A system for change detection and human recognition in voxel space using the Microsoft Kinect sensor

T. Gill and J. M. Keller
Electrical and Computer Engineering
University of Missouri
Columbia, MO
(tsg334 and kellerj)@missouri.edu

D. T. Anderson
Electrical and Computer Engineering
Mississippi State University
Mississippi State, MS
anderson@ece.msstate.edu

R. H. Luke III
U.S. Army Night Vision and Electronic Sensors Directorate – Countermine Div.
Fort Belvoir, VA
robert.h.luke@us.army.mil

Abstract—**Previously, we put forth a new computer vision system for indoor well-being monitoring of elderly populations based on the use of multiple stereo camera pairs. That approach involves combining the strengths of image space with three dimensional volume element (voxel) space techniques. However, that system is fundamentally limited because it is based on color imagery from visible light cameras. In this article, we extend our prior research and consider a new, inexpensive infrared depth camera device, the Microsoft Kinect. Advantages, such as the ability to operate 24-7 in low-to-no light conditions, and shortcomings are detailed. In addition, we discuss necessary algorithmic extensions to our mixed image and voxel space framework for the Kinect sensor. Experiments are performed in a laboratory designed to resemble an elders living quarter. Vision findings are evaluated using our prior high-level linguistic summarization of human activity work. Preliminary results indicate that the Kinect sensor does indeed work in a wider range of operating conditions and it can produce activity descriptions that match that of a human.**

Keywords; human activity recognition, voxel space, infrared depth camera, Microsoft Kinect sensor.

I. INTRODUCTION

The goal of this research is low-cost, real-time passive monitoring technologies for the elderly. Technology has the potential to benefit mankind by enabling elders to live longer, healthier, independent lives. In order to realize such a future, we must design new sensors, algorithms and theories in cross-disciplinary fields such as Engineering, Nursing, Physical Therapy, etc. Of particular interest to the elderly is adverse event detection, e.g., fall detection [1], risk assessment, e.g., fall prediction [2-4], human gait [5-7], as well as detecting the early onset of illness and/or functional decline. Ideally, sensor measurements would be obtained passively, in the course of normal daily activity [8]. Computer vision is one technology that has the potential to address the topics listed above. In [9], we report elderly focus group findings that indicated that elders are interested in video technologies as long as the video data is not stored, not viewed, and privacy protection mechanism are put into place (e.g., image silhouettes [1,10,11]).

While video technology is promising, a combination of real world and theoretical issues must first be tackled. For example, computational vision has been around since at least the 1950's. While advancements have been made in focused areas such as face detection [12-14] and automated target recognition [15-18], low-level core topics such as change detection and activity recognition in complex, loosely constrained environments is still unsolved. While the majority of humans perform this task with little apparent effort, rigorous (mathematical) explanations remain in the infancy stage. While we discuss applications to eldercare, this work stands to influence a wider field of human monitoring areas. One example is surveillance, a vital activity for security in a number of locations including airports, banks, military installations and other public buildings [19-21].

Before automated video monitoring becomes a reality, cost must be tackled. In [22,23], we present a multi-stereo camera pair computer vision system. At a minimum, two stereo camera pairs are required. However, quality stereo vision cameras still remain over a thousand dollars. The price of these sensors must drop considerably before adoption can start to become a reality (at least in the eldercare domain). In this article, we consider a new imaging sensor, the Microsoft Kinect. The Kinect is based on software technology developed by Rare and an infrared (IR) depth camera technology developed by the Israeli developer PrimeSense. The economic angle of this sensor is its low cost, $150 USD. However, cost is not the only benefit of the Kinect. Herein, we discuss sensing and computer vision advantages of the Kinect versus a passive sensor that detects visible light.

The remainder of the article is organized as such. In section II, our prior work in the area of multiple stereo camera pair change detection and human recognition is summarized. In section III, we discuss the Microsoft Kinect sensor and focus on its strengths and weaknesses in the context of well-being monitoring. Next, we describe extensions to our prior work for the Kinect. This is followed by summarization of prior stereo vision system performance and its comparison to the current proposed work in the context of human activity analysis.

II. METHODS: PRIOR WORK

In [22,23], we present a new change detection and human recognition system for complex, loosely constrained indoor environments. That system is designed around a combination of image and volume element (voxel) space computer vision techniques. We showed that the system acquires more accurate results than single camera silhouette techniques while operating in a wider range of scenarios. However, as discussed, the stereo vision sensor is fundamentally limited. It, like most, is based on collecting and processing imagery acquired by visible light cameras. In addition, the proposed vision framework is still, at the moment, too expensive to be deployed in a wide area for elderly populations with low-to-no income.

A. *Benefits of Prior Well-Being Video Monitoring System*

The work detailed in [22,23] has the following advantages. We summarize our contributions and put them in the context of current deficiencies in the state-of-the-art. It is important to know which, if any, areas are strengthened or weakened by the work described herein. Again, refer to [22,23] for additional explanation regarding the following summarized topics.

- *Real-time voxel modeling of an environment*

The system described in [22,23] yields full three dimensional voxel solid models for an entire environment in real-time. This is different from prior silhouette-based work or point cloud representations in the case of stereo vision.

- *Robustness to illumination changes and shadows*

Image space approaches do not intrinsically address significant and abrupt changes in lighting. They instead generally include techniques to explicitly identify lighting changes and shadows. These results are then factored into the other parts of the vision system. Some adapt background models, others explore different color models/spaces, and others try to build several models that describe a range of lighting possibilities and then determine which is the most appropriate to the current setting. Each approach has significant shortcomings and are ultimately ill-equipped to scale to real-world phenomena.

In contrast to these image space processing techniques, depth values produced from passive stereo vision cameras is robust to changes in lighting and shadows. Image space algorithms that use depth information instead of color have shown improved results. Extending the use of depth to a full three-dimensional voxel space further resolves the accuracy of scene modeling by allowing the fusion of multiple stereo pairs [22,23].

- *Human classification using both image and voxel space*

A large field of research exists related to human identification in video. Face detection is the probably the most prevalent of these techniques [12-14]. Some approaches are accurate and fast enough to be implemented in consumer digital cameras and assist with tasks such as automatically setting exposure or focus. Unfortunately, the majority of these techniques require the face to be aimed at the camera and have a fairly substancial number of pixels over the face area. Neither of these attributes will be guaranteed in our (eldercare) setting.

In [22,23], human detection is based on the design and fusion of weak classifiers. Specifically, skin color from image space and head shape and height from voxel space is utilized. These features are readily available and their fusion improves the robustness of finding the head of any standing person under a wide spectrum of conditions.

- *False alarm reduction in change detection output*

Image space-based change detection outputs any change from the background model whether it is human or nonhuman. Many higher level systems, e.g. human activity recognition, require only human change detection and segmentation. Because our system is able to track a person and analyze objects in terms of three dimensional shape and human characteristics, nonhuman objects can be better segmented from the human and not classified as change when moved. An example of this is when a person moves a chair. Both the human and the chair have moved. However, we demonstrate that the chair can be quickly differentiated and in most cases removed from the change detection output.

- *Improved robustness with respect to occlusion*

Using only a single camera and image space, one is unable to directly construct correct objects as the intersection of voxel spaces when occlusion is present in one or more cameras. An example is the loss of a person's legs when a chair or table is occluding them in one or more cameras. This is a common scenario in indoor living quarters and it can drastically impact subsequent activity analysis. The procedures detailed in [22,23] can segment drastically improved full three dimensional models in the presence of occlusion. The only constraint is that volumes be separable in at least one stereo camera pair. This improvement is due to the use of stereo vision and the fact that we construct entire voxel environments, a proposed blanketed set operation for downward angled camera viewing, and the way in which we address change detection in voxel space.

B. *Summarization of Prior Computer Vision Algorithms*

In order to familiarize the reader with our prior work, made publically available at [22], and provide a context for topics discussed in subsequent sections, a high-level algorithmic overview of the work in [22,23] is detailed in algorithms 1 and 2. The next few sub-sections expand on these two algorithms.

Alg. 1: Multiple Stereo Calibration (Preprocessing) [22,23]

(1) Find lensing parameters and perform epipolar rectification
(2) Find transformation matrix of stereo camera 2 to camera 1 space
(3) Find transformation matrix from stereo camera 1 to world space
(4) Create voxel-pixel list for each pixel of right camera in each stereo pair

Alg. 2: Algorithmic System Overview (Runtime) [22,23]

WHILE NOT DONE
// Stereo vision and voxel reconstruction
(1) Collect images from all stereo cameras
(2) Build individual voxel spaces for each camera pair
(3) Build intersected (global) voxel space
(4) Build blanketed set
// Change detection and volume segmentation
(5) Remove background voxels from blanketed voxel space
(6) Segment objects
// Human detection
(7) Build color histogram for all objects
(8) Find human using head shape (with height restriction) and skin color or color histogram similarity
(9) Update human color histogram
// Background update
(10) Remove human voxels from global voxel space
(11) Update background using nonhuman blanketed space

C. *Stereo Vision for Real-Time Voxel Scene Construction*

In algorithm 1, the (pre-processing) calibration of multiple stereo camera pairs is outlined. Namely, a set of transformation matrices are found to take information from each camera into each other and ultimately a global world space. Additionally, each camera is responsible for constructing a set of per-pixel voxel lists. That is, for each image pixel, the set of world space voxels that intersect a view frustum (i.e., truncated pyramid) extruding outward from the pixel is found. This is an enabling step for subsequent real-time voxel world construction. At a later moment in time, once all stereo images are collected and correspondence is computed for each stereo vision pair, voxel pixel lists can be indexed given a current depth image. All later voxel operations break down into relatively simple set theoretic calculations that can be carried out in parallel.

At run-time, each stereo vision camera pair collects their images and depth maps are constructed. For each pixel, the set of voxels behind the current depth is found (i.e., those voxels in the pixel-voxel list with a distance greater than the current depth). For a single camera, these refined pixel voxel sets are combined (their union is calculated). The individual per-stereo camera pair voxel sets are then intersected to create a single global voxel world (shown in figure 1 (c)). In [22,23], we detail a set theoretic operation based on the visible shell. The visible shell is a voxel set that intersects the stereo point clouds as well as also intersects the multi-camera global voxel world. Next, we calculate the umbra in the world down direction using the visible shell and global voxel world. This result is then post-processed using mathematical morphology, namely opening (figure 1(d)). The final result is a significant reduction in both non-visible back-projection error as well as noise.

D. *Change Detection and Volume Segmentation*

The result of the prior section is a single voxel scene at a specific moment in time from the standpoint of multiple stereo camera pairs. In order to discover humans, we combine change detection with automatic voxel region segmentation. We build and maintain a background (non-human) model. Specifically, this model is a set of per-voxel probabilities that describe the likelihood that a voxel belongs to the background. In [22,23], we describe how to initially estimate a model and subsequently how to use the results of our system, i.e., all non-human voxels, to update the voxel probability background model.

At each new moment in time, the background model is *hardened* and used to remove all (assumed) static background voxels. The result is an approximation of recent change. Next, a set of mathematical morphology operations is used in order to segment the change detection results and yield a candidate set of potential human islands (one or more joined voxel objects). Like most adaptive online approaches, this system feeds back prior system decisions.

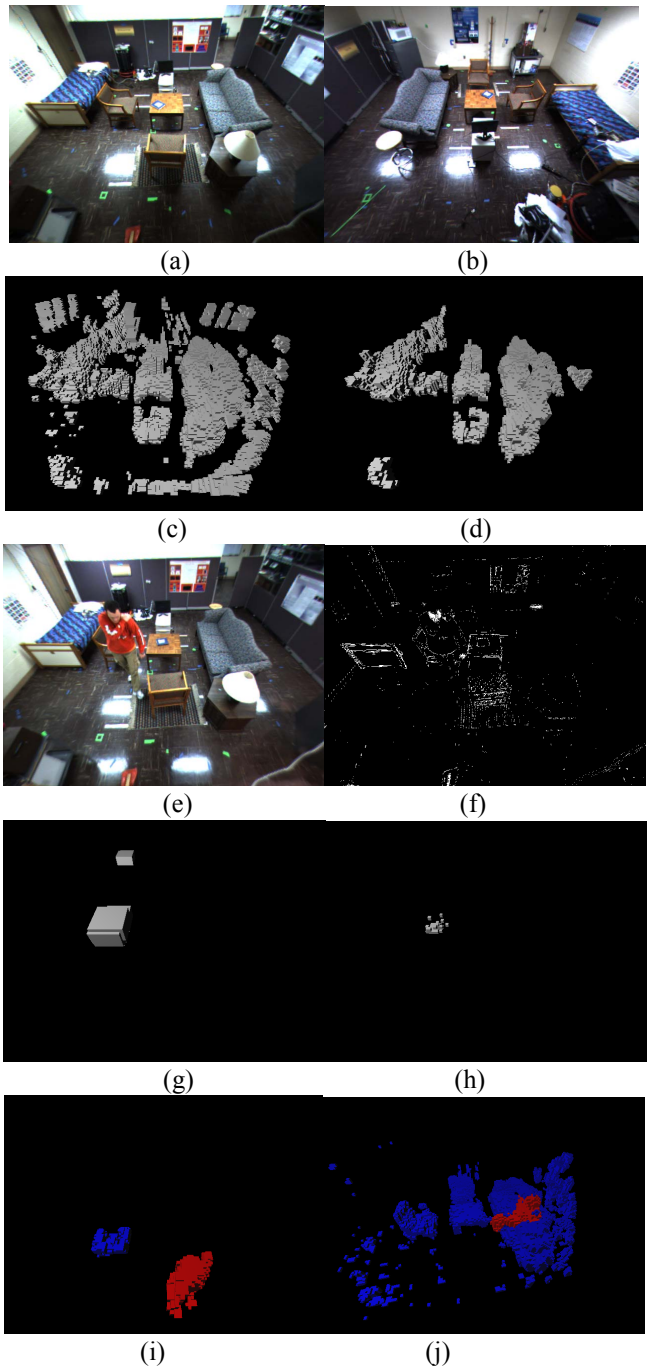

Fig. 1. (a) and (b) are raw images from two stereo vision camera pairs. (c) and (d) are the initial and refined combined multi-camera voxel spaces. (e) and (f) are the weak skin classifier in image space and the corresponding voxels. (g) and (h) are kernel head detection voxels and intersected head/skin voxel regions. (i) is an example segmentation result, where red is human and blue is recently moved non-human objects. In (j), a human is sitting on the couch, where red is the human and blue is the rest of the voxel scene.

E. *Human Detection*

In the prior sub-section, we described change detection and automatic region segmentation. From these results we look for humans. The majority of existing techniques require the face to be significantly aimed at the camera and have a relatively large number of pixels over the face area. Neither of these attributes is guaranteed in our (eldercare) setting. We instead engage a set of weak classifiers in both image space as well as voxel space to find people. Namely, a weak skin detector in image space is used and all voxels in the visible voxel set corresponding to the skin set if found (shown in figure 1(e)-(f)). In addition, a three dimensional voxel kernel head detector is engaged (shown in

figure 1(g)). The results of the skin and head finder are intersected in voxel space (shown in figure 1(h)). Any island segmented earlier from change detection results that intersects this set is selected. If that object greater than a minimum height, then the system calls it human (shown in figure 1(i)-(j)). This system assumes that a person will walk into the scene, or they stand up and move around at some point.

Subsequently, in [22,23] we detail a way to build a human color descriptor and a similarity measure (using volume range and a color comparison) for detecting future instances of that object (human) when it is not detected by our initial people finder. For example, if someone walks over to a couch then sits down, their volumetric region will be segmented by our change detection system but not identified by our combination of weak classifiers in image and voxel space. However, if that person was upright and walking around at some point, then its color descriptor was found. That information can now be used to find the person in the case that they are not found otherwise. Again, in [22,23] the specifics of this methodology are detailed.

F. World Voxel Model and Updating

We have already partially alluded to the following. In order to build a background model and detect future change and find quality automatically segmented voxel regions, updating rules are needed to feed back results from prior time steps. In [22,23], we begin by considering all current non-human voxels in the blanketed set. The background probabilities are updated using this information. That is, a combination of static and non-static non-human information is used to estimate the next time steps background. Static regions will persist over time and solidify their presence in the background map. Moving regions will lead to low probabilities in the background model and most likely fail to solidify their presence.

III. MICROSOFT KINECT SENSOR

The Kinect, released by Microsoft in North America on November the 4th, 2010, was designed to allow controller free game play on the Microsoft Xbox 360. As already stated, Microsoft uses the Israeli company PrimeSense sensor suite. This platform contains both an RGB camera (visible light) and an infrared (IR) sensitive camera. The sensor has an IR laser and diffraction grating that actively sends out unique patterns to be recognized using the IR camera. The IR laser and IR camera form a stereo pair. The depth data returned by the device (at 30 frames per second) is an 11-bit 640x480 image. The precision of the depth depends on the distance, where precision decreases from approximately two centimeters at two meters to approximately ten centimeters at six meters. While the Kinect software can support motion tracking, gesture, face and voice recognition, we do not currently use these features. Microsoft has just recently made the system available via a software development kit.

While the technical specifications have not been disclosed, the open source community have reversed engineered the inner workings of the sensor. A likely explanation involves the use of correlation-based matching of patterns as a result of the IR laser and diffraction grating. The minimum range is speculated to be somewhere around one meter and Microsoft recommends a user be around two meters from the device. In this work, we used the libfreenect open source library [24] to access the Microsoft Kinect sensor. The work described in [25] is used to transform the raw Kinect depth values into meters.

A. Advantages

Advantages of the Kinect sensor include the following.

- *Low-cost*

The Microsoft Kinect sensor currently retails for approximately $150 USD. When compared to a stereo camera system, this is approximately over an order of magnitude decrease in price. When compared to infrared imagers, this is approximately over two orders of magnitude decrease in price. For a domain like eldercare, this is significant. The method detailed in this article requires at least two cameras per monitored workspace.

- *Depth information*

The Kinect sensor has a number of advantages over traditional stereo vision. Namely, it operates in IR and on the basis of an actively emitted diffraction grating pattern. As such, it does not have the same set of shortcomings as passive depth sensing. An example is solving correspondence in stereo vision for a flat homogeneous region of little-to-no texture (e.g., a flat white wall with no texture). However, the Kinect can recognize such surfaces due to the active emission of a pattern.

- *Operation 24 hours a day*

One of the biggest advantages of the Kinect in the domain of well-being monitoring is its ability to operate in low-to-no light conditions. This makes it possible to recognize human activity (e.g., falls) in the day as well as night.

- *Shadows and illumination*

As already mentioned, a very serious problem associated with the processing of color imagery from visible light cameras is temporal and spatial variation of illumination. Every day examples are turning light sources on and off, emission from televisions, natural lighting, etc. In addition, shadows also exist in color imagery and they generally require separate procedures to identify or remove such artifacts. In [22,23], we used stereo vision to circumvent this topic of shadows and illumination. In this work, the Kinect sensor also has this advantage.

B. Disadvantages

Disadvantages of the Kinect sensor include the following.

- *Limited range*

A serious limitation of the Kinect is its depth range. However, for many indoor monitoring environments, such as the work discussed in this paper, this constraint may not prove to be too large of a factor relative to all the other sensor benefits.

- *Natural lighting and halogen light*

It has been found [26] that the Kinect sensor performs poorly in natural lighting and halogen light. The sensor works ideally in dim, but not completely dark conditions [26]. If a room has big windows, then it is recommended to shade them. While the

Kinect is able to operate in the dark, face recognition and other tasks benefit from both color and IR camera processing.

C. Side Notes

During our investigation, the following three observations were made. Two Kinect's were installed in a single room. The sensors were placed approximately five meters apart and they were viewing the same monitoring area. They were also rotated (about the x-y "floor" plane) approximately 180 degrees. Little effect, if any, was observed when the second Kinect was turned on. We also observed that certain types of clothing fail to reflect enough IR light back to the device to allow an estimate of depth at those pixels to be made. Additional analysis needs to be performed to fully understand what exact types of materials are ideal for sensing with respect to the Kinect. On a final note, an additional drawback of the Kinect sensor is its limited field of view, approximately 60 degrees.

IV. Microsoft Kinect Sensor Versus Stereo Vision

The Kinect can be used "as is" as a low-cost replacement for our prior multi-stereo camera computer vision well-being monitoring framework [22,23]. That is, depth information is extracted along with corresponding color imagery. Advantages include depth estimation from the Kinect using their structured light and low-cost. However, the Kinect can be used to obtain much more. Namely, our prior vision algorithm can be extended in order to achieve 24-7 operation. The necessary modifications are detailed in the following few sections.

A. Calibration and Color and Depth Image Acquisition

First, each Kinect sensor needs calibration. That is, we must estimate intrinsic and extrinsic camera parameters. One option is to use a checkerboard calibration pattern and supplemental IR backlighting [5]. Once again, we relied upon a set of camera parameters identified by the open source community [25] and the libfreenect [24] API for interfacing with the sensor.

Depth is acquired using the IR camera-laser pair. However, the Kinect also has a RGB (visible light) camera. Therefore, we must first transform the RGB color imagery into the IR camera space. The IR and RGB cameras are separated by a small baseline amount. A checkerboard pattern can be used to help determine the six degree of freedom (DOF) transform between these two cameras. Again, we used the common parameter set found by the open source community [25].

Lastly, calibration of the depth values returned from the Kinect is needed. The depth data returned from the Kinect must be transformed to obtain usable and accurate distances. In [5], Stone et al. describes a method to estimate and transform the Kinect sensor values. Instead, we used a method detailed in [25], which is based on the reverse engineering of the inner workings of the sensor. Figure 2 shows two Kinect RGB color images (which have been transformed into the depth camera) and the depth maps (scaled for display).

B. Human Detection

In [22,23], we proposed a *mixed* computer vision system for combining image space techniques (weak skin detector) with voxel techniques (head kernel and height filtering). Both color and depth information is required to realize this approach. Our extension for the Kinect is detailed in algorithm 3.

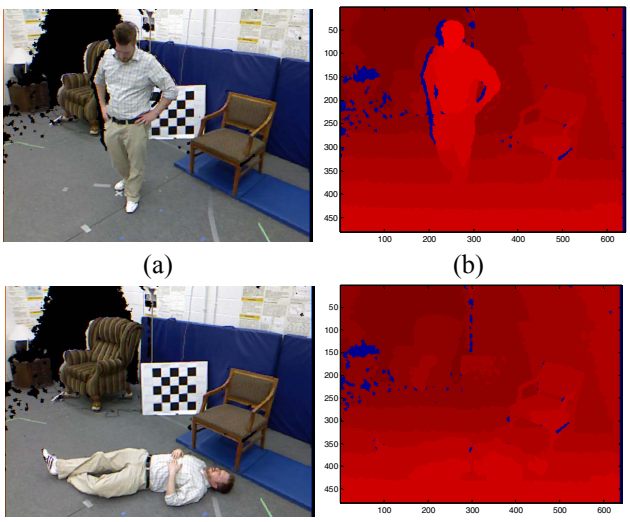

Fig. 2. Example Kinect color RGB and depth images (scaled for display). The lighter the red, the closer the depth. Blue values indicate no Kinect depth value. The RGB image shown is the projection of the raw color image onto the Kinect depth image (so color-depth association can be made).

In algorithm 3, the first extension involves checking for the presence of enough visible light using the RGB color camera. We start by calculating the image histogram of the grayscale image. To detect too low of visible light conditions, we employ the following check. If δ percent of the histogram resides in the interval [0,β], then we declare too low of light and the prior color-based system is bypassed. In the case of too low of illumination, the skin and color histogram similarity sub-systems are disabled. Instead, we are forced to rely solely on

Alg. 3: Extension to Algorithm 2 (Runtime) for the Kinect

WHILE NOT DONE
// Kinect depth and voxel reconstruction
(1) Collect color and depth images from Kinect sensors
(2) Build individual voxel spaces
(3) Build intersected (global) voxel space
(4) Build blanketed set
// Change detection and volume segmentation
(5) Remove background voxels from blanketed voxel space
(6) Segment objects
// Human detection
IF ((7) Enough visible light is present in RGB camera)
(8) Build color histogram for all objects
(9) Find human using head shape (with height restriction) and skin color or color histogram similarity
(10) Update human color histogram
ELSE
(8) Find human using head shape (with height restriction)
END IF
// Background update
(11) Remove human voxels from gobal voxel space
(12) Update background using nonhuman blanketed space

the head shape kernel in voxel space. However, this is an extremely important wide range of operational conditions that we were unable to address before.

Our rationalization is the following. If *enough* visible light is present, a combination of color and depth information is used. The utility of this approach was shown in [22,23]. However, in extremely low light environments, humans do not tend to perform complex activities such as those performed during the day. A system that detects people via depth-based change detection and head shape (with height restrictions) in three dimensional voxel space is practical and useful.

V. Preliminary Experiments and Results

The goal of the experiments put forth herein is to show that the Microsoft Kinect sensor can be used for high-quality, low-cost, 24-7 well-being monitoring in indoor environments. In order to support our claims, we take a two step approach. This is necessary because video is collected from different sensors with different capture rates (even when both collect the same number of frames per second, FPS) and no common three dimensional ground truth is available. In prior work, this was not a problem because we could compare our system results with others due to the use of a single common RGB camera.

A. Summary of Prior Experiments and Results

We start by summarizing our prior experiments and results. This demonstrates the effectiveness and superiority of our prior proposed mixed image space and world (voxel) space stereo vision computer vision solution (i.e., the framework that the Kinect sensor is integrated into). This work is compared to two similar existing methods, Stauffer and Grimson's Gaussian mixture models (GMMs) [27] and the work of Li et al [28].

We begin by noting that creating a ground truth for human tracking in three-dimensional space is not practical. In contrast, an image space ground truth can be created relatively easily by hand. The majority of comparative studies to date are also done in image space. The three dimensional output of our system can be projected back into image space, and directly compared to important existing image space algorithms. Figure 3 is the results of back projecting our three dimensional results into image space and its comparison to others and a hand segmented ground truth.

In total, six sequences, each with 1,200 frames, were collected at a rate of four frames per second to demonstrate a range of possible activities in a single person scene. Each sequence is five minutes long. Two subjects are used throughout the experiments. Subject one is used in sets 1, 2, 4 and 6, while subject two is used in sets 3 and 5. All data sets, ground truth and output for the stereo vision experimentation can be found at http://cirl.missouri.edu/vision/.

The combined statistics (table I) of all sequences displays the significant advantage of our system over the previous systems. The true positive rate is over ten percent higher than either algorithm, while true negative is also higher than each. It should also be noted that the ground truth data had an average of 6,576 foreground pixels and 300,624 background pixels per test image. So, a 1% change in foreground classification accuracy results in a change of roughly 66 pixels, while a 1% change in background classification accuracy results in a change of 3,006 pixels.

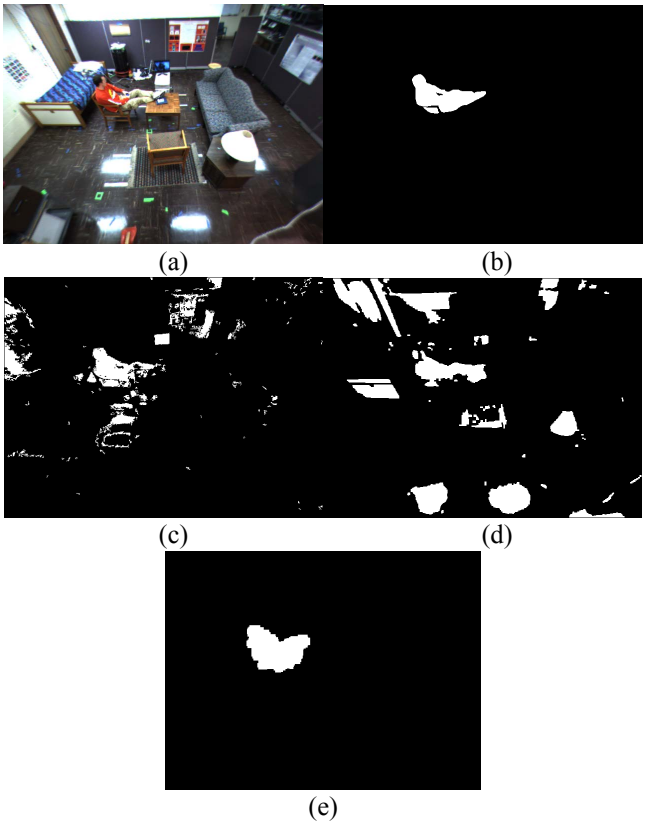

Fig. 3. Method for comparing a world space human recognition system to image space algorithms. At each frame, our system findings (in voxel space) are back-projected into image space (e). Example image showed is the state of the systems after a lighting change. (a) The original image. (b) The hand segmented ground truth. (c) The output of GMM [27]. (d) The output of Li et al [28]. (e) Projection of our findings into image space.

It should also be noted that the result of our system is a three-dimensional model of the human and objects that were moved. This provides a richer world-space representation for subsequent higher level processing. On a final note, while the image space results above show impressive quantitative improvement, the real advantage of this work resides in world space. Image space is only used for comparative analysis. Video demonstrating qualitative results can be found at http://cirl.missouri.edu/vision/.

Table I: Summary of Prior Experiments and Results

		Ground Truth	
		Foreground	*Background*
Our Stereo Approach [22,23]	*Foreground*	84.9%	1.3%
	Background	15.1%	98.7%
Stauffer and Grimson GMM [27]	*Foreground*	70.4%	8.6%
	Background	29.6%	91.4%
Li et al. [28]	*Foreground*	71%	2.3%
	Background	29%	97.7%

B. *Preliminary Findings For the Kinect*

The following experiments were designed for two reasons. First, we show that our proposed Kinect extension does operate in low-to-no light scenarios (a logical conclusion, but neverless a feature that we must verify). However, we can only compare Kinect data to a human ground truth acquired manually (verbal indication of activity being performed and when relative to the start of a video sequence). Second, while our prior stereo vision and the Kinect approach are similar, i.e., both operate on the basis of depth as well as color information. There is no direct way to compare them. We have different sensors and a ground truth must be provided for each independently. We reproduced the majority of our prior stereo vision experiments using the Kinect and the two approaches are compared on this basis.

It is not practical (or safe at that) to perform data collections of complex activities (e.g., falls) in low-to-no light conditions. Therefore, data sequences were collected herein with enough illumination to safely maneuver the environment, but not bright enough to be resolved in color images. Two student researchers were used. During capture, activities were manually recorded relative to the start of video capture.

However, before the low-to-no visible light experiments can be discussed, we must first summarize our prior linguistic summarization of human activity from video work [1,10]. The human activity analysis module is used here to generate a finite set of activity decisions that can be compared to human ground truth. Our prior hierarchical soft computing vision framework is able to address: human recognition, behavior inference and information reduction/complexity management. One benefit of that approach is the summarization of video in a *natural* way that domain experts, e.g., nurses, can more easily understand. That is, a fewer number of temporal linguistic descriptions are produced versus thousands of individual image decisions or numeric summaries of activity over time. In [1], we showed that linguistic summarizations can be used computationally to recognize higher-level complex human behavior. In particular, our work is focused on abnormal event detection, specifically fall recognition. Natural language terms are modeled using linguistic variables (fuzzy set theory). Fuzzy logic is used to infer human activity from features of voxel objects and temporal partitions of fuzzy membership time series. Example summarizations produced include "[The resident] has [fallen] in the [kitchen] in the [early morning]" or "[The resident] is [lying on the couch] in the [livingroom] for a [long time] in the [late afternoon]".

The first Kinect data set is three videos, approximately five minutes long each. Two Kinects were used and the monitored work space is shown in figure 2. In this data, an insignificant level of visible light was present (enough to ensure navigation and the safety of a subject). The subject walked back and forth between the chairs, varying sitting down, walking and standing (i.e., simple activities). The Kinect-based system and behavior recognition module found all of the ground truth identified with no false alarms. We acknowledge that this experiment is relatively simple. However, in low light settings, one would not expect a subject (elder) to perform complex behaviors. It is possible that the subject is sleep walking or doing something else suspicious or dangerous. Depending on the level of assisted care required, a system might generate an alert when an elder is walking around in the middle of the night in low-to-no light. Second, our application of interest is fall recognition. However, we did not have our subjects simulate any falls in the low light conditions due to the fact that the activity is unsafe in such a context. However, we did have the subjects lie on the ground and then we turned off the lights. We processed the data for a single moment and verified that we could indeed see the person lying on the ground in voxel space. Neverless, these experiments reinforce the notion that the Kinect can be used to monitor humans in low-to-no light conditions (something our other visible light sensors are not able to do).

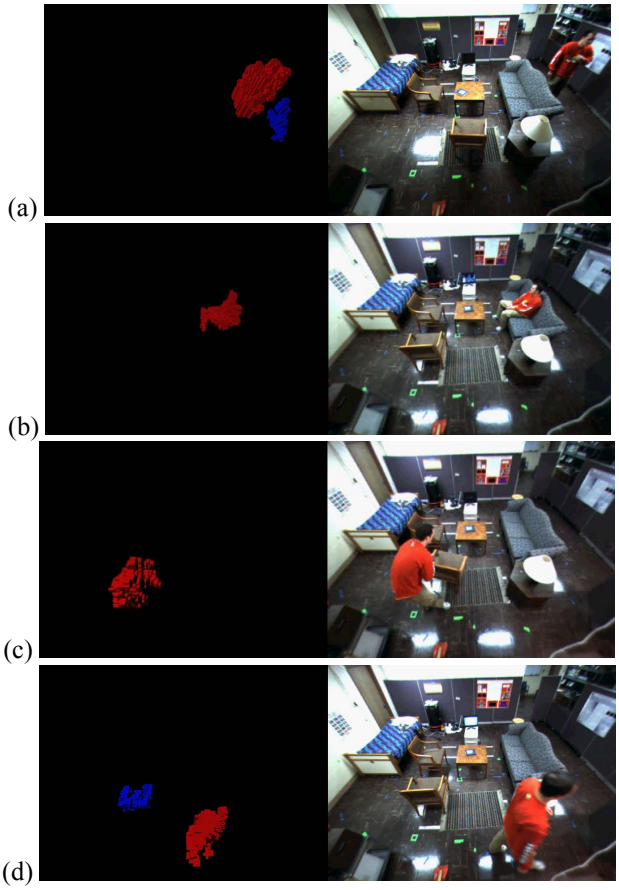

Fig. 4. Frames qualitatively demonstrating the success of our proposed stereo vision system under different challenging scenarios such as initial detection of the human, tracking, objects being constantly moved, lighting changes, and rapidly changing visual content on a monitor screen. Again, red are voxels labeled as human and blue is voxels labeled as change but not human.

Table II: Performance of Kinect system for sufficient visible light scenarios.

		Ground Truth	
		Foreground	*Background*
Kinect Approach	*Foreground*	86.3%	2.5%
	Background	13.7%	97.5%

In the second set of experiments (table II), we collected three videos, approximately five minutes long each. Again, two Kinects were used. We reproduced the majority of scenarios

discussed in [22,23] (i.e., the data set used to produce table I). Specifically, 20 images were used for ground truth. We produced a binary image mask using the RGB color image for the image locations of the human. As table II shows, the Kinect and stereo vision systems (table I) are indeed very similar in terms of their performance when enough visible light is present and both color and depth information is being used.

VI. Conclusion and Future Work

In summary, we presented an extension to our prior multi-stereo camera computer vision system for monitoring the well-being of elders in indoor environments. Namely, we discussed how a new low-cost color and depth imager, the Microsoft Kinect sensor, can be used to extend the operational range of our previous approach to 24-7. This sensor also has advantages over traditional stereo vision from color imagery due to the Kinect structured light. Advantages and disadvantages of the infrared sensor and diffraction grating-based depth estimation procedure are investigated. In addition, we described necessary extensions to our prior computer vision algorithms which were initially designed for visible light cameras. Preliminary results are very encouraging. Experiments indicate that the Kinect can be used in low-to-no visible light scenarios and the sensor performs very similar to stereo vision in situations where sufficient visible light is present and mixed image and voxel space algorithms are engaged.

References

[1] D. T. Anderson, R. H. Luke, J. M. Keller, M. Skubic, M. Rantz, and M. Aud, "Linguistic Summarization of Video for Fall Detection Using Voxel Person and Fuzzy Logic," *Computer Vision and Image Understanding*, vol. 113, pp. 80-89, 2009.

[2] D. Hodgins, "The Importance of Measuring Human Gait," Medical Device Technology,vol. 19 (5), pp. 44-47, 2008.

[3] B. Maki, "Gait changes in older adults: predictors of falls or indicators of fear," *Journal of the American Geriatrics Society*, vol. 45 (3), pp. 313–20, 1997.

[4] J. Hausdorff, D. Rios, H. Edelberg, "Gait variability and fall risk in community-living older adults: a 1-year prospective study," *Arch Phys Med. Rehabilitation*, pp. 1050–1056, 2001.

[5] E. Stone & M. Skubic, "Evaluation of an Inexpensive Depth Camera for Passive In-Home Fall Risk Assessment," *Proceedings, Pervasive Health Conference*, 2011, Best Paper Award.

[6] E. Stone, D. T. Anderson, M. Skubic and J. M. Keller, "Extracting Footfalls from Voxel Data," *Proceedings, 32nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Buenos Aires*, pp. 1119-1122, 2010.

[7] E. Stone, D. T. Anderson, M. Skubic, J. M. Keller, "Footfall extraction and visualization from voxel data," *Proceedings, International Society for Gerontechnology 7th World Confreence*, pp. 27-30, 2010.

[8] G. Demiris, M. J. Rantz, M. Aud, K. D. Marek, H. W. Tyrer, M. Skubic and A. Hussam, "Older Adults' Attitudes Towards and Perceptions of 'Smarthome' Technologies: a Pilot Study," *Medical Informatics and The Internet in Medicine*, vol. 29 (2), pp. 87-94, 2004.

[9] G. Demiris, O. Parker, J. Giger, M. Skubic and M. Rantz, "Older adults' privacy considerations for vision based recognition methods of eldercare applications," *Technology and Health Care*, vol. 17, pp. 41-48, 2009.

[10] D. T. Anderson, R. H. Luke, J. M. Keller, M. Skubic, M. Rantz and M. Aud, "Modeling Human Activity from Voxel Person Using Fuzzy Logic," *IEEE Transactions on Fuzzy Systems*, vol. 17 (1), pp. 39-49, 2009.

[11] R. H. Luke, "Moving Object Segmentation from Video Using Fused Color and Texture Features in Indoor Environments", *Technical Report*, University of Missouri, *http://cirl.missouri.edu/vision/*, 2008.

[12] P. Viola, "Robust real-time face detection," *International Journal on Computer Vision*, vol. 57, pp. 137-154, 2004.

[13] M. Yang, D. Kriegman, and N. Ahuja, "Detecting faces in images: a survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, pp. 34-58, 2002.

[14] R. Osuna, R. Freund, F. Girosi, "Training support vector machines: an application to face detection," in *Computer Vision and Pattern Recognition*, pp. 130-136, 1997.

[15] H. Frigui and P. Gader, "Detection and Discrimination of Land Mines in Ground-Penetrating Radar Based on Edge Histogram Descriptors and a Possibilistic K-Nearest Neighbor Classifier," *IEEE Transactions on Fuzzy Systems*, vol. 17, pp. 185-199, 2009.

[16] W. Tsaipei, M. K. James, D. G. Paul, and S. Ozy, "Frequency Subband Processing and Feature Analysis of Forward-Looking Ground-Penetrating Radar Signals for Land-Mine Detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, pp. 718-729, 2007.

[17] M. Popescu, P. Gader, and J. M. Keller, "Fuzzy spatial pattern processing using linguistic hidden Markov models," *IEEE Transactions on Fuzzy Systems*, vol. 14, pp. 81-92, 2006.

[18] D. T. Anderson, J. M. Keller, and O. Sjahputera, "Algorithm fusion in forward-looking long-wave infrared imagery for buried explosive hazard detection," i*n Detection and Sensing of Mines, Explosive Objects, and Obscured Targets XVI*, pp. 801722-801722, 2011.

[19] D. Makris and T. Ellis, "Learning semantic scene models from observing activity in visual surveillance," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol 35 (3), pp. 397-408, 2005.

[20] D. Thirde, M. Borg, J. Ferryman, F. Fusier, V. Valentin, F. Bremond, and M. Thonnat, "A Real-Time Scene Understanding System for Airport Apron Monitoring," in *IEEE International Conference on Computer Vision Systems*, pp. 26, 2006.

[21] T. Raty, "Survey on Contemporary Remote Surveillance Systems for Public Safety," *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol 40 (5), pp. 493-515, 2010.

[22] R. H. Luke, D. T. Anderson and J. M. Keller, "A system for change detection and human recognition in voxel space using stereo vision," Tech. Report, University of Missouri, *http://cirl.missouri.edu/vision/,* 2011.

[23] R. H. Luke, "A system for change detection and human recognition in voxel space using stereo vision," Ph.D. Thesis, Electrical and Computer Engineering, University of Missouri, 2010.

[24] OpenKinect.org, http://openkinect.org/wiki/Main_Page, 2011.

[25] ROS.org, http://www.ros.org/wiki/kinect_calibration/technical, 2011.

[26] POPSCI, http://www.popsci.com/gadgets/article/2010-11/how-set-your-living-room-microsoft-kinect, 2011.

[27] C. Stauffer and W. Grimson, "Adaptive Background Mixture Models for Real-Time Tracking," in *Computer Society Conference on Computer Vision and Pattern Recognition*, 1999.

[28] L. Li, W. Huang, I. Gu and Q. Tian, "Foreground Object Detection from Videos Containing Complex Background," in *Proceedings of the eleventh ACM international conference on Multimedia*, pp. 2-10, 2003.