Segmentation and Linguistic Summarization of Voxel Environments using Stereo Vision and Genetic Algorithms

Derek T. Anderson, Robert H. Luke, *Student Members, IEEE*, and James M. Keller, *Fellow, IEEE*

***Abstract*—For reasons such as computational complexity, spatial and temporal information reduction, and human understandability, it is important that computer vision systems be equipped with the means to summarize their content in a natural language. Such rich descriptions are of use by both humans and computers for describing, recognizing, and tracking objects, activity, and their interactions at a desired level of abstraction. A genetic algorithm is introduced here for segmenting non-human objects deemed relevant to human activity analysis in stereo vision acquired voxel environments. This approach is of use in an Eldercare setting as it relates to monitoring the "well-being" of residents through acquiring and detecting deviations in patterns of typical behavior as well as recognizing abnormal events, such as fall detection.**

I. Introduction

Our objective is the utilization of passive monitoring technologies and computational intelligence for assisting elders with "aging in place". This includes the discovery of long-term patterns of typical behavior and detection of deviation from such behavior [1][2], as well as short-term adverse event detection, e.g. falls [3][4][5]. This research is part of a interdisciplinary collaboration between Engineers, Nurses, and other Health Care individuals at the University of Missouri [6][7]. 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.

Video sensors are a rich source of information that can be used to monitor a scene. High-level computer vision systems performing human activity analysis must be provided reliable information regarding the whereabouts of people. Privacy is preserved by not viewing or storing the raw video. Instead, binary silhouette maps, which represent the pixels a person occupies in an image, are typically used. Focus groups at the "aging in place" facility of residential apartments known as TigerPlace [8] indicate that elderly residents are willing to consider technologies such as silhouette-based images for abnormal event detection such as falls [9].

Martin et al. [10] present 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 changes in sleep patterns over several months.

Johnson and Sixsmith [11] use an infrared array technology to acquire a low resolution thermal image of the resident and they track the human using elliptical-contour gradient-tracking. Falls are detected using a neural network that took the vertical velocity of the person as input. Their fall classification results are poor, only capturing around one-third of falls. However, no non-fall scenarios resulted in a fall alert. In [12] we present a silhouette non-voxel-based method for fall detection using hidden Markov models (HMMs). Thome and Miguet have similar work in the area of hierarchical HMMs for fall detection [13]. The reason for moving away from such an approach is due to shortcomings in statistical inference for this domain. One desires a human understandable confidence in the performance of an activity, not an un-interpretable numerical value that is only of use as it relates to selecting the most likely model from a candidate set (from which a new activity may not even belong).

In [3][4][5] we present a soft computing approach to the (temporal) linguistic summarization of human activity. The system is demonstrated for elderly fall detection. While image silhouettes were initially used to build voxel humans for activity analysis, our more recent silhouette-free stereo vision human voxel object system [14] should be used in its place. An advantage of our activity analysis system is that rules can be inserted, removed, and modified by domain experts, such as nurses, based on cognitive and/or physical information regarding each specific resident. Summaries are generated at different levels of abstraction. First, state, which occurs at a single moment/frame in time, is reasoned about, followed by activity, which occurs over a time interval (seconds, minutes, hours, etc). The system yields results in a format (linguistic) that caregivers can more easily utilize as well as it is of use for adverse event detection. For example, "[Derek] is *on the ground* in the living room for *awhile* in the *mid afternoon*" or "[Derek] *has fallen* in the living room in the *early morning*".

Adequately segmenting humans from imagery over long time periods in general loosely constrained environments is still an area of active research. Many phenomena make an environment uncontrolled, e.g. changes in lighting, shadows, movement of non-human objects (books, trees, blinds, etc). To date, a large number of screen space change detection

Derek Anderson and Robert Luke, {dtaxtd,rhl3db}@mail.missouri.edu, are pre-doctoral biomedical informatics research fellows funded by the National Library of Medicine (T15 LM07089). All authors, including James M. Keller, kellerj@missouri.edu, are with the Electrical and Computer Engineering Department at the University of Missouri, Columbia, MO, 65211, USA.

978-1-4244-8126-2/10/$26.00 ©2010 IEEE

algorithms have been proposed to address, but do not solve, these problems. Examples include Gaussian Mixture Models [15], Eigen Backgrounds [16], and Wallflower [17].

In [14], we introduce a computer vision system that reliably segments and classifies humans in unconstrained indoor environments using stereo vision and color. This system is different from most because change detection is not carried out in image space, i.e. image silhouettes. Rather, segmentation and change detection is performed in three-dimensional volume element (voxel) space. A classifier is used to identify human voxel objects and they are used to improve background update. This silhouette-free approach is consistent with our prior silhouette work in the regard that its output, voxel objects, preserves privacy. Figure 1 shows the difference in stereo-vision acquired voxel scenes and the standard method of intersecting silhouette back-projections.

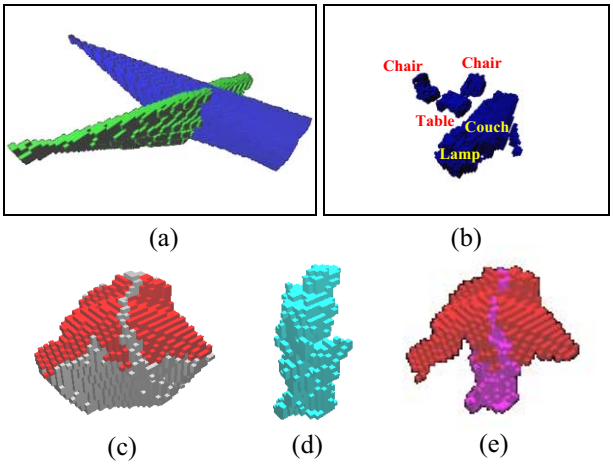

Fig. 1. Multiple approaches to voxel scene modeling. In (a), human image silhouettes are back-projected from the cameras. In (b), a full voxel scene is created using depth-based back-projection. In (c), multiple silhouette back-projections, e.g. (a), are intersected to acquire only the human. In (c), red is voxels visible by the camera and gray is non-visible back-projected voxels. In (d), cyan is a segmented stereo vision acquired human object and (e) shows the substantial difference in visible voxel sets (magenta is the visible voxel set for stereo-vision acquired voxel person).

The inference of many human actions require interactions with non-human objects, e.g. couches and chairs. To properly infer these behaviors, significant objects in a scene must first be labeled. Performing this by hand is tedious, time consuming and only works for static scene objects. This paper describes a technique to automatically segment and label important non-human objects in light of object movement using minimal user interaction.

II. Island Segmentation

First, a few stereo and voxel concepts need introducing. A voxel is specified according to its center location, $\bar{v}_{(i,j,k)}$, and axis *widths,* which are typically the same for all dimensions. Voxel width is important. The larger the voxel, the lower the object resolution, and the smaller the voxel, the

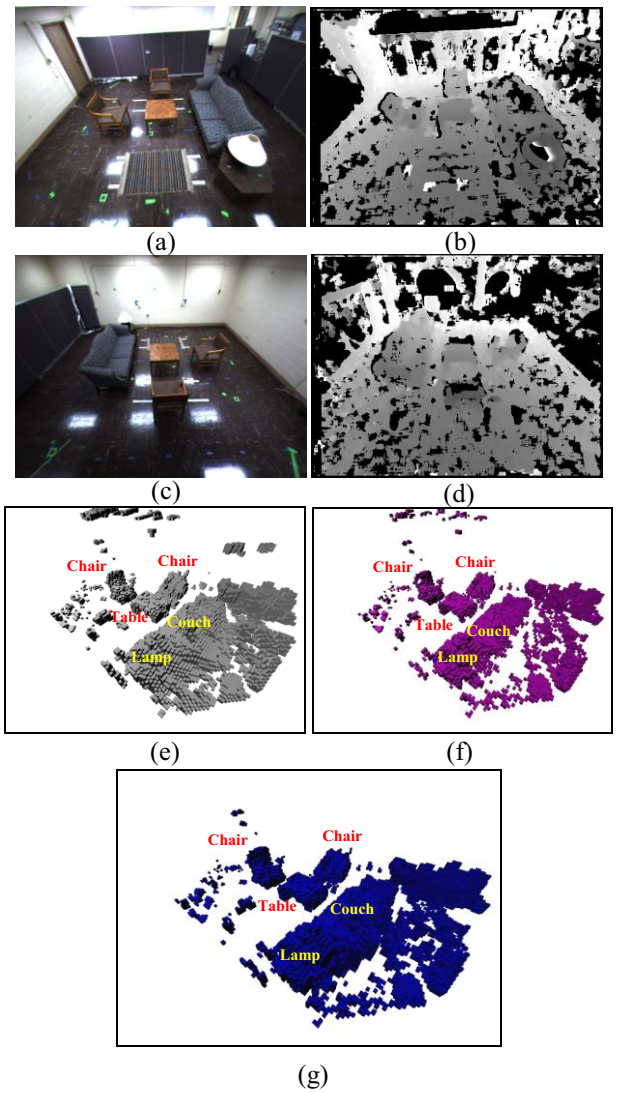

Fig. 2. Right images for (a) camera 1 and (c) camera 2 in a two stereo pair setup and their respective, (b) and (d), depth maps (where darker colors indicate closer depth). Image (e) is the intersected space, (f) is the visible shell of the intersected space, and (g) is the blanketed set.

greater the computational load and memory requirements. Voxel width needs to be determined for each application and environment. Each pixel in a camera is associated with a world-space view ray and view frustum [14]. The frustum is only calculated for the right camera in each stereo vision pair. A sorted (ascending) list, according to distance to the camera, of intersecting voxels, with respect to the view frustum, in voxel environment E, per pixel, per right camera in each stereo pair is built offline. At time t, each stereo vision pair, $1 < c \le C$, builds a depth map (shown in figure 2). Full details regarding the above can be found in [14].

The visible shell, $E_{c,V}$, is the set of voxels associated with a given depth map. A fast way of computing the visible shell is that for each ray in the right camera for each stereo pair, one uses the corresponding depth to index a single voxel. In the case of a frustum, the result is a set. The back-projection of a depth map is the additive solid modeling process of

selecting all voxels in E "behind" the visible shell, that is all voxels in a particular pixels view frustum behind a given depth value, is $E_{c,P}$. The fusion of stereo pair voxel scenes, for sake of environment refinement, is

$$E_P = \bigcap_{c=1}^{C} E_{c,P}, \quad (1)$$

$$E_V = E_P \wedge \bigcup_{c=1}^{C} E_{c,V}. \quad (2)$$

Human image silhouette extraction is far less robust when compared to the problem of correspondence in stereo vision [18]. Additionally, stereo vision is superior to the back-projection of image silhouettes in the regard that full voxel environments are built using stereo vision, which enables segmentation work such as this article. Because stereo vision results in point clouds, proximity and clustering of voxels can be factored into change detection versus solely relying on image-space. Our procedure is also superior to image silhouettes as it relates to shadows (does not change depth), illumination changes (addressed by correspondence), and occlusion [14]. With silhouettes, if an object is fully occluded in any camera view, then no voxel object will be constructed. However, in our approach, visibility and separation is only required in at least one camera pair.

A voxel operator of importance here is the blanketed set (shown in figures 2 and 3) [14],

$$E_B = \{\bar{v}_{(i,j,k)} \in E \mid \bar{v}_{(i,j,k)} \in \Omega(E_V), \bar{v}_{(i,j,k)} \in E_P\}, \quad (3)$$

where $\Omega(E_V)$ is the Umbra of E_V,

$$\Omega(E_V) = \{\bar{v}_{(i,j,k)} \in E \mid \exists \bar{g} \in E_V \text{ s.t. } g_k \geq v_k\}, \quad (4)$$

where $g_k \geq v_k$ is a check for existence of a visible shell voxel *above* $\bar{v}_{(i,j,k)}$ (with respect to world up direction). For a predominant downward viewing scene, i.e. cameras installed on the ceilings angled downward, the blanketed set is of particular utility in the regard of visible and non-visible back-projection error minimization [14]. These situations occur most often in non-ideal viewing conditions (which is the case generally as the dot product of intersecting voxel camera rays approach zero). Figure 3 is the blanketed set for a segmented human object in a non-ideal viewing location.

The system presented in this paper is a collection of algorithms. The first algorithm (A1) is stereo vision based voxel scene construction. In order to segment and monitor humans, the primary objective of our research, knowledge of a scene is required. Here, this knowledge comes in the form of an assumed human free voxel background set, E_Q. Algorithm 2 is change detection from E_Q. We present an adaptive algorithm in [14] for building and updating E_Q (steps A2.1 and A2.4). This algorithm addresses difficult scenarios, e.g. humans, or other moving objects, in the scene or lighting changes during E_Q construction and updating.

A1: ALGORITHM 1[14]
STEREO VISION FOR VOXEL SCENE CREATION

(1) Acquire images from all cameras in a scene
(2) Compute correspondence and acquire depth images
(3) Use the right cameras from each stereo pair to build the back-projected and visible shell sets ($E_{c,P}$ and $E_{c,V}$)
(4) Build the full intersected voxel space using Eq (1)
(5) Build the visible shell of E_P using Eq (2)
(6) Build the blanketed set E_B using Eqs (3,4)

A2: ALGORITHM 2 [14]
VOXEL SPACE CHANGE DETECTION

(1) Build a background model E_Q
WHILE NOT DONE
(2) Run algorithm 1
(3) Compute $E_F = E_B \wedge \bar{E}_Q$
(4) Update E_Q

Our current focus is the monitoring of single resident environments. If a scene contains multiple residents, it is assumed that help is available for adverse events, e.g. a fall. At each time step, we assume that the human is identified. Our stereo vision-based human segmentation work [14] is too long to reproduce for sake of compactness of this article. In summary, the algorithm operates on E_F and produces E_H (human object). Morphology is used first to clean up E_F, then connected components is used to find voxel islands. Next, a classifier, which utilizes features such as skin color, height, and head shape, is engaged to find humans from the set of islands. If a human is not found at a given frame, then we use color matching to see if any island is similar enough to a previously found human (in the case that the human classifier fails at a given frame).

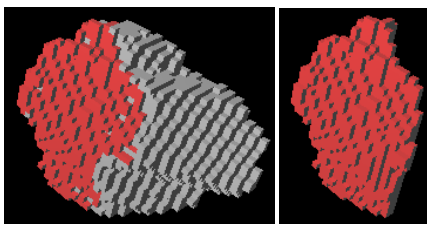

Fig. 3. Red is visible shell voxels and gray is non-visible voxels. The left object is the full intersected set and the right is the blanketed set. The great majority of gray voxels in the left image are incorrect (non-visible error). The blanket set on the right far more resembles the actual human.

New to this work is algorithm 3, the unsupervised segmentation of (assumed) non-human islands (shown in figure 4). Morphological opening, using a 3x3x3 kernel of ones, helps eliminate small islands of erroneous voxels. It also separates regions joined by a thin connection of voxels. Connected components, with 6-point connectivity, is then performed. At each time step, algorithm 3 yields a set of voxel islands, $I_m\,(1<m\le M)$. In addition, each island has an associated visible shell, $IS_m = E_V \wedge I_m$.

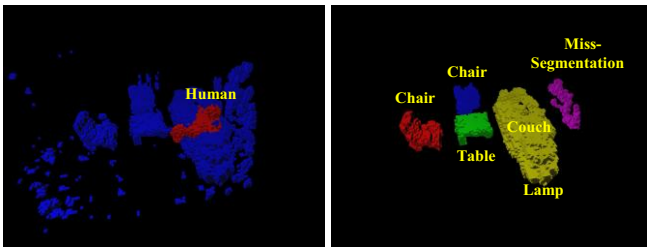

Fig. 4. (left) Voxel space with a human (shown in red) in it and (right) the same voxel space after the human is identified and removed [14] and voxel islands are found. In this figure, voxel islands are arbitrarily color coded.

III. Feature Extraction

Once voxel islands are extracted, labels need to be inferred. For our eldercare application, the assumption is that the scene is an indoor single resident environment and important large objects include furniture. Living rooms typically have couches, which are usually large volumetric objects that are about three feet tall, are longer in the width than length direction, chairs are generally present, which are about three feet tall, smaller in volume, etc. The point is that prior information, specifically linguistic, is present regarding a scene. Additionally, linguistic output is desired, e.g. "the scene contains one *tall big* couch". To take advantage of such knowledge, and produce such output, we propose one extracts features from islands, models these features using linguistic variables, and uses an appropriate fuzzy aggregation technique (fuzzy logic, aggregation operator, fuzzy integrals, etc) for label inference.

Currently, features extracted from island I_m include the following. The voxel-based volume is $VOL_m = |I_m|$, where $|\bullet|$ is the cardinality (number of elements) of a set. The volume is useful for linguistically describing objects in terms of words such as *small*, *medium*, *large*, etc.

The next feature is an islands *width* to *length* ratio in the x-y (world ground) plane. Let island I'_m be the set containing only the x and y components from I_m. The eigen vectors, $\vec{U}_{m,k}$, and values, $V_{m,k}$, for $k \in \{1,2\}$, are computed for the covariance matrix, $COV(I'_m)$,

---

A3: ALGORITHM 3
NON-HUMAN ISLAND SEGMENTATION

---

(1) Build a background model E_Q

WHILE NOT DONE

(2) Run algorithm 1 and acquire a voxel scene

(3) Find the human (E_H) and remove it ($E_W = E_B \wedge \overline{E}_H$)

(4) Segment islands in E_W

(a) Perform morphological opening

(b) Use connected components algorithm to acquire the set of voxel islands

(c) Discard all voxel islands with too few of points

(5) Update E_Q using E_W

$$COV(I'_m) = \frac{\sum_{\forall \vec{v}_a \in I'_m} (\vec{v}_a - \vec{\mu}_m)(\vec{v}_a - \vec{\mu}_m)^T}{|I'_m| - 1}, \qquad (5)$$

where $\vec{\mu}_m$ is the mean of I'_m. Assuming the eigen values are sorted in descending order, the eigen ratio used here is

$$ER_m = \frac{V_{m,1}}{V_{m,2}}. \qquad (6)$$

This feature is used to abstractly describe the shape of an object. For example, it should be *near one* for a chair, whose length and width should be approximately equal, and *big* for a couch, whose length is generally larger than width.

IV. Object Labeling

Labeling is important for at least two reasons. First, as it relates to human computer interaction, namely linguistic summarization of scenes, labeling yields results consistent with a human's vocabulary. Second, these labels are of use computationally. A system can use the label confidences to help infer different types of activity [4][5]. For example, "[Derek] is in a sitting pose" and "[Derek] is on the couch", thus the inferred human-object interaction might be "[Derek] is sitting on the couch".

The features outlined above are the basis for recognizing the common living room objects chair and couch. These are often of importance in terms of recognizing different types of falls and determining daily information such as the overall amount of activity. For example, if someone is lying on their couch all day long, this might be an indicator of little overall daily activity. Additionally, the above objects need to be monitored in the case that an elder falls while sitting down (or getting up respectively), possibly because of being disoriented or decline in functional ability. Living room quarters are analyzed here because it is less common to find

video equipment installed in bedrooms or bathrooms for obvious privacy reasons. Fuzzy set theory, introduced by Lotfi A. Zadeh in 1965, is an extension of classical set theory [20]. One of the more well known branches of fuzzy set theory is fuzzy logic, a powerful framework for performing automated reasoning, was introduced in 1973 [21]. In this work, features and labels are linguistic variables. A linguistic variable takes on fuzzy values (terms) that are represented as words and modeled as fuzzy subsets of an appropriate domain. An example is the linguistic variable voxel volume of an object, which can assume the terms *small*, *medium*, and *large*.

Fuzzy sets are modeled herein as trapezoidal membership functions, which are defined with respect to four ordered points, $\{a,b,c,d\}$. The linguistic variables and terms for object volume and eigen ratio are reported in Table 1.

TABLE I
LINGUISTIC VARIABLES AND TERMS

Variable	Term	Trapezoid Parameters
Volume	*Small*	{0,1000,2500,3500}
	Large	{4000,7000,15000,15000}
EigenRatio	*Little*	{1,1,2,4}
	Big	{2,6,10,10}

We currently use a straightforward labeling procedure. If the volume is *large* and the eigen ratio is *big* then the object is a couch. Respectively, if the volume is *small* and the eigen ratio is *little* then the object is a chair. We could use a fuzzy inference system, but there are only two rules, and one respectively for each desired output. Membership is presently calculated using a t-norm ($\wedge$),

$$\mu^{\text{couch}} = \pi^{\text{volume}}_{\text{large}}(\text{VOL}_m) \wedge \pi^{\text{eienratio}}_{\text{big}}(\text{ER}_m), \quad (7)$$

$$\mu^{\text{chair}} = \pi^{\text{volume}}_{\text{small}}(\text{VOL}_m) \wedge \pi^{\text{eienratio}}_{\text{little}}(\text{ER}_m). \quad (8)$$

Remember, algorithm 3 is designed to return only large (assumed) non-human islands and the environment is a living room. Thus, there are typically few objects that will ever have features similar to a couch or chair. The inference of objects can be made more sophisticated (fuzzy logic, fuzzy integrals, etc) in future work if the problem requires more objects that are harder to distinguish.

V. OBJECT SEGMENTATION IN CLUTTERED ENVIRONMENTS

The primary difficulty resides in voxel islands that are not single objects. In this section, we present an island-object matching procedure and genetic algorithm (GA) for approximating the location of a set of known reference voxel objects in clutter. Thus, if we know a scene contains two chairs and a couch, what are the locations of these objects at any particular instance in time?

An assumption is that the system knows ahead of time what objects are relevant to monitoring human activity in a particular scene. Additionally, we assume these objects are not removed from the scene at any moment. In future works, we will address ways to relax this assumption. For our application (eldercare) this assumption, i.e. a living room always contains a couch, chairs, etc, is rational.

A4: ALGORITHM 4
SEMI-SUPERVISED VOXEL SCENE INITILIZATION

(1) Run algorithm 3 and get a single voxel environment
FOR $m = 1$ to M
  (2) Extract features from I_m
  (3) Compute fuzzy label memberships for I_m
  (4) Report label (term) with maximum membership value
  (5) If user indicates the island is of any importance
    (a) If the object needs further segmentation
      (i) Have user lasso objects, create blanketed depth back-projections, label objects, and record objects
    (b) Else, if the label is correct, record the object, otherwise ask the human to label the object

The predicament is that algorithm 3 regularly fails to segment objects in cluttered environments. On the opposite extreme, it is too time consuming, and a potential threat to resident privacy, for a human to be in the loop continuously. What we propose is a middle ground. That is, the automated system takes an initial guess at the scene, and then the human verifies, or corrects, the outcome. Thus, if one voxel island is actually multiple objects, the human will break the objects apart (shown in figure 5). We propose using the right images from all stereo pairs. In the case of a segmentation error, the human lassos the image sub region in each right image that is an object, those regions are depth back-projected, and the system either re-classifies the object or the human provides the label. This procedure is only performed at system initialization. At future time steps, the system automatically segments and locates important objects.

Algorithm 4 (system initialization) yields N blanketed objects, $\{O_1,...,O_N\}$, and visible shells, $OS_n = E_V \wedge O_n$. Algorithm 5 approximates the current location of reference objects at each future time step. Step A5.5 is early identification of all islands that are known reference objects (i.e. islands free of clutter). We begin by using knowledge

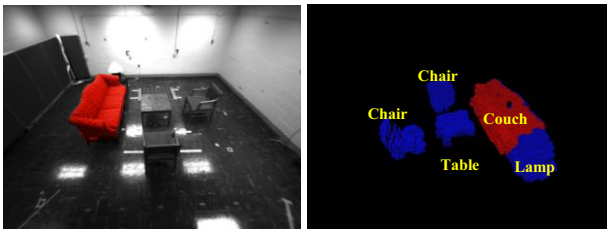

Fig. 5. (left) Users image mask (shown in red) for the couch. The right image is the depth-based back-projection (couch, which is initially part of a table, lamp, and couch island, is shown in red).

of our specific environment. Furniture is not generally stacked on top of other objects. Additionally, furniture has a specific up direction, meaning a couch is never flipped over or located on its side. Therefore, object rotation is restrict to $[0,360]$, indicating the rotation amount about the z-axis (world up direction). An object is centered at an island according to their respective means. We then *sweep*, i.e. perform a check at a user specified angle offset amount, such as every one degree, a set of candidate rotations. The angle that makes an object *best match* an island is selected. The metric for object-island matching is:

$$m(O_n, I_m) = s(O_n, I_m)c(OS_n, IS_m), \quad (9)$$

$$s(O_n, I_m) = \frac{|O_n \wedge I_m|}{|O_n|}, \quad (10)$$

$$c(OS_n, IS_m) = \frac{\sum_{i=1}^{|OS_n|} \Phi(OS_{n,i}, IS_m)}{|OS_n|}, \quad (11)$$

where $\Phi(OS_{n,i}, IS_m)$ is 0 if the ith voxel in OS_n, $OS_{n,i}$, has no corresponding voxel, $IS_{m,k}$, in IS_m (meaning no object-island voxel pair can be found that have the same voxel center location). Otherwise, the result is $1 - d(OS_{n,i}, IS_{m,k})$. The distance, $d(OS_{n,i}, IS_{m,k})$, of two shell voxels is the Euclidean distance of their RGB color values. Voxel shell color is determined by averaging the image space pixel RGB colors for each visible shell voxel (many image space view vectors can intersect a single visible shell voxel) [14]. Thus, $s(O_n, I_m)$ is a measure of object shape/volume similarity and $c(OS_n, IS_m)$ is a measure of shell shape/color similarity. If $m(O_n, I_m) > \vartheta$, $\vartheta \in [0,1]$, then I_m is labeled as O_n.

GAs are an optimization procedure inspired by evolution. The general format of a GA is outlined in algorithm 6 [22]. The optimization task must first be coded into a chromosome representation. After A5.5, for each object not yet identified, there are $N' \leq N$ of them, the goal is to approximate the location of remaining objects in remaining islands. Let K be the set of recognized objects from A5.5 and $U = O - K$. The non-stacking object assumption is used to restrict the initially unbounded full three-dimensional space for translation of an object's mean to a single, or local set of, two-dimensional x-y plane(s) relative to an objects mean. The search space is reduced even further by restricting it to just the island voxel set in a current object's mean x-y plane(s). Rotation is a free parameter, but it is not included in the chromosome. We search for the best rotation based on the current mean offset, instead of adding rotation to the chromosome and further complicating (expanding) the size of the search space.

Chromosome C_j is of length $|U|*2$, i.e. N' objects and their x and y mean translations. Fitness is a combination of island-object matching and a penalty for volume overlap among the current locations of all objects. Objects should not intersect in the solution. The fitness of C_j is

$$f(C_j) = \frac{T_1}{2} + \frac{T_2}{2} - T_3, \quad (12)$$

$$T_1 = \frac{\sum_{k \in K} s(O_k, E_W)}{2|K|} + \frac{\sum_{k \in K} m(O_k, E_W)}{2|K|}, \quad (13)$$

$$T_2 = \frac{\sum_{u \in U} s(O_u, E_W)}{2|U|} + \frac{\sum_{u \in U} m(O_u, E_W)}{2|U|}, \quad (14)$$

$$T_3 = \max_{\substack{\forall(a,b) \\ a,b \in \{1,\ldots,N\} \\ a \neq b}} \left(\frac{|O_a \wedge O_b|}{|O_a \vee O_b|} \right). \quad (15)$$

Elitism is used in GA parent selection. Standard two point crossover, with probability λ_1, is used and the crossover point must be an event offset. Four mutation operators are

---

A5: ALGORITHM 5
NON-HUMAN VOXEL OBJECT SEGMENTATION

---

(1) Build background model E_Q

WHILE NOT DONE

(2) Run algorithm 1 and produce the blanketed set (E_B)

(3) Segment the human, E_H, and remove it from

(a) Blanketed set, $E_W = E_B \wedge \overline{E}_H$

(b) Visible set, $\widetilde{E}_W = (E_B \wedge E_V) \wedge \overline{E}_H$

(4) Segment islands for E_W (A3.4)

(5) Early stage identification of islands that match known objects, i.e. according to $m(O_n, I_m) > \vartheta$

(6) GA for finding remaining object locations (A6)

(7) Update E_Q

A6: ALGORITHM 6 [22]
GENERAL STRUCTURE OF A GENETIC ALGORITHM

---

(1) Set generation counter q = 1

(2) Create initial generation $P_1 = \{C_1, ..., C_J\}$

UNTIL CONVERGENCE (e.g. maximum iterations)

(3) Evaluate the fitness of each chromosome

(4) Increment the generation counter q = q + 1

(5) Select parents from P_{q-1}

(6) Recombine selected parents by cross-over to form P_q'

(7) Mutate offspring P_q'

(8) Select the new generation P_q from the previous generation P_{q-1} and the offspring P_q'

performed. The first mutation operator, with probability λ_2, is pseudo-random selection of all object mean offsets in the constrained plane(s) of the blanketed island set. This operation encourages the most amount of exploration. The second operator, with probability λ_3, involves keeping $J-1$ objects fixed and pseudo-randomly *jumping* (random position assignment) one object. The idea is to keep a potentially good solution intact and search for the position of only one object. The third operator, with probability λ_4, is the shifting around of a single object for position refinement. That is, an offset of $U[-1,1]\lambda_5$ is calculated, where λ_5 is a user specified maximum shift amount and $U[-1,1]$ is a random number generated from a uniform probability distribution over $[-1,1]$. The translated position is restricted to a location from the blanketed object mean plane(s). If a location is generated outside of this set, then the nearest voxel in the blanketed set is selected. The last mutation operator, with probability λ_6, is a variation on the third operator. Instead of searching a large window for a good translation, this operator searches a very local window for fine position refinement. That is, it is the shifting of one object by a small $U[-1,1]\lambda_7$ amount. Thus, mutation operator three moves an object into new areas for searching, while operator four helps refine an already good solution.

VI. Linguistic Summarization of Scenes

At least three types of linguistic summarization of scenes, for human understanding as well as computational purposes, e.g. fall detection [3][4][5], exist.

1. *Objects*: The (spatial) segmentation and labeling of sub-regions in a scene, e.g. couch, chair, human, etc. An example is "the scene contains one *tall* human, two *small* chairs and one *large* couch".

2. *Human activity*: The (temporal) summarization of human behavior [4]. An example is "[Derek] is *exercising* in the living room for a *moderate amount of time* in the *early morning*".

3. *Human-object interaction*: The (spatial and temporal) summarization of the interaction between humans and scene objects. Examples include: "[Derek] is *on the couch*" or "[Derek] is *lying on the couch* in the living room *taking a nap* in the *late afternoon*".

The proposed system is a series of summarizations. The first step is human segmentation [14], followed by non-human object segmentation (this article). Both steps are the spatial reduction of C stereo vision camera pairs, each of image size A rows by B columns, thus $2*A*B*C$ pixels. For example, a 640x480 two stereo vision camera system contains 1,228,800 pixels. Segmentation reduces this large stream of pixels into a significantly smaller number, e.g. one human, two chairs, and one couch. The combination of [14], this work, and [3][4][5] make the listed three summarization (spatial and temporal) categories possible. Prior [5], we were greatly restricted to assuming that object locations were known and that objects did not move.

VII. Preliminary Results

Data was captured in the Computational Intelligence Research Laboratory at the University of Missouri. A lab is used due to the severity of the activities being analyzed, e.g. falls [4][23], and in particular the target elderly population. Here, we show preliminary results for summarization of this lab environment, which is setup to resemble a living room. The proposed system is able to address difficult situations, such as surfaces that appear to be moving, e.g. a monitor or television, namely due to stereo vision, as well as actual moving objects. The user indicated, via A4, that the scene depicted in figure 2 contains two chairs and a couch relevant to tracking human activity. In general, the scene has two chairs and a coffee table that are well separated from other objects. There is also an end table, lamp, and couch that are cluttered (i.e. island that is not a single object).

Two hundred voxel scenes, collected in a dynamic and complex environment, are analyzed here. That is, clutter is present, objects are moving, and lighting is constantly undergoing change. A difficulty resides in the fact that no ground truth exists. That is, the real-world volume that objects occupy is not known. For one or two objects it might be possible to approximate the ground truth; however this is not the case for scenes with many moving objects. A human is required to verify object matching. In addition to the human"s qualitative assessment of object matching, we also provide a quantitative measure for the amount of voxel object overlap. That is, overlap between an image-space segmented voxel object and the final GA learned voxel objects. This measure helps support the user"s qualitative answer and informs us about the accuracy of the GA in terms of object positioning. The GA parameters used here are $\lambda_1 = 0.3$, $\lambda_2 = 0.5$, $\lambda_3 = 0.3$, $\lambda_4 = 0.1$, $\lambda_5 = 25$, $\lambda_6 = 0.1$, $\lambda_7 = 1$, population of size 100, and 50 generations. The (qualitative) result of algorithm 3, island segmentation, is shown in table 2 and the (qualitative and quantitative) results for algorithm 5 are shown in table 3.

TABLE II
Algorithm 3 For Island Segmentation of Relevant Objects

	Couch	Chair 1	Chair 2
Island is an Object	0%	100%	100%

TABLE III
Algorithm 5 For object Segmentation of Relevant Objects

	Couch	Chair 1	Chair 2
Qualitative	100%	100%	100%
Quantitative	96%	90%	86%

The result of linguistically summarizing the scene, found by combining this articles work with our previous activity summarization work [3][4][5][14], is shown below. For sake of redundancy and article length, only a selective subset of the total number of summaries are reported.

"The scene contains two *small* volume chairs and one *large* volume couch"
"Bob is *walking fast* in the living room for a *short amount of time*"
"Bob is *sitting on the couch* for a *brief time*"
"Bob is *walking slowly* in the living room for a *moderate amount of time*"
"Bob is *sitting in a chair* for a *brief time*"
"Bob is *walking slowly* in the living room for a moderate amount of time"
…

VIII. Conclusion

This article extends our prior linguistic summarization of human activity and stereo vision-based voxel human segmentation work to non-human object segmentation, scene description, and human-object summarization. Specifically, this problem is analyzed in the context of eldercare and a home environment, namely living room quarters. This research is of importance as it relates to describing the contents of scenes to humans as well as its inclusion in automated activity analysis. A genetic algorithm is used to search for the location of objects at an arbitrary moment in time in light of clutter. Encouraging preliminary quantitative and qualitative results were demonstrated.

Future work includes collecting and exhaustively analyzing a larger data set from the home of elders, in light of different types of objects and varying degrees of clutter. We are also working on extending the fuzzy labeling and summarization component for the case of a larger set of more complex objects and richer descriptions (i.e. inclusion of color, texture, etc). Additionally, we are looking into the matter of relaxing the semi-supervised first step of system initialization (by using an object database or segmentation and tracking over time for dynamically building a database) and we are looking for ways to relax the assumption that objects cannot exit or enter a scene after the system has started. On a final note, there is the matter of extending the activity recognition framework for addressing richer types of human-object description and analysis.

References

[1] S. Wang, M. Skubic, Y. Zhu, "Activity Density Map Dis-similarity Comparison for Eldercare Monitoring," *IEEE Eng. in Medicine and Biology Society Conf.*, Minneapolis, MN, Sep 2-6, 2009.

[2] I. Sledge, J.M. Keller, G. Alexander, "Emergent Trend Detection in Diurnal Activity," *IEEE Eng. in Medicine and Biology Society Conf.*, pp. 3815-3818, Vancouver, August, 2008.

[3] D. Anderson, R. H. Luke, J. M. Keller, M. Skubic, M. Rantz, M Aud, "Modeling Human Activity From Voxel Person Using Fuzzy Logic," *IEEE Transactions on Fuzzy Systems*, vol. 17, no. 1, 2009, pp. 39-49.

[4] D. Anderson, R. H. Luke, J. M. Keller, M. Skubic, M. Rantz, M Aud, "Linguistic Summarization of Video for Fall Detection Using Voxel Person and Fuzzy Logic," *Comp. Vision and Image Understanding*, Vol. 113, pp. 80-89, 2008.

[5] D. Anderson, R. H. Luke, J. M. Keller, M. Skubic, "Extension of a Soft-Computing Framework for Activity Analysis from Linguistic Summarizations of Video," *IEEE Intl. Conf. on Fuzzy Systems (FUZZ-IEEE), WCCI*, pp. 1404-1410, Hong Kong, June 2008.

[6] 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," *Intl. Congress of the European Federation of Medical Informatics*, pp. 45-50, 2006.

[7] 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," *IEEE EMBS Intl. Special Topic Conf. on Information Technology in Biomedicine*, pp. 1-4, 2006.

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

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

[10] T. Martin, B. Majeed, L. Beum-Seuk, N. Clarke, "Fuzzy ambient intelligence for next generation telecare," *IEEE Intl. Conf. on Fuzzy Systems*, vol. 15, 2006, pp. 894–901.

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

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

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

[14] R. H. Luke, D. T. Anderson, J. M. Keller, "Human Change Detection in Voxel Space using Stereo Vision," under review by *Computer Vision and Image Understanding*, 2010

[15] 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.

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

[17] K. Toyama, J. Krumm, B. Brumitt, B. Meyers, "Wallflower: principles and practice of background maintenance," *IEEE Intl. Conf. on Computer Vision*, vol. 1, 1999, pp. 255–261.

[18] M.Z. Brown, D. Burschka, G.D. Hager, "Advances in Computational Stereo," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 25, pp. 993-1008, 2003.

[19] D. Anderson, R. H. Luke, E. Stone, and J. M. Keller, "Fuzzy Voxel Model for Human Activity Analysis," *Intl. Fuzzy Systems Association* (IFSA), Lisbon, Portugal, July 2009.

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

[21] 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.

[22] Engelbrecht, A. (2007) *Computational Intelligence: An Introduction,* Second Edition. John Wiley, Chichester.

[23] Popescu M, Li Y, Skubic M, Rantz M, "An Acoustic Fall Detector System that Uses Sound Height Information to Reduce the False Alarm Rate," *Intl. IEEE EMBS Conf.*, Vancouver, British Columbia, Canada, August 20-24, 2008.