# Evaluation of a Video-Based Fall Recognition System for Elders Using Voxel Space

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

***Abstract*—Video is a rich source of information that can be used to passively monitor the activity of elders. The amount of information contained in video is significantly greater than other sensing technologies such as RFID tags and motion sensors. Privacy of residents is preserved by not using the raw video, but instead, extracting binary silhouette maps, which represent the pixels a person occupies in an image. Silhouettes acquired from multiple cameras viewing the same scene are used to build a three-dimensional object whose activity is linguistically summarized for activity monitoring. These linguistic summarizations are used for abnormal event detection, specifically for the automated detection of falls. In this paper, we present three measures for system performance evaluation and discuss successes and difficulties in video-based human activity recognition of falls.**

## I. Introduction

We are researching passive monitoring technologies for assisting elders with "aging in place". This includes adverse event detection from video for activities such as falls [1][2]. Privacy is preserved by not using the raw video, but extracting binary silhouette maps, which represent the pixels a person occupies in an image. Focus groups at the "aging in place" facility of residential apartments known as TigerPlace [3] indicate that elderly residents are willing to consider silhouette-based images for abnormal event detection such as falls [4].

A reliable video-based monitoring system must be able to discriminate between similar appearing activities, such as a subject on the floor stretching or sleeping versus having fallen. While these tasks are often relatively simple for a human, they are extremely difficult for an automated system. This is a high level 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. We have designed a soft computing approach to human activity monitoring [5][2], in which knowledge is explicit and linguistic. Rules for activity monitoring can be inserted, removed, and modified by domain experts, such as nurses, based on cognitive and/or physical information regarding each specific resident. In addition, the system produces linguistically summarized information in a natural language format that caregivers can utilize.

In an eldercare context, false alarms can be expensive. A false alarm may result in an alert being generated, such as a fall, requiring the intervention of a caregiver. Too many false alarms could result in a loss of trust, or worse, loss of use of the system. However, missing a single fall is the worst case scenario. Identifying an acceptable false alarm rate and understanding the conditions in which many false alarms occur is of vital use for the long term success of an automated system. In this paper, we identify and evaluate three measures for the assessment of various types of information and fall classification in our video system.

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 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".

Thome and Miguet demonstrated a fall detection procedure that uses Hierarchical Hidden Markov Models (HHMM) [7]. The HHMM is hand designed and operates on an observation sequence of rectified angles. Johnson and Sixsmith [8] 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 alert.

## II. Linguistic Summarization Of Activity

Our first step in human activity analysis is silhouette extraction (shown in figure 1). This is an image processing and computer vision classification task, in which the objective is to discover the pixels in the current image that belong to the human. This is not a simple task and has been the subject of much research over the years [9][10]. Objects move in a scene, illumination changes occur, and shadows and other phenomenon such as reflections further complicate automated extraction. Our silhouette extraction system is adaptive and fuses texture and color information [11]. The camera is assumed to be stationary and a background model is constructed. As each new image is acquired, features are extracted and locations that do not belong to the background are identified and labeled as silhouette.

Manuscript received April 30, 2008. This work was supported in part by the National Science Foundation (ITR award llS-0428420) and the Administration on Aging (90AM3013). D. Anderson and R. Luke are pre-doctoral biomedical informatics research fellows funded by the National Library of Medicine (T15 LM07089).

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

M. Rantz, and M. Aud are with the Sinclair School of Nursing at the University of Missouri, Columbia, MO 65211 USA (e-mails: rantzm@missouri.edu and audm@health.missouri.edu).

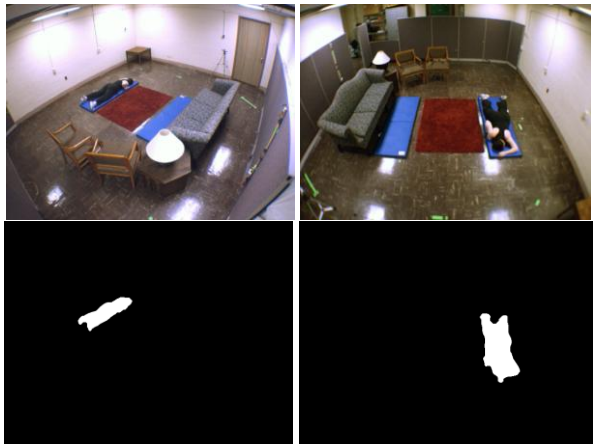

Fig. 1: Fall shown from two cameras monitoring the same scene. White regions in the bottom two images, i.e. the silhouettes, are the human.

In [5] we presented a robust method for the construction of a three-dimensional object, specifically a representation of the human called voxel person, from the back projection of silhouettes from multiple cameras viewing the same scene. 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 volume, and the pixels from which a particular voxel is viewable are recorded. Corresponding silhouettes, those with the closest time stamps, are acquired. For each camera a new list is constructed, i.e. the set of all voxels in foreground regions, the silhouette. The next step is the intersection of these new voxel lists, which results in voxel person. The procedure is illustrated in figure 2.

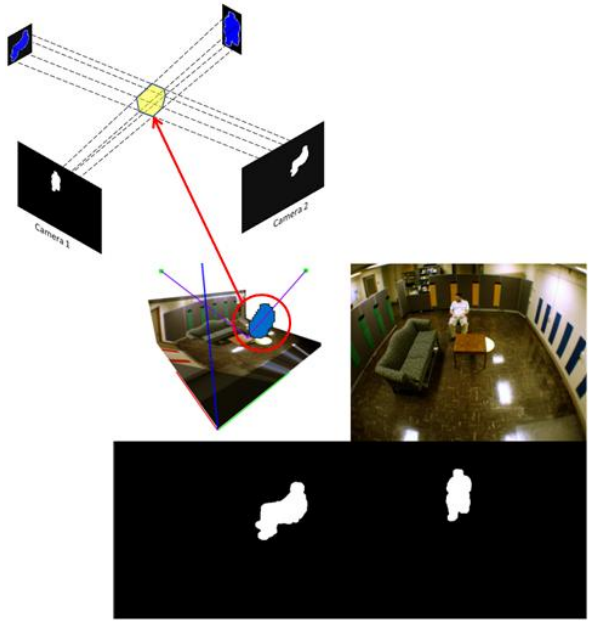


Fig. 2: Back projection of two-dimensional image plane silhouettes in three-dimensional space for voxel person construction.

We use fuzzy logic for human activity analysis. Fuzzy set theory, introduced by Lotfi A. Zadeh in 1965, is an extension of classical set theory [12]. One of the more well-known branches of fuzzy set theory is fuzzy logic, introduced by Zadeh in 1973 [13]. 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, a feature that is tracked and helps with determining when the subject is on the ground. This variable can assume the terms *low*, *medium*, and *high*.

Our first step in monitoring human activity from video involves acquiring confidences in the states of voxel person (e.g., upright, on-the-ground), a frame-by-frame decision process [5]. The current set of states (level one quantities that are later used to recognize activity) include: on-the-ground, upright, in-between, on-the-couch, and on-the-chair [5][14], as described below.

**Upright**: This state is generally characterized by voxel person having a large height, his 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**: 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**: 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.

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

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

Our next step is linguistic summarization of this information and the recognition of specific activities, e.g., falls [2]. This second stage uses domain expert knowledge regarding activities to produce a confidence in the occurrence of an activity. Rules allow for the recognition of common performances of an activity, as well as the ability to model special cases. 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 3 illustrates our activity recognition framework.

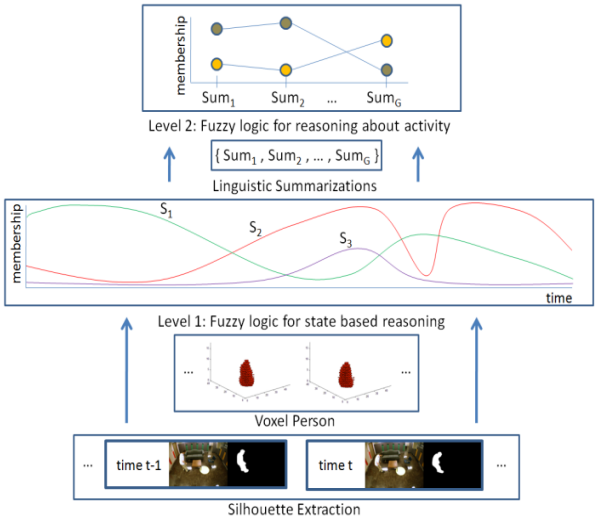


Fig. 3: Activity recognition framework, which utilizes a hierarchy of fuzzy logic systems based on voxel person. The first level is reasoning about the state of the individual. Linguistic summarizations are produced and fuzzy logic is used again to reason about human activity.

Decisions regarding the current activity can be made at each time step, but the result is too much information. 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.

## III. Fall Detection

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 was captured in our lab using students as subjects. The rule base for recognizing falls, validated by nurses, can be found in [2]. Features, described in [2][5][14], are extracted from voxel person and used in the rule base. Example features used to reason about the state of voxel person include the height (indicates if the subject is on the ground) and a quick recent change in acceleration (looking for a quick change in speed at the beginning of a fall). Figure 4 shows these two features for an example fall sequence and figure 5 shows the automated decision making output.

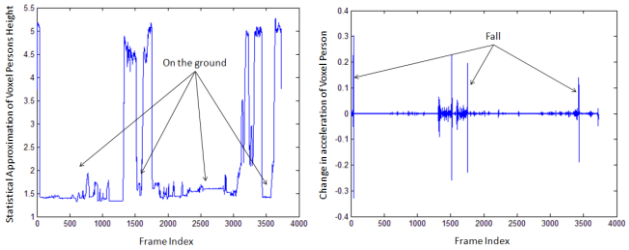


Fig. 4: Two features used to monitor the activity of voxel person. (left) The statistical approximation of voxel person's height feature, which indicates whether the subject is on the ground or upright. (right) The change in acceleration of voxel person feature, which is one source of information that indicates a potential fall.

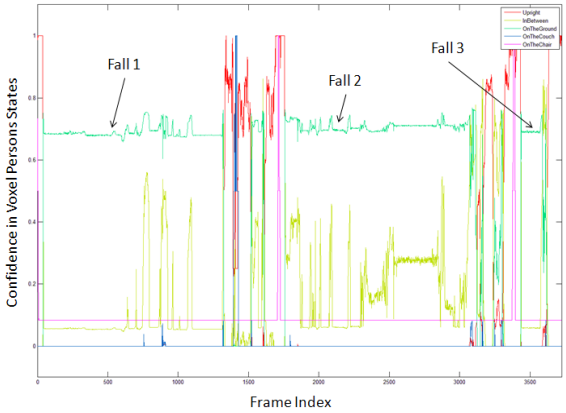


Fig. 5: Activity sequence, same as figure 4, which consists of 3650 frames (approximately 12 minutes using a capture rate of 5 fps). The output of reasoning about the state of voxel person (level one of fuzzy inference) is plotted over the time/frame domain. Labels indicate where the second level of fuzzy inference classified a fall. This sequence contained three falls, all of which were correctly recognized.

## IV. System Evaluation Metrics

The data set analyzed in this paper was manually hand segmented to acquire a ground truth for comparison against the automated systems results. Only the activities that the system tracks were hand segmented. The beginning and ending frames for each activity are identified. There are multiple ways to evaluate the performance of the system given the ground truth and the outputs at each level of fuzzy inference. The three metrics identified and evaluated here for this data set are:

**Metric 1:** Matching between the frame-by-frame state decisions (according to the fuzzy state with the maximum membership value at each frame in the first level of fuzzy inference) and the frame-by-frame ground truth labels. The human indicated the start and end time frames and all frames in this interval are automatically assigned the same label.

**Metric 2:** Matching between linguistic summarizations produced by processing the first level of fuzzy inference results and the hand annotated data. This measures how successful the summarization system is in terms of correspondence with what a human produced. However, this metric does not indicate how much the system summaries and the ground truth intervals overlap. When the first and second metric are analyzed

together, an understanding of how much the linguistic summarization and ground truth intervals overlap is possible. This is important because falls need to be recognized in a timely manner. For this metric, a zero score is the best.

**Metric 3:** Matching between the fall detection produced by the second level of inference and when a fall occurred, as noted by the manual segmentation. This is a measure evaluating the success of the second level of fuzzy inference.

## V. Results

The fall data set consists of eighteen sequences. The camera capture rate was 3 fps and a total of 5512 frames were captured (approximately 30 minutes). The two subjects walked around the room, stood still, kneeled, fell, and sat on the couch and the chair (example images are shown in figures 6 and 7). Kneeling, lying on the couch, and sitting on the chair with feet on a coffee table were included to show that some common activities that might appear as a fall are not misclassified by our system (examples in figure 8). 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 after which the person got back up, falls where the person stayed down on the ground but attempted to get back up, and falls where the person simulated a severe injury and laid on the ground motionless.

Fig. 6: Walking, kneeling, and sitting on the couch and chair.

Fig. 7: Walking, falling, and stretching.

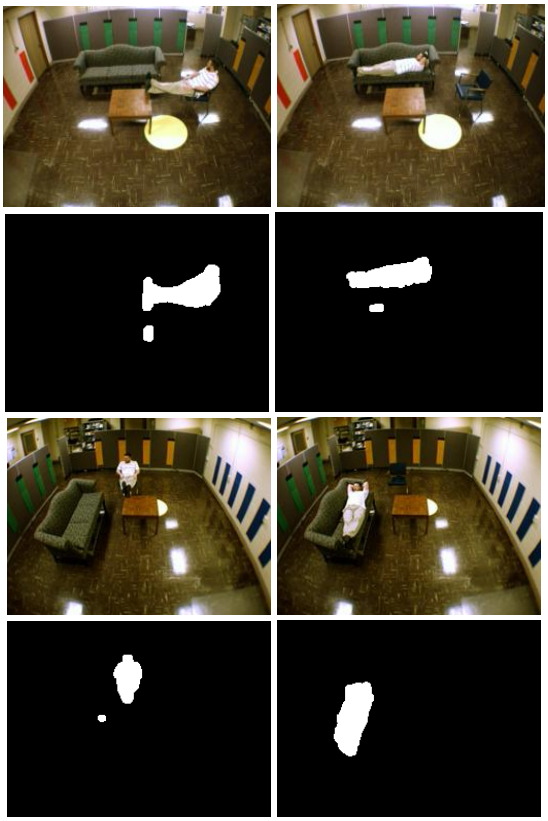

Fig. 8: Lying on the couch and sitting on the chair with feet up activities, which could be misinterpreted as a fall, are not recognized as a fall in our system. Rules for falls and knowledge about three-dimensional voxel person helps with the elimination of many false alarms.

Metric 1, table 1, shows the evaluation of the system from the standpoint of frame-by-frame state decisions.

Table 1. Comparison of frame-by-frame state decisions between the system, s, and the ground truth, t, (Metric 1).

| t \ s | on the chair | on the couch | upright | on the ground | in between |
|---|---|---|---|---|---|
| on the chair | 0.989 | 0.000 | 0.011 | 0.000 | 0.000 |
| on the couch | 0.000 | 1.000 | 0.000 | 0.000 | 0.000 |
| upright | 0.000 | 0.004 | 0.827 | 0.169 | 0.000 |
| on the ground | 0.000 | 0.000 | 0.017 | 0.976 | 0.006 |
| in between | 0.000 | 0.000 | 0.513 | 0.010 | 0.477 |

The results in table 1 are shown as percentages and they indicate the frequency at which our system agrees or disagrees with the human's labels (each row sums to one, within numerical precision of the displayed numbers). The results show that the system captures nearly all of the on-the-chair and all of the on-the-couch states (those activities that mostly depend on the spatial location in the room of voxel person and a static object). These activities still involve reasoning about the pose of the subject, but the position in the room contexutalizes the activity and simplifies identification. A room is segmented by a human into regions that are used to help track and reason about activity. Figure 9 shows an example apartment segmentation and illustrates how we track voxel person interacting with scene regions.

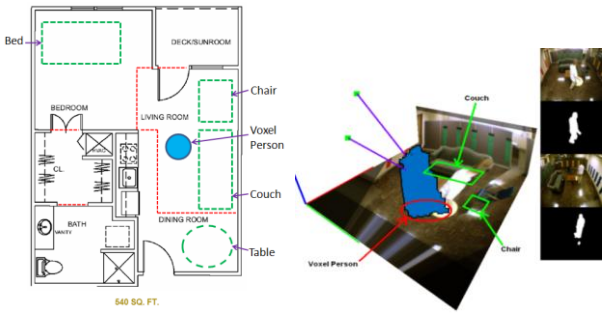


Fig. 9: A room is segmented by a human into different regions and large static objects are identified. Voxel person is projected onto the x-y plane and a measure of region overlap is produced for tracking [14].

As more activities are identified for the couch and chair, the rates will most likely decrease some, but these tasks are clearly distinguishable from the majority of activities that the subject performs at various locations in the apartment.

The upright state has good classification (82.7%); however, some time intervals were called on-the-ground (16.9%). These situtations are 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 the feature extraction process to take into account factors such as the viewing ray angles and the distance of a voxel to the camera 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 comparing this to the nurse's system.

The on-the-ground state was recognized in 97.6% of the image frames, which is critical for fall recognition. The in-between state had little similiarity with on-the-ground, but it was essentially 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 really detect the time intervals when someone was half way between upright and on-the-ground.

Table 2 is the system evaluation results according to Metric 2, which compares the linguistic summarizations to the ground truth labels.

Table 2. Comparison of linguistic summarizations, s, to ground truth labels, t, (Metric 2), computed as s-t.

| on the chair | on the couch | upright | on the ground | in between |
|---|---|---|---|---|
| 0 | 0 | -31 | 2 | 5 |

Negative numbers in table 2 indicates fewer linguistic summarizations than labeled intervals were found. Positive numbers indicate that we generated more summaries than there were labels, and zero values indicate that there were the same number of summaries as labels. The results show that, in activities that involve interaction with the chair and the couch (static scene regions/objects), the automated system finds what the human identified.

Table 2 shows that a fair number of upright time intervals went undetected. This is mostly because linguistic summarizations that are too short in time duration are removed by our system. These periods are possibly due to incorrect silhouette segmentation, inaccuracies in fuzzy inference, or high frequency activity that is not related to fall detection. Elders do not generally perform extremely quick activities, such as being on the ground for only one second. We remove linguistic summarizations less than two seconds.

We discover more on-the-ground and in-between states than human labelings. After looking at the level one inference results, this is because of time intervals of incorrectly inferred activity and silhouette segmentation error (bad features extracted, hence, incorrect inference). As a result, the linguistic summarizations produced by our system are segmented into a larger number of smaller summaries. This is not ideal from a report generation standpoint, but from a recognition standpoint it is not bad as long as we are still able to automatically recognize falls from these summaries (which we are able to do).

Metric 3 is the most important; it shows how many times the second level of inference correctly classified a fall. Sixteen short time period sequences were evaluated (30 seconds to 1 minute in duration each). In 12 of these sequences the subject walked into the room, went over to the mat, and then fell to the ground (where the falls were performed in different fashions, such as to the front, the side, etc). Each fall was successfully detected; there were no false alarms. Four of the 16 sequences were non-fall activities, such as bending down to tie one's shoes and also tripping and getting back up immediately. Nurses indicated that they would like to get a summary when someone is on the ground for a short amount of time, but they do not want to have an alert generated. We did not call any of these false alarm situations a fall.

Two longer time period sequences were evaluated for falls. There were no falls in the first sequence, which was approximately 7 minutes in duration, and the system correctly did not classify any falls. In the second sequence, approximately 11 minutes in duration, there were four on the ground periods, however, only two falls. In the first fall the subject stayed on the ground for a long time period and in the second case the subject fell and repeatedly tried to get back up but was not able to. In both of these situations our system correctly classified the fall. There were two on the ground activities that were intended to look like a fall. In the first case the subject tripped and got right back up and in the second case the subject went to the ground but was able to make it back up in a reasonable amount of time (quantified by the fuzzy sets that the nurses picked). In both of these cases the system correctly did not flag a fall.

Even though there were more on-the-ground linguistic summaries than there were on the ground activities, the system correctly classified all of the falls.

## VI. Conclusions

In this paper, we demonstrated the performance of a video-based activity analysis system for assisting elders with "aging in place". The primary activity analyzed was falling, which is a relatively short time activity. The

system is built using soft computing and activities are recognized using linguistic summarizations of activity. The system's knowledge is expressed using linguistic variables, which helps in understanding its successes and failures, and more importantly, identifying and fixing problems. The linguistic summaries also provide a rich set of reduced descriptions about the video sequence in a language and format that users can understand. The metrics that we introduced indicate that there is still some work to be done with respect to matching the exact number of lingusitic summarizations and the hand labeled activities and the frame-by-frame decisions made. However, the system generated an adequate number of on-the-ground summaries, which enabled the second level of inference to correctly classify falls and distinguish between fall and non-fall activities.

## VII. Future Work

Many of the quantities used in this work are based on empirical observations and domain knowledge from nurses. As mentioned above, we are investigating using training data to determine fuzzy sets and fuzzy rules. This should help with some of the deficiencies observed in evaluation metrics 1 and 2. This will also provide a comparison between domain experts and an automated way of rule and/or fuzzy set acquisition.

We just captured a larger dataset of falls using stunt actors that can be used to learn the system parameters and further test the system under a wider range of activities and subjects. To make sure that the actors performed the falls in a similar fashion to the way that elders fall, nurses coached the stunt actors. We showed great discriminatory ability for the range of activities that were included in the data set analyzed in this paper. However, we captured a more complicated and larger set of false alarms activities in the stunt actor data set. We will analyze these situations, measure how the system performs, and recommend corrections based on these findings.

## References


[1] 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.

[2] D. Anderson et al, "Linguistic Summarization of Activities from Video for Fall Detection Using Voxel Person and Fuzzy Logic," Under Review by *Computer Vision and Image Understanding*.

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

[4] G. Demiris et al, "Older adults' attitudes towards and perceptions of 'smart home' technologies: a pilot study," in *Medical Inf. and the Internet in Medicine*, 2004.

[5] D. Anderson et al, "Modeling Human Activity From Voxel Person Using Fuzzy Logic," Under Review by *IEEE Transactions on Fuzzy Systems*.

[6] T. Martin et al, "Fuzzy ambient intelligence for next generation telecare," in *IEEE Int. Conf. on Fuzzy Systems*, 2006, pp. 894-901.

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

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

[9] C. Stauffer and W.E.L. Grimson, "Learning patterns of activity using real-time tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, pp. 747-757, 2000.

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

[11] R. H. Luke et al, "Moving Object Segmentation from Video Using Fused Color and Texture Features in Indoor Environments," Under Review by *IEEE Transactions on Image Processing*, 2008.

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

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

[14] D. Anderson et al, "Extension of a Soft-Computing Framework for Activity Analysis from Linguistic Summarizations of Video," *IEEE International Conference on Fuzzy Systems*, WCCI 2008.