inference to classify voxel person's present membership in each state, and ultimately, to recognize human activity.

The fuzzy sets and rules used for fall recognition were designed by our engineering and nursing team, but they could be learned from appropriate training sequences. There are a total of 24 rules at the moment for determining the state of voxel person from frame to frame [12]. The result is a set of temporal confidence curves that encode the resident's activity. Figure 5 shows the plot of the state memberships over time for a fall scenario.

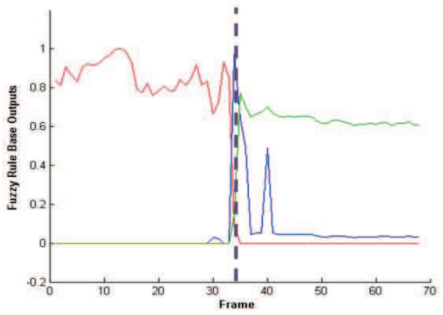

Fig. 5: Fuzzy inference outputs plotted for a voxel person fall. The x-axis is time, measured in frames, and the y-axis is the fuzzy inference outputs. The red curve is **upright**, the blue curve is **in-between**, and the green curve is **on-the-ground**. The frame rate was 3 per second, so the above plot is approximately 23 seconds of activity.

III. Fall Detection from Linguistic Summarizations

Decisions regarding the current activity can be made at each time step based on the **upright**, **in-between**, and **on-the-ground** membership values, but the result is too much information, requiring a human expert to interpret. Our goal is to linguistically summarize the temporal activity of voxel person. The objective is to take seconds, minutes, hours, and even days of resident activity and produce temporal linguistic summarizations, such as "the resident has fallen in the living room for a long time" or "the resident made and ate lunch shortly after noon". This is a situation in which less detail is more meaningful. Reporting activity for every frame results in information overload. Linguistic summarization is designed to increase the understanding of the system output, reporting a reduced set of conditions that characterizes a time interval, and temporally describes the duration that voxel person was in a state or performed a particular activity. The linguistic summarizations of voxel person's activity can help in informing nurses, residents, residents' families, and other approved individuals about the general welfare of the resident, as well as assist in an automated or manual form of determining potential cognitive or functional decline.

Linguistic summarization from video, outlined in [15], is the generation of meaningful human understandable information of the form

$$X_c \text{ is } S_i \text{ in } P_k \text{ for } T_j.$$

The object of interest, voxel person here, is denoted as X_c ($0 < c \leq C$, 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 three dimensional voxel objects. A state of voxel person is S_i. Important world locations, P_k ($0 < k \leq K$, where K is the number of locations), 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 ($0 < j \leq J$, 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 location in the apartment, L_i, can be crisp or fuzzy (we use a crisp L_i).

An example linguistic summarization is "Derek is **on-the-ground** in the living room for a moderate amount of time". The individual's name, Derek, is included in the linguistic summarization. This personalizes the summarizations, which increases readability for an end user or health care individual interested in analyzing the activity of the resident.

In order to compute linguistic summarizations from video, the fuzzy inference outputs for voxel person's state are temporally segmented. This involves identifying the state with the maximum membership value at each step, and collapsing consecutive identical state labels together for reduction, which results is a collection of intervals. Intervals that only lasted a second or two are removed because we are tracking elderly activity and are not looking for high-frequency information, such as a fall that lasted a fraction of a second. The linguistic summarizations are produced for each segmented interval by looking at information such as: the average state membership value for an interval, the location in the apartment, and the interval length is converted into a fuzzy set defined over the time domain. An example of this procedure for approximately 11 minutes of video is shown in Figure 6.

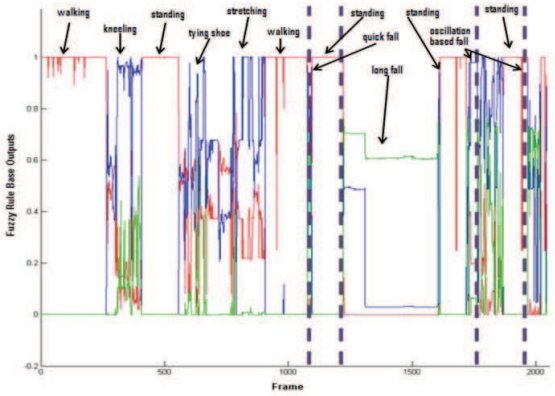

Fig. 6: Approximately 11 minutes of video analysis, 2,042 frames total, in which the person performed various activities, including: walking, standing, kneeling, tying shoes, stretching, and fallen. A total of 4 falls occurred and 38 linguistic summarizations were produced. The **upright** membership 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.

Features are extracted from the linguistic summarizations and a second level of fuzzy inference is used to recognize activities. The features extracted from the linguistic summarizations include: average **on-the-ground** membership, duration spent **on-the-ground**, quick change in acceleration, motion during **on-the-ground**, and the confidence in recent oscillating behavior between the **on-the-ground** and **in-between** states. In [15], we used 13 rules to build a confidence in a single activity, **fall**, which is evaluated whenever an **on-the-ground** summarization is found. A decision regarding a fall is made by picking a confidence threshold. The linguistic variable fall has the terms low, medium, and high, and we discovered that a value of 0.7, mostly in the high term, worked well.

IV. Activity Recognition System Extension

Two levels of fuzzy logic perform well for fall detection, but the question remains: does the framework extend well to different activities? The following section shows that this framework is indeed extendable, and it extends rather easily. In this section we show new features extracted from voxel person and show that knowledge about spatial context improves the ability to infer new activities based on awareness of the resident's position in the apartment and/or static object being used. The new activities being recognized, based on what elders perform regularly on a daily basis, include: **standing**, **walking**, **motionless-on-the-**

chair, and **lying-motionless-on-the-couch**. The new states required to recognize these activities include:

On-the-chair (S_4): This state is characterized by voxel person being on the chair. Activities that involve this state are, for example, sitting on the chair and/or lying on the chair.

On-the-couch (S_5): This state is more specific than **on-the-chair**. It is generally characterized by voxel person being on the couch, having a low similarity with the ground plane normal, a high centroid height, and a high minimum height.

These two states are different from the previous three because they are based on voxel person interacting with a static object in the scene. In order to recognize these states we need new voxel person features. Suppose that the set of voxels belonging to the subject at time t is $\boldsymbol{V'_t} = \{\boldsymbol{V_{t,1}}, \ldots, \boldsymbol{V_{t,P}}\}$, where the j[th] voxel is $\boldsymbol{V_{t,j}} = \langle x_{t,j}, y_{t,j}, z_{t,j}\rangle^{\boldsymbol{t}}$ and P is the number of voxels in voxel person at time t. The centroid of voxel person at time t is

$$\boldsymbol{\mu_t} = \left(\frac{1}{P}\right)\sum_{j=1}^{P} \boldsymbol{V_{t,j}}.$$

An eigen-based descriptor is used for robust identification of the minimum height of voxel person. The covariance matrix, used to find the eigen information, is

$$\sum_t = \left(\frac{1}{P-1}\right)\sum_{j=1}^{P}(\boldsymbol{V_{t,j}} - \boldsymbol{\mu_t}) * (\boldsymbol{V_{t,j}} - \boldsymbol{\mu_t})^t.$$

The eigenvectors, $\boldsymbol{eigvec_{t,k}}$, where $k = \{1,2,3\}$, are scaled by their respective eigenvalues, $eigval_{t,k}$, and are added to the voxel person centroid, i.e.

$$\boldsymbol{eigheight_{t,k}} = \boldsymbol{\mu_t} + 2\sqrt{eigval_{t,k}} * \boldsymbol{eigvec_{t,k}}.$$

The eigenvalues are sorted in decreasing order. For each eigenvector, we generate $\boldsymbol{eigheight_{t,k+3}}$, which is in the opposite direction of $\boldsymbol{eigvec_{t,k}}$, hence

$$\boldsymbol{eigheight_{t,k+3}} = \boldsymbol{\mu_t} + 2\sqrt{eigval_{t,k}} * (-1) * \boldsymbol{eigvec_{t,k}}.$$

The minimum z value, i.e. voxel person's eigen-based minimum height, from the $\boldsymbol{eigheight_{t,k}}$ set is recorded. The linguistic variable for minimum height has the terms low = [-0.15 -0.1 0.1 0.15], medium = [0.1 0.15 0.25 0.3], and high = [0.25 0.3 9.95 10]. The fuzzy sets in this paper are represented as trapezoidal membership functions, which are defined with respect to 4 points.

Next, we identify static objects and general regions of interest in the apartment. For example, meal preparation should occur in the kitchen by the sink and lying down watching TV will most likely occur on the couch or chair. The apartment blueprint is acquired and the map is spatially partitioned by a user. This is shown for an apartment blueprint from TigerPlace in Figure 7.

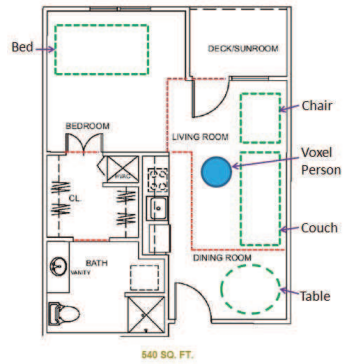

Fig. 7: TigerPlace apartment blueprint with region and object labeling. Green regions are static objects and red lines partition regions of interest.

The most discriminate plane for detecting interaction with large static regions and/or objects is the ground (x-y) plane. The region partitions, such as dining room, living room, and bedroom are used in the generation of linguistic summaries. Figure 7 shows a few important objects, such as the couch, table, chair, and bed. Voxel person is projected onto the ground plane at each frame and the amount of object overlap is calculated. Figure 8 illustrates this general procedure.

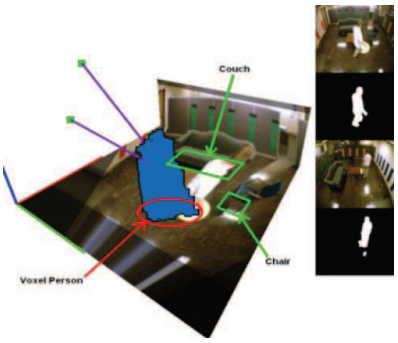

Fig. 8: Object region labeling and an illustration showing how voxel person is projected onto the ground plane for object overlap computation.

The amount of voxel person-object overlap, for object $\boldsymbol{O_i}$, where $\boldsymbol{O_i} = \langle x_i, y_i, z_i\rangle^{\boldsymbol{t}}$, is

$$\left(\frac{1}{\mathrm{P}}\right)\sum_{\mathrm{j=1}}^{\mathrm{P}} \mathrm{CONTAINED}(\mathbf{V_{t,j}}, \mathbf{O_i}),$$

where CONTAINED is a binary check for containment of voxel $\boldsymbol{V_{t,j}}$ within the object $\boldsymbol{O_i}$. The linguistic variable containership has the terms low = [-0.2 -0.1 0.1 0.2], medium = [0.1 0.2 0.4 0.5], and high = [0.4 0.5 1.5 1.6].

Six new rules for identifying **on-the-chair** and **on-the-couch** are shown in Table 1. Antecedent 1 (A1) is voxel person overlap with the chair, A2 is voxel person overlap with the couch, A3 is the eigen-based minimum height, A4 is the similarity between voxel person's primary orientation and the ground plane normal (a feature for distinguishing between **upright** and **on-the-ground**), and A5 is the height of voxel persons centroid. Calculations for A4 and A5 and

their fuzzy sets are in [12]. All antecedents are coded as L=low, M=medium, and H=high in Table 1. Consequent 1 (C1) is **on-the-chair** and C2 is **on-the-couch**. The two consequents have 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].

Table 1. New fuzzy rules for state recognition

Rule		A1	A2	A3	A4	A5		C1	C2
1		H						H	
2		M						M	
3	IF	L					Then	L	
4			H	H	L	H			H
5			M	H	L	H			M
6			L	H	L	H			L

The next step involves recognizing new activities from linguistic summarizations produced based on these new frame-by-frame voxel person state classifications. As mentioned above, the new elderly activities being recognized are **standing**, **walking**, **motionless-on-the-chair**, and **lying-motionless-on-the-couch**. Each variable has the terms very low = [-0.5 0 0 0.5], low = [0 0.25 0.25 0.5], medium = [0 0.5 0.5 1], and high = [0.5 1 1 1.5]. Rules for recognizing the new activities are shown in Table 2. Antecedent 1 (A1) is the time duration of the linguistic summarization in seconds, and the terms are brief = [-1 1 1 2], short = [1 5 10 15], moderate = [10 120 480 720], and long = [480 900 86400 86400] (represented in Table 2 as B, S, M2, and L2 respectively). A2 is motion during the summary, which is the average magnitude of the motion vectors observed (calculation in [15]), and the terms are low = [-0.2 0 0 0.2] and high = [0.1 0.4 100 102]. A3 is **upright**, A4 is **on-the-couch**, A5 is **on-the-chair**, C1 is **standing**, C2 is **walking**, C3 is **lying-motionless-on-the-couch**, and C4 is **motionless-on-the-chair**. As mentioned above, the consequents have the terms low, medium, and high.

Table 2. New fuzzy rules for activity recognition

Rule		A1	A2	A3	A4	A5		C1	C2	C3	C4
1		S	L	H				H			
2		S	L	M				M			
3		M2	L	H				H			
4		S	H	H					H		
5	If	S	H	M			Then		M		
6		M2	H	H					H		
7		S	L		H					M	
8		M2	L		H					H	
9		S	L			H					M
10		M2	L			H					H

Standing is generally identified by voxel person being in an **upright** state and having low motion. **Walking** is identified by voxel person having an **upright** state, having a high motion (we only use low and high for describing motion currently). **Motionless-on-the-chair** is identified by voxel person being mostly in the **on-the-chair** state and having low motion. **Lying-motionless-on-the-couch** is identified by voxel person being in the **on-the-couch** and having low motion. The confidences of these events are based on the time duration of the respective linguistic summarization.

V. RESULTS

All data was captured in the Computational Intelligence Laboratory at the University of Missouri-Columbia. 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. Data sets and movies illustrating our processing of these sequences will be posted at http://cirl.missouri.edu/fallrecognition. Nineteen sequences were analyzed, the camera capture rate was 3 fps, and a total of 3352 frames were captured (approximately 18 minutes).

Many of the sequences analyzed here were also processed in our previous fall detection papers [12][15]. This is done to show that new activities can be recognized, old activities are still correctly recognized, and false alarms are few if any per-activity type. The person varied walking and standing, possibly kneeled and tied their shoes, fell at some point, and sat on the couch and the chair. The kneeling, lying on the couch, and sitting on the chair with feet up on a coffee table were included in order to show the discriminate power of this system and features extracted from voxel person. Falls were performed differently, meaning that sometimes the person fell forward, sometimes backwards, and also to the side. Fall scenarios also included falls that lasted for only a couple of seconds and then 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.

There were no falls in any of the new sequences. We still recognized all of the old falls, as reported before, and successfully did not call any new activity a fall, even in light of lying on the couch and sitting on the chair with ones feet on the coffee table. Figure 9 represents approximately 2 and ½ minutes of fuzzy state memberships, where all of the new activities were performed at some point in the sequence. Figure 10 shows a few frames from this sequence.

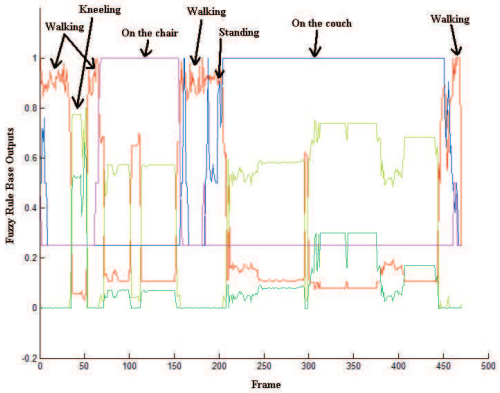

Fig. 9: State membership plots for approximately 2 and ½ minutes of video analysis, 470 frames, in which the person was walking, standing, kneeling, falls, sitting on the chair, and sitting on the couch. Annotations are provided in order to show where a human believed the different activities occurred. **Upright** is shown in red, yellow is **in-between**, green is **on-the-ground**, blue is **on-the-couch**, and purple is **on-the-chair**.

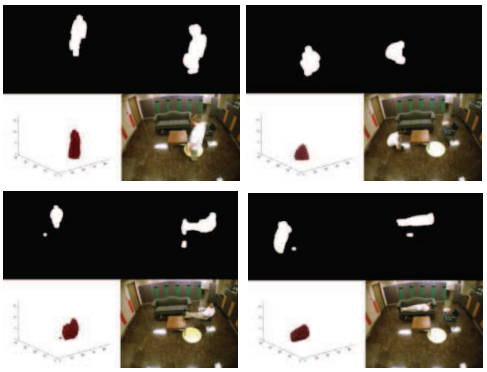

Fig. 10: Frames 22 (walking), 47 (kneeling), 120 (sitting on the chair), and 375 (lying on the couch) from the sequence shown in Figure 9.

The linguistic summarizations extracted for the sequence shown in Figure 9 are

Derek is **walking** in the lab for a moderate time
Derek is **in-between** in the lab for a short time
Derek is **walking** in the lab for a short time
Derek is **motionless-on-the-chair** in the lab for a moderate time
Derek is **walking** in the lab for a short time
Derek is **standing** in the lab for a short time
Derek is **lying-motionless-on-the-couch** in the lab for a moderate time
Derek is **walking** in the lab for a short time.

These summaries match the annotated video sequence. Recognizing all of the frames where the subject is **walking** or **standing** has turned out to be the most difficult. Pauses in walking appear to be moments of standing, making labeling a rather subjective process, and many standing and walking activities tend to be filtered out because of our check for activities that occurred too quickly (our criteria was to remove activities less than 2 seconds). The rational for removing short time summaries was to filter out potential noise and eliminate activities that elders should not be performing (high frequency activity). Figure 11 shows the subject lying on the couch, which could easily be interpreted as a fall for a procedure only using a single camera. The combination of voxel person and fuzzy rule based reasoning helps to eliminate such potential false alarms.

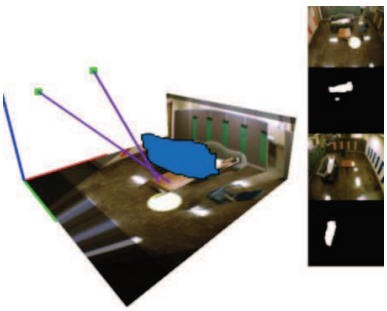

Fig. 11: Voxel person lying on the couch. Determining the subject has not fallen, but is elevated on an object performing a safe activity, is difficult given the limited information in a single two dimensional silhouette. However, this can be successfully identified using the three dimensional voxel person representation and fuzzy logic.

VI. Future Work

Many of the quantities used in this work are based on empirical observations and domain knowledge from nurses. We are looking towards using training data to determine some of the fuzzy sets and fuzzy rules. This will require a database of activity captured from the elderly and assistance in interpreting the data by nurses and other caregivers. We are currently preparing to capture a larger dataset of falls using stunt actors. To make sure that these actors perform the falls in a similar fashion to the way that elders fall, nurses will coach the stunt actors.

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. We are building a framework for computing with words so that important linguistic queries about the well-being of the resident over longer time periods can be performed using the linguistic summaries gathered from video and even linguistic summarizations from simpler non-video based household sensors

References

[1] M. Alwan et al, "Facilitating interdisciplinary design specification of "smart homes" for aging in place," in *Proc. Int. Congress of the European Federation of Medical Informatics*, 2006, pp. 45-50.

[2] G. Demiris et al, "An evaluation protocol of a smart home application for older adults," in *Proc. Int. Conference Addressing Information Technology and Communications in Health*, 2007, pp. 319-323.

[3] G. Demiris et al, "Smart home sensors for the elderly: a model for participatory formative evaluation," in *Proc. IEEE EMBS Int. Special Topic Conf. on Information Technology in Biomedicine*, 2006, pp. 1-4.

[4] M. Rantz et al, "TigerPlace, a state-academic-private project to revolutionize traditional long term care," Submitted to *Journal of Housing for the Elderly*, 2007.

[5] G. Demiris, M. Rantz, M. Aud, K. Marek, H. Tyrer, M. Skubic, and A. Hussam, "Older adults' attitudes towards and perceptions of 'smart home' technologies: a pilot study," in *Medical Inf. and the Internet in Medicine*, 2004.

[6] T. Martin, B. Majeed, L. Beum-Seuk, and N. Clarke, "Fuzzy ambient intelligence for next generation telecare," in *IEEE Int. Conf. on Fuzzy Systems*, 2006, pp. 894-901.

[7] D. Anderson et al, "Recognizing falls from silhouette," in *28th Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society*, 2006, pp. 6388–6391.

[8] N. Thome and S. Miguet, "A HHMM-based approach for robust fall detection," in *9th International Conference on Control, Automation, Robotics and Vision*, 2006.

[9] N. Johnson and A. Sixsmith, "Simbad: smart inactivity monitor using array-based detector," in *Gerontechnology*, 2002.

[10] C.P. Town. "Ontology-driven Bayesian networks for dynamic scene understanding." in *Proc. Int. Workshop on Detection and Recognition of Events in Video*, 2004.

[11] R. H. Luke et al, "Silhouette extraction, refinement and fusion from color and texture features," Under Review by *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2007.

[12] D. Anderson, R.H. Luke, J. M. Keller, M. Skubic, "Modeling Human Activity From Voxel Person Using Fuzzy Logic," Under Review by *IEEE Transactions on Fuzzy Systems*.

[13] L. Zadeh, "Fuzzy sets," *Information Control*, pp. 338-353, 1965.

[14] L. A. Zadeh, "Outline of a new approach to the analysis of complex systems and decision processes," in *IEEE Transactions on System, Man, and Cybernetics*, 1973.

[15] D. Anderson, R.H. Luke, J. M. Keller, M. Skubic, "Linguistic Summarization of Activities from Video for Fall Detection Using Voxel Person and Fuzzy Logic," Under Review by *Computer Vision and Image Understanding*.