Modeling Human Activity From Voxel Person Using Fuzzy Logic

*Derek Anderson, *Robert H. Luke, *James M. Keller, *Marjorie Skubic, #Marilyn Rantz, and #Myra Aud

*Department of Electrical and Computer Engineering
#Sinclair School of Nursing
University of Missouri, Columbia, MO, 65211

dtaxtd@mizzou.edu, rhl3db@mizzou.edu, kellerj@missouri.edu, skubicm@missouri.edu,
rantzm@health.missouri.edu, audm@health.missouri.edu

ABSTRACT

As part of an interdisciplinary collaboration on eldercare monitoring, a sensor suite for the home has been augmented with video cameras. Multiple cameras are used to view the same environment and the world is quantized into non-overlapping volume elements (voxels). Through the use of silhouettes, a privacy protected image representation of the human, acquired from multiple cameras, a three-dimensional representation of the human is built in real-time, called voxel person. Features are extracted from voxel person and fuzzy logic is used to reason about the membership degree of a predetermined number of states at each frame. Fuzzy logic enables human activity, which is inherently fuzzy and case based, to be reliably modeled. Membership values provide the foundation for rejecting unknown activities, something that nearly all current approaches are insufficient in doing. We discuss temporal fuzzy confidence curves for the common elderly abnormal activity of falling. The automated system is also compared to a ground truth acquired by a human. The proposed soft-computing activity analysis framework is extremely flexible. Rules can be modified, added, or removed, allowing per-resident customization based on knowledge about their cognitive and functionality ability. To the best of our knowledge, this is a new application of fuzzy logic in a novel approach to modeling and monitoring human activity, in particular the well-being of an elderly resident, from video.

1. INTRODUCTION

Human activity analysis from video is an open problem that has been studied intensely within the areas of video surveillance [1][2], homeland security [3][4], and more recently, eldercare [5][6], to name a few. 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.

The video sensors mentioned above are only one component in a larger sensor network that is intended to help elders live longer, healthier independent lives. A variety of sensors are dispersed throughout the home in order to capture information such as binary indications of motion in different areas, activity and appliances used in the kitchen, bed sensors for restlessness analysis and more. These technologies are part of a large interdisciplinary collaboration between Engineers, Nurses, and other Health Care individuals at the University of Missouri-Columbia [7][8][9]. This collaboration is unique because these technologies are being installed at the Tigerplace facility [10], which is a group of residential apartments for "aging in place". These technologies are being deployed and surveys are being conducted to not only test the effectiveness of the tools and processes, but the realistic integration of technologies into these elders' lives. Focus groups indicate that residents are willing to consider silhouette-based images for abnormal event detection such as falls [11].

2. BACKGROUND

Martin et al. presented a soft-computing approach to monitor the "well-being" of elders over long time periods using non-video sensors such as passive infrared, toggle switches, vibration, temperature, and pressure sensors [5]. Their system was installed in elderly resident's homes for periods ranging from 6 to 18 months. 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".

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 the construction of a background model and regions in subsequent images with significantly different characteristics are classified as foreground. The differencing task is usually formalized as a background subtraction procedure [2][12][13][14][15][16][17][18]. Stauffer and Grimson introduced an adaptive method for background subtraction that uses a mixture of Gaussians per pixel with a real-time online approximation to the model update [2]. Oliver et al. [12] used Principal Component Analysis (PCA) to build a background model, called an eigenbackground, which helps with handling a range of scene variation (illumination, weather, etc). However, the method is limited because no model update method was presented.

After the human is segmented, a larger problem arises regarding the higher level processing of this information for recognizing activity, as well as deviations from normal patterns of activity. In the area of short-term activity recognition, we used Hidden Markov Models (HMMs) for fall detection [6]. HMMs are the classical workhorses in the areas of symbol and speech recognition [19][20][21]. They are a popular statistical tool that has the advantage of unsupervised learning, the Baum-Welch algorithm, and temporal patterns can be recognized through statistical inference [19][22][23]. Our preliminary results indicate that HMMs can be used in some situations for fall detection using a single camera solution and features calculated from silhouettes. However, for reasons discussed in more depth in the next few sections, we have moved from tracking the person in two to three dimensions, for the generation of features that work for a wider variety of activity and environmental conditions, and fuzzy logic is used instead of statistical inference for the production of interpretable confidences.

Sixsmith and Johnson used an infrared array technology to acquire a low resolution thermal image of the resident and they then tracked the human using an elliptical-contour gradient-tracking scheme [24]. Fall detection involved using a neural network that took the vertical velocity of the subject 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. Thome and Miguet recently demonstrated a technique that used Hierarchical HMMs (HHMM) for video based fall detection [25]. The most interesting part of this research is the feature that they use in the HHMM. 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.

3. APPROACHES TO MODELING HUMAN ACTIVIY

The most widely used approaches to modeling human activity, not just for fall detection, include HMMs [6][12] and its variants (Hierarchical HMMs [25], Entropic-HMMs [26], coupled-HMMs [12][27], etc), graphical models [28], and dynamic Bayesian networks (also known as dynamic graphical models) [29]. There are several fundamental limitations with these popular yet powerful approaches which have led us in a different direction. First, Expectation-Maximization (EM), known as the Baum-Welch procedure for HMMs [19], is a well known technique for learning parameters in hidden variable models. A model with S states has S*S transition probabilities, S initial state probabilities, M mixtures for each state, and for a Gaussian with dimension D, there are D components in each mean and D*D components in each covariance matrix. This amounts to a large number of parameters to estimate for what is still a relatively small model. It is not commonly advertised and acknowledged that there is a severe problem related to the fact that most of the model parameters are only supported by a relatively small subset of the data. While research is being conducted in the area of learning model structure, such as through the use of entropy minimization [26], the majority of researchers still manually specify the model and its sparcity structure or conduct ad-hoc heuristic searches, making this procedure not as automated as most lead it on to be. All of this limits the predicative power of these learned models and their generalizability to a large set of sequences not included in the training set.

These approaches provide a way to compute the likelihood of a model given an observation sequence (the inference task). These likelihoods are useful for comparing which model is the most likely from a set, typically small, of trained models (pick one of K). However, this value is not something that can be easily interpreted as a confidence that the activity occurred, making the reliable rejection of activity not possible for most non-trivial real-world problems. This is especially the case in long observation sequences, even if scaling is applied [19].

Some elect to train a small number of models to encompass all activities that the system should ignore (i.e. not one of the K to be recognized), called garbage or filter models [30]. The learning and representation of a large number of unknown activities using a small number of models is not a theoretically sound and computationally tractable approach. In the case of speech and symbol recognition, tasks that can in many situations assume that an observation sequence was generated from one of K known models, likelihood values can be compared and the most likely model identified. However, human activity analysis does not have the flexibility to assume that an observation sequence came from one of K known models. The selection of a threshold for model acceptance is the most unreliable approach. Many elect to go with an equally sub adequate approach, acceptance based on the formulation of a ratio formed between the top two models, which still requires the selection of a ratio threshold.

What is really needed in the area of human activity analysis is not another non-interpretable likelihood value or the ad hoc training of garbage models, but a confidence value that can be understood and reliably used to reject 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 is necessary in order to address the deficiencies just mentioned and the inherent uncertainty related to modeling and inferring human activity. Fuzzy sets are used for modeling features extracted from the human, the human's state and subsequent activity, and fuzzy logic is used for inferring the state and activity from the human's features. The systems output is 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 decision making process, which is substantially different from a likelihood calculation based on the product of a potentially large number of conditional probabilities (each of which reside in the [0,1] range). A fuzzy approach also has the advantage that the rules and linguistic variables are understandable and simplify addition, removal, and modification of the knowledge structure. In this paper, we outline the first step in the reliable recognition of human activity through the recognition of the state of voxel person over time for fall recognition.

4. SILHOUETTE SEGMENTATION

Silhouette segmentation is a change detection procedure. We assume, as most, that the camera is stationary and a model of the background is built. As each new image is acquired, features are extracted and locations that have significantly changed from the background are identified. Most start at the pixel processing level, a low level computer vision task, and the results are then used by medium to high level algorithms to identify regions of interest, such as the largest connected component, and in some cases, identification of the body regions. This is not a simple task since objects move in a scene, illumination changes occur, and shadows and other phenomenon such as reflections further complicate the automated extraction procedures. Our silhouette extraction system is adaptive, incorporates both texture and color information, and performs shadow removal. Figure 1 shows our silhouette segmentation system [31].

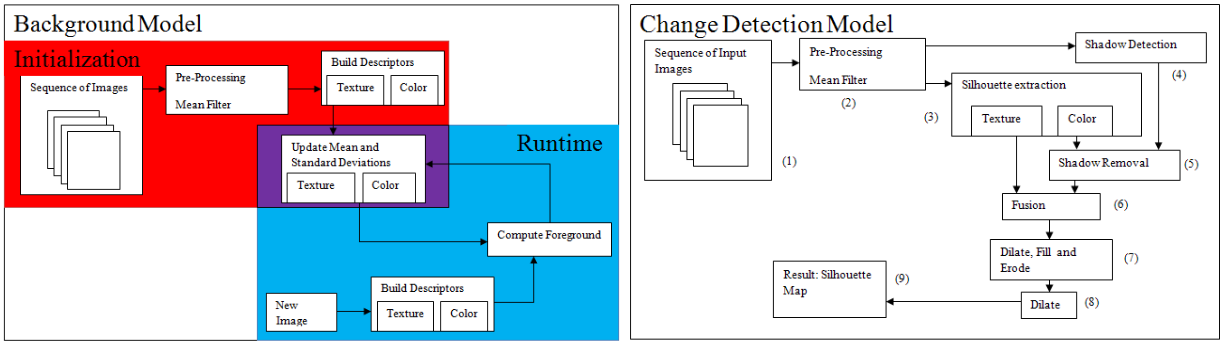


Fig. 1: Background model construction and change detection process for silhouette segmentation from a single stationary camera.

The background model is first built using a user specified number of images, which is typically somewhere around 10 frames. It is best, but not necessary because the system is adaptive, to initialize the background model using a sequence of images contain only the background and not the human. Images are pre-processed for pixel noise removal, color and texture features based on histograms of texture and color are extracted, and the mean and standard deviation of a single Gaussian is recorded for each pixel. Each new image is passed through the change detection model. The images are pre-processed, shadows are removed using a modified Hue, Saturation, and Value (HSV) color space procedure (where hue is the perceived color, saturation describes the amount of color present, and value is related to brightness), and color and texture features are extracted. Shadows are removed from the color features, which have a greater tendency to register shadows given the selected color space and feature descriptor, the texture results, which are computed using different color spaces, are fused using the Yager union, then morphological and logical operations are performed on the results to remove noise and clean up the silhouette. Morphological dilation is performed in order to expand areas of the silhouette, which assists in the creation of connected silhouettes. A fill operation takes the dilated result and makes regions that are surrounded by a connected silhouette region foreground. Lastly, these results are eroded to reduce them to a size and shape that is more like the original silhouette. Further details about each specific stage can be found in [31].

5. VOXEL PERSON

Multiple cameras viewing the same environment are crucial for the recognition of activity. Different viewpoints assist in resolving many issues due to single camera object occlusion and it makes the reconstruction 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 environment, a voxel (volume element) is an element resulting from a discretization of three-dimensional space. Voxels are typically non-overlapping cubes. The set of voxels belonging to voxel person at time t are $\boldsymbol{V'_t} = \{\boldsymbol{V_{t,1}}, \boldsymbol{V_{t,2}}, \dots, \boldsymbol{V_{t,P}}\}$, where the center position of the jth voxel at time t is $\boldsymbol{V_{t,j}} = \langle \boldsymbol{x_j}, \boldsymbol{y_j}, \boldsymbol{z_j} \rangle^{\boldsymbol{t}}$. Each image has its capture time recorded. The silhouettes for each camera that are the closest in time are used to build the current voxel person. Construction of voxel person from a single camera results in a 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. Voxel person, according to camera i $(1 \leq i \leq C)$ at time t is $\boldsymbol{V^i_t}$, whose cardinality, $|\boldsymbol{V^i_t}|$, is $\boldsymbol{P_i}$. The planar extensions of voxel person from multiple cameras, $\{\boldsymbol{V^1_t}, \ \dots \boldsymbol{V^C_t}\}$, are combined using an operation, such as intersection, $\boldsymbol{V'_t} = \wedge_{\boldsymbol{i=1}}^{\boldsymbol{C}} \boldsymbol{V^i_t}$, to assemble a more accurate object representation.

Reconstruction of three-dimensional objects, both solid representations and hulls, from two-dimensional images through back-projection is not a new concept. Object reconstruction has been studied within computer graphics, computer vision, biomedicine, and even in a variety of forms in the activity analysis domain [32][33][34][35][36]. What separates our work from most, besides the use of silhouettes for back-projection, is the way in which voxel person is used and how its shape is refined using knowledge about the environment. We use a relatively low resolution object, for computational efficiency, and it is only used to obtain features for activity analysis. The object is not explicitly tracked, segmentation of object regions is not attempted, and the goal is not to build a highly detailed surface or solid representation. Advances have been made in each respective area, but no approach to date is either real-time or mature enough to be included in a real-world system that runs unsupervised for long time periods. Not building a detailed object representation, which is difficult to obtain using silhouettes produced by a robust change detection system, further helps preserve the resident's privacy. A wide range of activities, especially in an eldercare domain, and for fall detection in particular, do not require high detail for reasoning about activity that involves the movement of the entire body.

Voxels that correspond to walls, floor, ceiling, or other static objects or surfaces are removed. This procedure eliminates areas of little interest that shadows and reflections are projected onto, which complicate feature extraction. The volume of voxel person can be approximated by summing up the number of voxels, $|\boldsymbol{V}_t'|$, or through generating the covariance matrix for voxel person and computing its determinant. In our eldercare application, each resident's profile, such as their height, width, and length, or a prototypical set of these parameters, can be modeled, and if voxel person is not within some acceptable volume range, segmentation can be assumed to have gone astray. This further helps to ensure that feature extraction is operating under more ideal conditions.

The use of two cameras oriented orthogonally with overlapping view volumes results in a silhouette intersection that defines the majority of voxel person's primary shape. The planar extension regions that the person was not occupying are removed by the intersection. The use of more than two cameras helps with further defining various parts of the body, typically extensions such as arms and feet, as well as eliminating troublesome viewing angles that result in a loss of body detail. The voxel person construction process is shown in Figure 2.

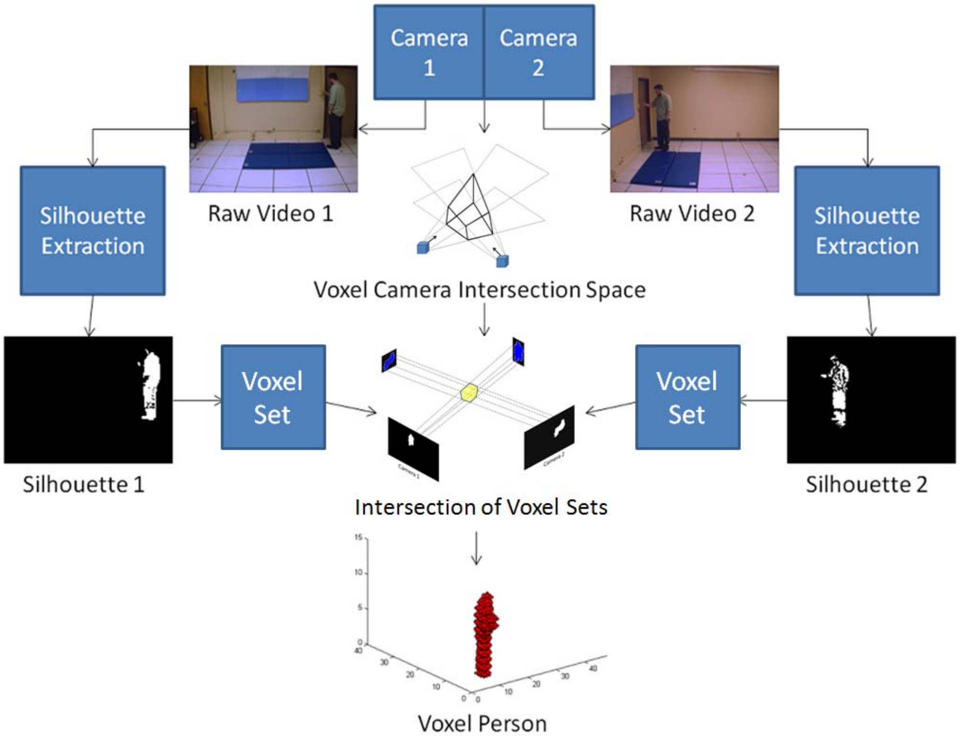
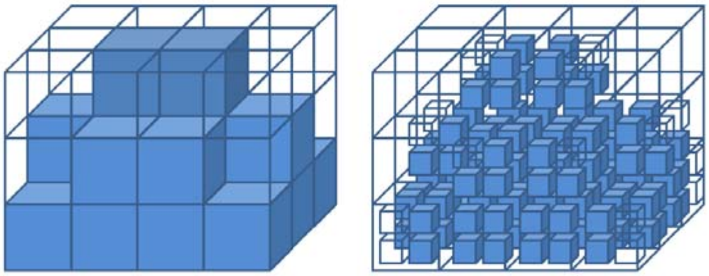


Fig. 2: Voxel person construction. Cameras capture the raw video from different viewpoints, silhouette extraction is performed for each camera, voxel sets are constructed from the silhouettes for each camera, and the voxel sets are intersected to compute voxel person.

Camera positions are recorded and intrinsic and rotation parameters are estimated [37]. For each pixel in the image plane, a set of voxels that its viewing ray intersects is identified. This is the enabling step that makes voxel person construction real-time. Voxel person construction is now just an indexing procedure. There is no need to recalculate voxel-pixel intersections while building voxel person from the silhouettes. The voxel-pixel test is a procedure that only has to be computed one time when the cameras are positioned in the room. If computational complexity of voxel-pixel is of concern, possibly because the cameras could be moved frequently or the possible voxel space is very large, spatial partitioning of voxel space, either an octree or binary spatial partition tree, can be used to speed up voxel-pixel set construction. One limitation with this approach is that it results in a fixed and usually low resolution representation of the object. However, voxels belonging to the object can be further subdivided and tested on the fly for increased object resolution (shown in Figure 3). This method is dynamic and it allows for resolution to be increased only in the areas that are needed, making an otherwise non-loadable and/or non-computable space now possible.



(a) (b)

Fig. 3: (a) Low resolution voxel space object. (b) Voxel resolution is increased in areas that are needed by subdividing only the volumes included in the low resolution representation. The subdivided voxels are then tested for intersection and a more detailed object can be created.

6. FEATURE EXTRACTION

In order to track the activity of voxel person, features must first be extracted at each time step. Our initial objective is fall detection; the features extracted are both spatial and temporal. The spatial features include voxel person's (a) centroid, (b) eigen-based height, and (c) the similarity of voxel person's primary orientation and the ground plane normal. The centroid of voxel person at time t, $\boldsymbol{\mu_t}$, is

$$\left(\frac{1}{P}\right)\sum_{j=1}^{P}\boldsymbol{V'_{t,j}}.$$

The eigen-based height is helpful for robustly identifying the activity according to voxel person's posture. The covariance matrix, used to find the eigen information, is

$$\left(\frac{1}{P-1}\right)\sum_{j=1}^{P}(\boldsymbol{V'_{t,j}}-\boldsymbol{\mu_t})*(\boldsymbol{V'_{t,j}}-\boldsymbol{\mu_t})^{t}.$$

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 maximum z value, i.e. voxel person's eigen-based height, from the $\boldsymbol{eigheight_{t,k}}$ set (total of 6 vectors) is recorded. The next feature is the similarity between voxel person's primary orientation, that is, the eigenvector with the largest corresponding eigenvalue, as well as the negative of that eigenvector, and the ground plane normal is computed,

$$gpsim_t = \max\left(\boldsymbol{eigenvec_{t,1}} \cdot \langle \mathbf{0}, \mathbf{0}, \mathbf{1} \rangle^{\boldsymbol{t}}, \boldsymbol{eigenvec_{t,4}} \cdot \langle \mathbf{0}, \mathbf{0}, \mathbf{1} \rangle^{\boldsymbol{t}}\right).$$

The $gpsim_t$ value helps in determining if the individual is **upright** (a value near 1), or if he or she is on the ground plane (a value near 0).

The features outlined above were selected due to their overall resilience to noise. Features such as the maximum height value could be used, but the eigen-based height is more reliable because using the covariance matrix to calculate the person's height utilizes statistics to automatically help remove some outliers that occur from incorrect silhouette segmentation or inaccuracies in the camera intersection construction of voxel person.

7. FUZZY LOGIC FOR STATE CLASSIFICATION

We are proposing a system to acquire a fuzzy confidence over time regarding the state of the resident from video. Our immediate goal is fall detection, a relatively short-time activity, but the framework is being developed for higher level reasoning about the resident's "well-being" over longer periods, such as days, weeks, month and even years, which depends entirely on this first level process. The next few sections detail the novel use of fuzzy logic for temporal pattern recognition in a human activity analysis setting.

The features extracted from voxel person are used to determine his or her current state. A finite set of states is identified ahead of time; the primary objective of processing in voxel space is to determine the membership degrees of each state at every time step. These state membership degrees are the input to activity analysis. An activity is a model defined according to specific state duration, frequency of state visitation, and state transition behavior. The collection of states, state i is denoted by S_i, that we have identified at the moment for fall recognition include

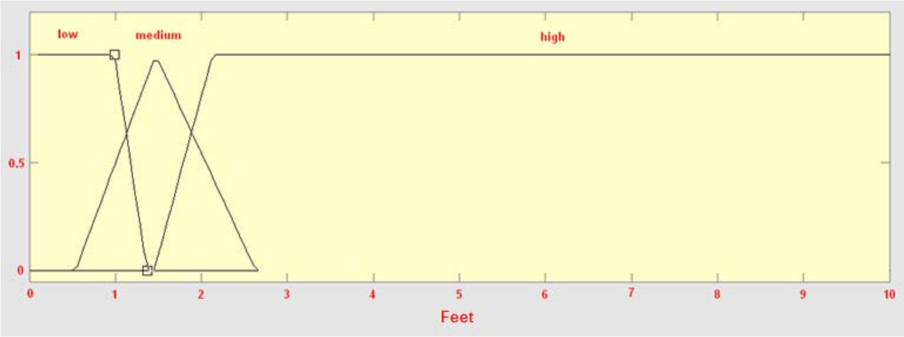
- **Upright** (S_1): This state is generally characterized by voxel person having a large height, its centroid being at a medium height, and a high similarity of the ground plane normal with voxel person's primary orientation. Activities that involve this state are, for example, standing, walking, and 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 some of the reason why we selected fuzzy inference for the determination of voxel person's present state.

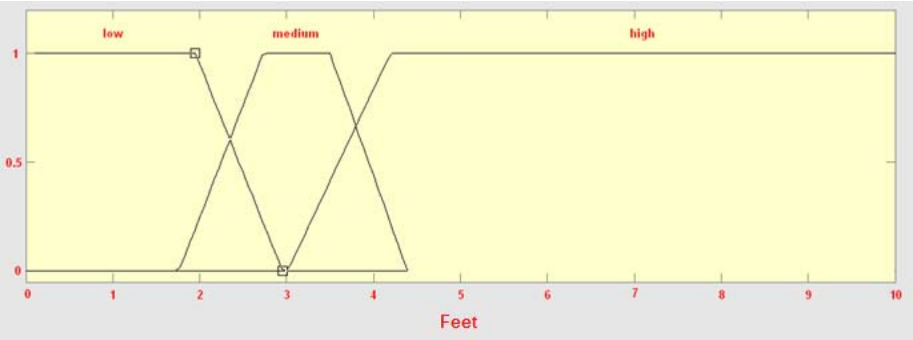
Fuzzy set theory, introduced by Lotfi A. Zadeh in 1965, is an extension of classical set theory [38]. One of the more well known branches of fuzzy set theory is fuzzy logic, introduced by Zadeh in 1973 [39]. Fuzzy logic is a powerful framework for performing automated reasoning. An inference engine operates 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.

Below, we define our linguistic antecedent and consequent variables, and the set of rules used for classifying the state of voxel person. The actual parameters of these fuzzy sets were chosen experimentally and validated by our nursing team, but they could be learned from appropriate training sequences. We use the standard Mamdani fuzzy inference system [39][40]. The fuzzy sets used below are trapezoidal membership functions, which are defined with respect to four points. The reason for selecting a trapezoidal to represent the fuzzy sets is that a triangular membership function can be formed by making the middle two points of a trapezoid equal. The linguistic antecedent variables and their respective terms are:

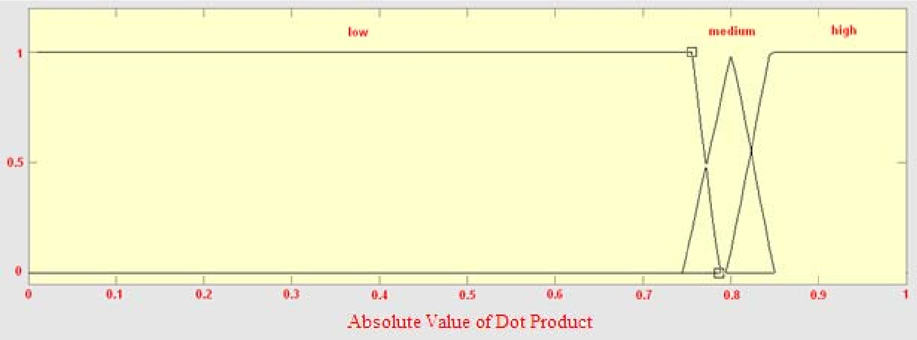
- Linguistic variable - Centroid



- Linguistic variable - Eigen-based height



- Linguistic variable - Max eigenvector and ground plane normal similarity



The three fuzzy consequent variables, **upright**, **in-between**, and **on-the-ground** are all defined with respect to the same 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], where the values [a b c d] represent the trapezoid left most point (a), the left central point (b), right central point (c), and right most point (d). The very low and high sets are centered at 0 and 1 respectively in order to help with the values that result from defuzzification. This ensures very low has a centroid at 0 and high has a centroid at 1. The fuzzy sets in the antecedent part of the rules below have the following symbolic mapping: L = low, M = medium, and H = high. The mapping for the fuzzy sets in the consequents for the rules listed below is: V = very low, L = low, M = medium, and H = high. These abbreviations simply make the rules displayable in a table. The antecedent variables are: centroid, eigen-based height, and max eigenvector and ground plane normal similarity. The consequent variable mappings are **upright**, **in-between**, and **on-the-ground**. The set of rules used to determine the state of voxel person is shown in Table 1.

Table 1. Fuzzy rules for state modeling

Rule		Centroid	Eigen Based Height	Normal Similarity		Upright	In Between	On the Ground
1	If	H	H	H	Then	L	V	V
2		M	H	H		L	L	V
3		L	H	H		V	L	L
4		H	M	H		V	H	V
5		M	M	H		V	H	L
6		L	M	H		V	H	H
7		M	L	H		V	L	H
8		L	L	H		V	V	M
9		H	H	M		L	V	V
10		M	H	M		L	L	V
11		L	H	M		L	H	V
12		H	M	M		L	H	V
13		M	M	M		L	H	V
14		L	M	M		V	H	L
15		L	M	L		V	L	H
16		L	L	M		V	L	M
17		H	H	L		H	V	V
18		M	H	L		M	V	V
19		L	H	L		L	L	V
20		H	M	L		M	L	V
21		M	M	L		L	L	V
22		L	M	L		L	H	V
23		M	L	L		V	H	L
24		L	L	L		V	L	H

It should be noted that these rules make it possible to detect when voxel person is lying on the ground, not just lying down anywhere in the room. If a person is lying on a couch or a bed, he or she should not have a low centroid and will not have a low height. The rule that would generally be dominant by voxel person lying on a bed or couch is rule 21. In this situation he or she would typically have a medium centroid and a medium height. In all of these mentioned situations, lying on the bed, lying on the ground, and lying on the couch, voxel person should generally have a low max eigenvector and ground plane normal similarity. In addition, further states could be identified and classified from these features. This approach makes it easy to add new rules for the recognition of new states.

The result of fuzzy inference, performed at each time step, is 3 defuzzified values (the centroid) corresponding to the confidence of **upright**, **in-between**, and **on-the-ground**. An example plot of the defuzzified outputs is illustrated in Figure 9. The camera capture rate was 3 frames per second and the 23 second scenario shows a subject falling and not getting back up.

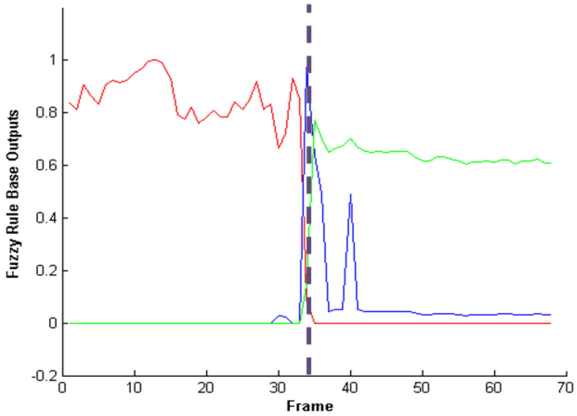
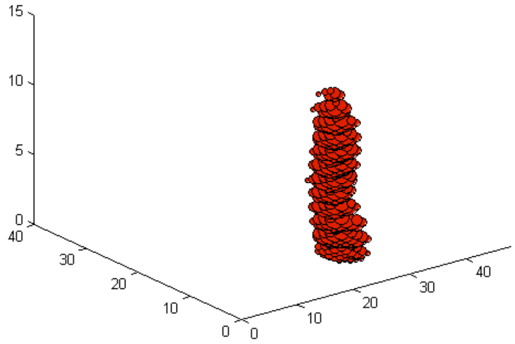
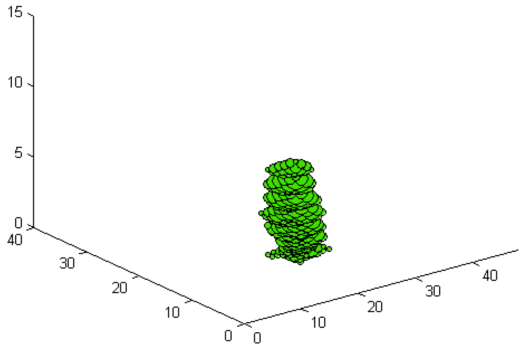


Fig. 9: Fuzzy inference outputs plotted for a voxel person fall. The x-axis is time, measured in frames, and the y-axis is the fuzzy inference outputs. The red curve is **upright**, the blue curve is **in-between**, the green curve is **on-the-ground**, and the dashed purple vertical line is where the human indicated a fall occurred. The frame rate was 3 per second, so the above plot is approximately 23 seconds of activity.

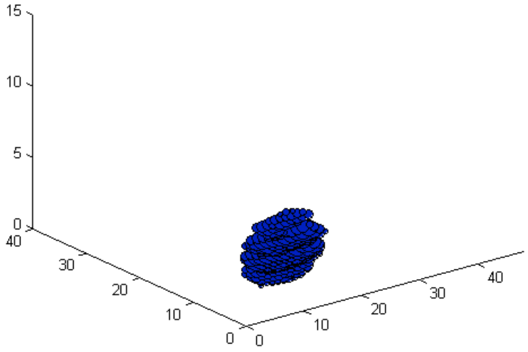
In the case of three states, voxel person can be color coded in order to illustrate the state memberships of the resident. The defuzzified consequent values, all in the interval [0,1], determine the amount of red, blue, and green in voxel person. Figure 10 is a sequence that shows the color coding of voxel person for the sequence shown in Figure 9. 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 fuzzy rule base outputs.



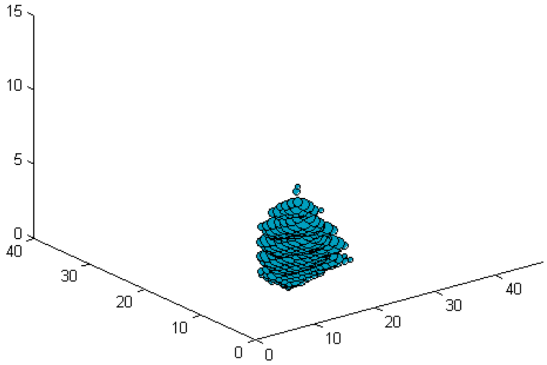
Frame 15: Upright



Frame 35: In Between



Frame 38: **On-the-ground**



Frame 40: **In-between** & **On-the-ground**
(Trying to Get Up)

Fig. 10: Color-coding of voxel person according to the defuzzified output values. Voxel persons color is a blend of the fuzzy rule system outputs. The **upright** state determines the amount of red, **in-between** is green, and **on-the-ground** adds blue.

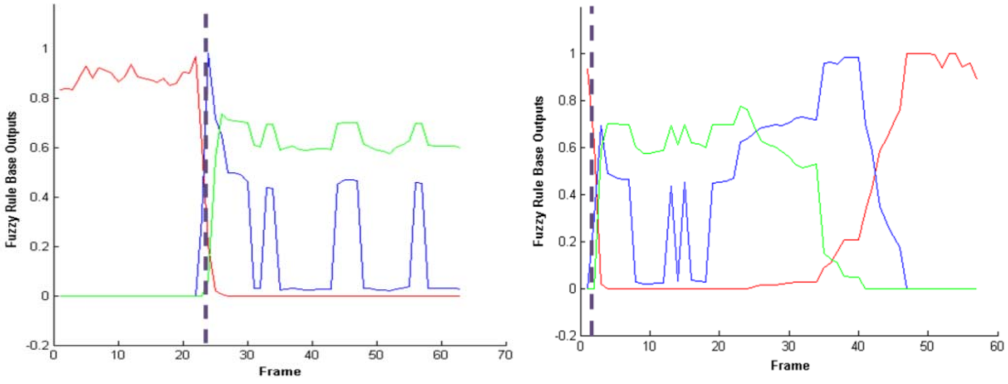
8. EXPERIMENTS AND DISCUSSION

In this section, we discuss the activity present in the temporal fuzzy confidence curves for different types of falls. This data set was hand segmented by a human to acquire a ground truth to compare the automated systems results against. Only activities that the system tracks were hand segmented. This comparison demonstrates how successful the fuzzy system in modelling the moment-by-moment, e.g., frame-by-frame, state according to a human.

All data was captured in the Computational Intelligence Laboratory at the University of Missouri. As mentioned above, movies illustrating the sequences and our processing of them can be found at http://cirl.missouri.edu/fallrecognition. Data is collected in a lab environment because of the severity of the activity being analyzed, and in particular the target elderly population. Sixteen short time period, 30 seconds to 1 minute in duration, fall activity sequences were studied. In 12 of these sequences the subject walked into the room, went over to a mat, and fell to the ground. Falls were performed differently, meaning that sometimes the person fell forward, sometimes backwards, and also to the side. Four of the 16 sequences were not falls that we wanted to recognize, as determined by the nurses, such as tripping and getting back up immediately and one being on the ground for too short of a time period. Two longer sequences, approximately 7 and 11 minutes, are included. The camera capture rate was 3 fps and a total of 5512 frames were analyzed (approximately 30 minutes). This is a sufficient number of frames to base the following statistics on and a reasonable amount of data to have a human hand segment. Types of falls include: 1) falls where the subject simulated a severe injury and laid on the ground motionless, 2) falls where the subject unsuccessfully attempted to get back up, and 3) and non-severe falls that lasted only a couple of seconds and then the subject got back up.

The first fall scenario, shown in Figure 9, is the subject simulating a severe fall. The subject was **on-the-ground** for a moderate amount of time, a large sudden change in acceleration of voxel person occurred before **on-the-ground**, and there was very little motion during the **on-the-ground** time period. This is the prototypical fall, which needs to result in the triggering of an alert and help dispatched.

In the next type of fall, shown in Figure 11(a), is where the subject fell and tried to make it back to an upright position (approximately frames 32, 45, and 58), but were unsuccessful. The subject was predominantly **on-the-ground** for a moderate amount of time, a large sudden change in acceleration of voxel person was detected before they were **on-the-ground**, but there was motion detected while they were **on-the-ground**. If the individual keeps trying to get back up, it is possible that the system could be confused about whether the subject has fallen, is making it back up, or is performing a non-fall activity (such as exercising on the ground). Moments in which the subject tries to make it back up but is unsuccessful can be detected by simultaneously monitoring the **on-the-ground** and **in-between** state behavior. In the case that the subject makes it to an **in-between** state and then back to an **on-the-ground** state, but never back to an **upright** state, oscillating behavior between the two states will occur. While recognizing and discriminating between activities is relatively simple for most humans, it is extremely difficult for an automated system. This is a high level case-based computer vision and image understanding task that requires information about the context, temporal activity, and even inference about the mental and/or physical state of a subject.



(a) (b)

Fig. 11: Fuzzy inference outputs plotted for two voxel person falls. (a) Sequence where the subject fell and tried to get back up three times. (b) Sequence where the subject fell and was able to get back up. Red is **upright**, blue is **in-between**, green is **on-the-ground,** and the dashed purple vertical line is where the human indicated a fall occurred.

At approximately frames 32, 45, and 58, the fuzzy membership for **on-the-ground** increases when one might expect it to decrease because the human is trying to move into the **in-between** state. Analysis of voxel person and the rule firings during these time periods shows that there is a problem with the calculation of $gpsim_t$ when the ratio of the top two eigenvalues is near one. This feature operates the best when the person is **upright** or lying **on-the-ground**, which is good for detecting many types of falls, but when the person is hunched over and the voxel object is near spherical in shape, there is not a clearly distinguishable primary orientation and the feature is not always stable. Deciding what to do in this situation is difficult and requires additional information. If the person is propped up against some object, such as a couch, and there was a quick acceleration change before that moment and there is little motion afterwards, then the confidence in fall might be high. The point is that this domain is inherently fuzzy and case based and as more activities or fall conditions are added more features need to be added to help further contextualize the decision making.

In the third fall example, shown in Figure 11(b), the subject went to the ground abruptly and then was able to make it to an upright position. 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 pictures, to look at later. The storage of voxel person, not the original image, helps in the preservation of resident privacy.

In [41] we present a higher level fuzzy logic framework that operates on temporal linguistic summarizations, extracted from temporal fuzzy confidence curves, for reasoning about human activity. We show that these linguistic summarizations make it possible to automatically detect a variety of falls that vary in terms of the method performed and the time scale at which it was observed. Fuzzy logic is utilized again and the rule base for recognizing falls is designed by nurses. Each fall discussed in this section is recognized by this system.

While figures 9 and 11 show what a few common types of falls look like according to the fuzzy state memberships over a short time duration, which is good for illustration purposes, they do not stress the massive amount of information that the system is responsible for processing. Our system is designed to continuously track human activity over long time periods (minutes, hours, days, weeks, and months). Figure 12 shows approximately 11 minutes (2,042 frames) of video analysis, which is still a relatively short amount of time, in which the subject performed various activities, including: walking, standing, kneeling, tying shoes, stretching, and all three fall types mentioned above. The activities in Figure 12 are labeled by a human so that the reader can observe the fuzzy state memberships during the different respective time intervals. Once again, the system proposed in [41] is able to detect the falls with no false alarms using the temporal fuzzy state memberships.

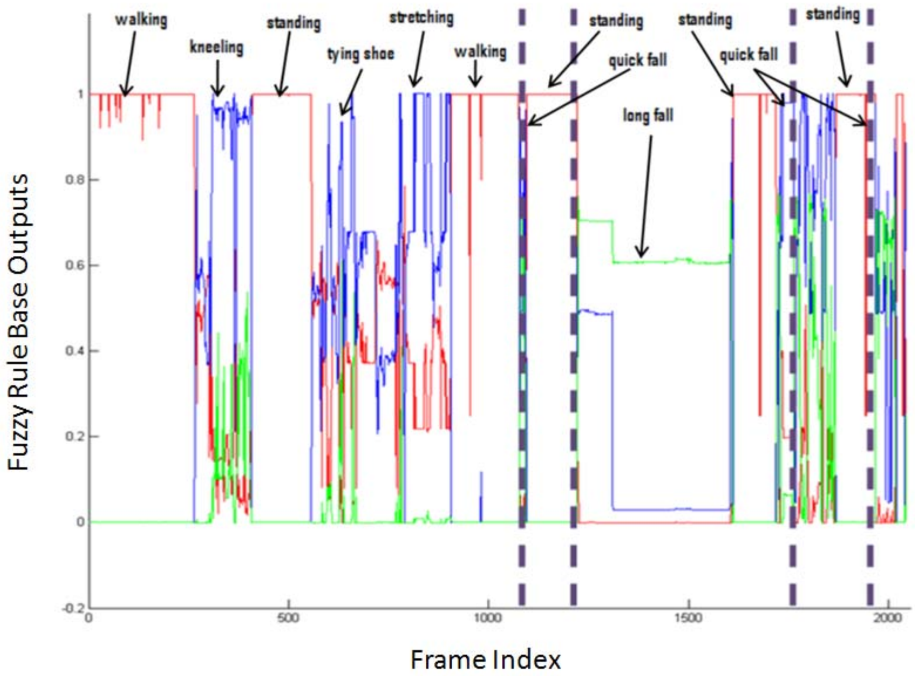
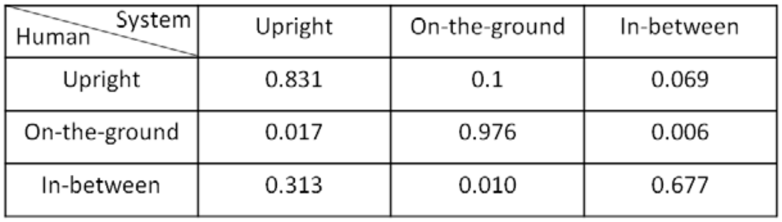


Fig. 12: Approximately 11 minutes of video analysis, 2,042 frames total. A total of 4 different falls occurred during this time period. The **upright** membership is shown in red, **in-between** membership is shown in blue, and **on-the-ground** is shown in green. Dashed vertical purple lines are the manually inserted moments where a fall occurred.

The human was asked to segment the video sequence according to the list of states the system recognizes. The human identified the start and end frames for each state and all frames in that interval are assigned the same state label. This ground truth is a crisp state labeling per frame. In order to compare the frame-by-frame state decisions of the fuzzy system to the human's labels the fuzzy state with the maximum membership value at each frame is used. Table 2 indicates how much of a particular state the fuzzy system recognized.

Table 2. Frame-by-frame comparison between the ground truths, acquired from the human, and the inferred state from our automated fuzzy system.



System Human	Upright	On-the-ground	In-between
Upright	0.831	0.1	0.069
On-the-ground	0.017	0.976	0.006
In-between	0.313	0.010	0.677

Table 2 shows that **on-the-ground** had the highest classification percentage, which is necessary for the recognition of falls, while **upright** had a classification rate of 83.1%. There are some instances (10%) in which our system incorrectly labeled **upright** as **on-the-ground**, which lowered the recognition rate of this state. This is due to two factors. The first and largest factor involves time intervals in which the subject moved into the far bounds of one or both of the cameras and the viewing angles make object reconstruction difficult. To address this, we are working on fuzzifying voxel person and feature extraction to take into account uncertainty related to the viewing ray angles and the distance of a voxel to a camera's focal plane. The second problem resides in the fuzzy sets used to build the rules. These sets were empirically defined by humans. Some situtations in the feature extraction process do not perfectly fit the empirically determined fuzzy sets. We will address this problem in the future by learning the fuzzy sets from training data and compare this to the nurse's system.

The **in-between** state had little similiarity with **on-the-ground**, but it was very similar to **upright**. This primarily has to do with the fuzzy sets used to classify that state. The automated system's fuzzy sets do not coorespond with the human's assesment of **in-between**. The human was quick to call someone **in-between**, while the fuzzy sets were designed to detect the time intervals when someone was half way between **upright** and **on-the-ground**. This may or may not be addressed depending on how well the system recognizes the higher level activities. It might not prove to be crucial to have the systems decision perfectly match that of the human.

9. CONCLUSION

In this paper, we presented a flexible soft-computing framework for modeling and monitoring human activity from video, in particular elderly falls. The result is human understandable information and meaningful confidences regarding activities for the sake of monitoring the "well-being" of a resident over different time scales. The importance of using fuzzy logic for modeling and monitoring human activity, which is significantly different from the majority of current human monitoring approaches, is the production of interpretable information that can be reliably used to infer and reject activity.

Silhouettes from multiple cameras are used to build a three-dimensional approximation of the human, i.e. voxel person. Voxel space provides a platform for the further refinement of our human model through the removal of additional shadows, reflective static surfaces, and error detection. Features are extracted from voxel person and used along with fuzzy inference to determine the temporal state of the resident. The resulting fuzzy rule base outputs can then be temporally processed to detect activity. Rules can be modified, added, or removed, allowing for per-resident customization based on knowledge about their cognitive and functionality ability.

The current system requires manual calibration by domain experts. However, the fuzzy sets and rules can be learned using a set of data that reflects the prototypical elder.

10. FUTURE WORK

As noted above, many of the quantities used in this work are based on empirical observations. We just finished collecting a larger dataset of falls using a stunt actor that was trained by nurses to perform activities, such as walking and falling, like an elderly person. This dataset, along with the one used in this paper, will be used to determine the fuzzy sets, fuzzy rules, and possibly some of the thresholds. In addition, it will be interesting to compare the learned system to the system designed by the nurses.

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. We are also working on converting this type-1 system into a type-2 fuzzy system. We are going to observe how type-2 fuzzy sets could benefit the activity analysis domain, by analyzing its effects on various components in the current system, such as: voxel person construction, feature extraction, fuzzy sets for features and states, and inference.

As mentioned in the results section, we are working on fuzzifying voxel person and feature extraction for addressing the uncertainty that arises from intrinsic camera parameters, the pixel view rays and their respective lensing amounts, and the location of voxels given a cameras position and focal plane. It is also possible that many of the components in our system could benefit if information regarding each cameras position and unique view of the scene was taken into account. For example, discovering the effect of placing of a camera on the ceiling looking down at the ground for detecting falls versus a camera located at chest height for observing an individual moving around and interacting with objects. We will investigate methods for assigning or learning camera specific confidences that can be used for object creation, feature extraction or making decisions with respect to states and/or activities.

The detection of falls is a form of short-term activity monitoring, but the work presented here is in no way limited to short-term activity recognition. 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.

We recently demonstrated the utility of low-cost high performance specialized graphics hardware called Graphics Processing Units (GPUs) for speeding up computation of the Fuzzy C-Means (FCM) clustering algorithm [42][43]. We showed that under particular clustering configurations, two orders of magnitude in speedup can be achieved. The NVIDIA 8800 GPU has 128 processing elements that operate in parallel and is capable of around 350 GFLOPS [44]. We have also shown a method to transfer fuzzy inference to the GPU, again achieving over two orders of magnitude under particular profiles [45][46]. We are currently working on a GPU solutions for silhouette extraction, voxel person construction (crisp and fuzzy voxel person), type-2 fuzzy inference, and feature extraction.

11. ACKNOWLEDGEMENTS

D. Anderson and R. 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).

12. REFERENCES

[1] W.P. Zajdel, “Bayesian visual surveillance: from object detection to distributed cameras,” PhD Dissertation, University of Amsterdam, 2006.

[2] C. Stauffer and W.E.L. Grimson, “Learning patterns of activity using real-time tracking,” in *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 22, pp. 747-757, 2000.

[3] A. Nusimow, “Intelligent video for homeland security application,” in *IEEE Conf. on Technologies for Homeland Security*, pp. 139-144, 2007.

[4] G.B. Garibotto, *Computer Vision and Pattern Recognition in Homeland Security Applications.* Heidelberg: Springer Berlin, 2007.

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

[6] D. Anderson, J.M. Keller, M. Skubic, X. Chen, and H. Zhihai, “Recognizing falls from silhouettes,” *28th Annual Intl. Conf. of the IEEE Engineering in Medicine and Biology Society*, pp. 6388–6391, 2006.

[7] G. Demiris, M. Skubic, M. Rantz, K. Courtney, M. Aud, H. Tyrer, Z. He, and J. Lee, "Facilitating interdisciplinary design specification of ‘smart homes’ for aging in place," in *Proc., Intl. Congress of the European Federation of Medical Informatics*, pp. 45-50, 2006.

[8] G. Demiris, K. Courtney, M. Skubic, and M. Rantz, "An evaluation protocol of a smart home application for older adults," in *Proc. Intl. Conf. Addressing Information Technology and Communications in Health*, pp. 319-323, 2007.

[9] G. Demiris, M. Skubic, M. Rantz, J. Keller, M. Aud, B. Hensel, and Z. He, "Smart home sensors for the elderly: a model for participatory formative evaluation," in *Proceedings, IEEE EMBS Intl. Special Topic Conf. on Information Technology in Biomedicine*, pp. 1-4, 2006.

[10] M. Rantz, R. Porter, D. Cheshier, D. Otto, C. Servey, R. Johnson, M. Skubic, H. Tyrer, Z. He, G. Demiris, J. Lee, G. Alexander, and G. Taylor, "TigerPlace, a state-academic-private project to revolutionize traditional long term care," *Journal of Housing for the Elderly*, 2007.

[11] 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,” *Medical Informatics and the Internet in Medicine*, 2004.

[12] N. M. Oliver, B. Rosario, and A. P. Pentland, "A Bayesian Computer Vision System for

Modeling Human Interactions," in *IEEE Tans. on Pattern Analysis and Machine Intelligence*, Vol. 22, pp. 831-843, 2000.

[13] T. Parag, A. Elgammal, and A. Mittal, "A framework for feature selection for background subtraction," in *IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, pp. 1916-1923, 2006.

[14] S. McKenna, S. Jabri, Z. Duric, H. Wechsler, and Z. Rosenfeld, "Tracking groups of people," *Computer Vision and Image Understanding*, Vol. 9, pp. 42-56, 2000.

[15] I. Haritaoglu, D. Harwood, and L. S. Davis, "W4: real-time surveillance of people and their activities," I*EEE Trans. Pattern Analysis and Machine Intelligence*, pp. 809-830, 2000.

[16] N. Ohta, "A statistical approach to background suppression for surveillance systems," in *Proc. of IEEE Intl. Conference on Computer Vision*, pp. 481-486, 2001.

[17] L. Wang, T. Tieniu, H. Ning, W. Hu, "Silhouette analysis-based gait recognition for human identification," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 25, pp. 1505-1518, 2003.

[18] L. Dar-Shyang, "Effective gaussian mixture learning for video background subtraction," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 27, pp. 827-832, 2005.

[19] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," Morgan Kaufmann Publishers Inc, Proc. of the IEEE, 1989.

[20] D. Anderson, D. Bailey, and M. Skubic, "Hidden Markov model symbol recognition for sketch-based interfaces," in *AAAI Fall Symp.: Making Pen-Based Interaction Intelligent and Natural*, pp. 15-21, 2004.

[21] R. Davis and T. M. Sezgin, "HMM-based efficient sketch recognition," in *Proc. of the Intl. Conf. on Intelligent User Interfaces*, Vol. 7, pp. 4564-4570, 2005.

[22] P. Gader and M. A. Mohamed, "Generalized hidden Markov models I: theoretical frameworks," *IEEE Trans. on Fuzzy Systems*, Vol. 8, pp. 67-81, 2002.

[23] J. Bilmes, "A gentle tutorial of the EM algorithm and its application to parameter estimation for gaussian mixture and hidden Markov models," Technical Report ICSI-TR-97-021, 1998.

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

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

[26] M. Brand and V. Kettnaker, "Discovery and Segmentation of Activities in Video," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 22, No. 8, 2000.

[27] M. Brand, N. Oliver, and A. Pentland, "Coupled Hidden Markov Models for Complex Action Recognition," *in Proc. IEEE Computer Vision and Pattern Recognition*, pp. 994-999, 1997.

[28] W.L. Buntine, "Operations for Learning with Graphical Models," *J. Artificial Intelligence Research*, pp. 159-225, 1994.

[29] K. Murphy, "Dynamic Bayesian Networks: Representation, Inference and Learning". PhD thesis, Dept. Computer Science, UC Berkeley, 2002.

[30] L.D. Wilcox and M.A. Bush, "Training and Search Algorithms for an Interactive Wordspotting System," *Proc. Int'l Conf. Acoustics, Speech, and Signal Processing*, Vol. 2, pp. 12-49, John Wiley & Sons, 1991.

[31] R. H. Luke, D. Anderson, J. M. Keller, and M. Skubic, "Moving Object Segmentation from Video Using Fused Color and Texture Features in Indoor Environments," Under review by *IEEE Transactions on Image Processing*, 2008.

[32] B. G. Baumgart, *Geometric Modeling for Computer Vision.* Technical Report AIM-249, Artificial Intelligence Laboratory, Stanford University, 1974.

[33] B. C. Vemuri and J. K. Aggarwal, “3-D model construction from multiple views using range and intensity data,” in *Proceedings of IEEE Conf. Computer Vision and Pattern Recognition*, pp. 435-437, 1986.

[34] A. Laurentini, “The visual hull concept for silhouette-based image understanding,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 16, pp. 150-162, 1994.

[35] G. Dudek and D. Daum, “On 3-D surface reconstruction using shape from shadows,” in *IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, pp. 461-468, 1998.

[36] M. Pardas and J. Landabaso, “Foreground regions extraction and characterization towards real-time object tracking,” in *Proc. of Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms*, 2005.

[37] J. Weng, P. Cohen, and M. Herniou, " Camera calibration with distortion models and accuracy evaluation”, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 14, pp. 965-980, 1992.

[38] L. Zadeh, “Fuzzy sets,” *Information Control*, pp. 338-353, 1965.

[39] L. A. Zadeh, “Outline of a new approach to the analysis of complex systems and decision processes,” *IEEE Trans. on System, Man, and Cybernetics*, 1973.

[40] E. H Mamdani and S. Assilian, “An experiment in linguistic synthesis with a fuzzy logic controller,” *Intl. Journal of Man-Machine Studies*, 1975.

[41] D. Anderson, R. H. Luke, J. M. Keller, M. Skubic, M. Rantz, and M Aud, “Linguistic summarization of activities for fall detection using voxel person and fuzzy logic,” Under review by *Computer Vision and Image Understanding*, 2007.

[42] D. Anderson, R. Luke, and J.M. Keller, “Speedup of Fuzzy Clustering Through Stream Processing on Graphics Processor Units,” *IEEE Transactions on Fuzzy Systems*, 2006

[43] D. Anderson, R. Luke, and J.M. Keller, “Incorporation of Non-Euclidean Distance Metrics into Fuzzy Clustering on Graphics Processing Units,” *Intl. Fuzzy Systems Association*, 2007

[44] Nvidia Corp., “GeForce 8800,” Nov. 2006, http://www.nvidia.com/page/geforce_8800.html

[45] N. Harvey, R. H. Luke, J. M. Keller, and D. Anderson, “Speedup of Fuzzy Logic through Stream Processing on Graphics Processing Units”, *in Proc. of IEEE Congress on Evolutionary Computation*, 2008.

[46] D. Anderson and S. Coupland, “Parallelisation of Fuzzy Inference on a Graphics Processor Unit Using the Compute Unified Device Architecture”, *UKCI 2008, the 8th Annual Workshop on Computational Intelligence*, 2008.