Fuzzy Voxel Object

Derek T. Anderson, Robert H. Luke, Erik Stone, and James M. Keller

Department of Electrical and Computer Engineering
University of Missouri, Columbia, USA
Email: {dtaxtd, rhl3db, erikstone}@mizzou.edu and kellerj@missouri.edu

Abstract*—In this paper, computer vision and fuzzy set theory are merged for the robust construction of three-dimensional objects using a small number of cameras and minimal a priori knowledge about the objects. This work extends our previously defined crisp model, which has been successfully used for recognizing and linguistically summarizing human activity. The objects true features more closely resemble the fuzzy object than those of the crisp object. This is demonstrated both empirically and through the comparison of features used in tracking human activity.*

Keywords—computer vision, human activity analysis, fuzzy objects, fuzzy voxel person.

1 Introduction

Information gathered from three-dimensional models is significantly more useful and robust than that gathered from two-dimensional sources. This is quite apparent in the area of human surveillance. Features, such as the centroid, height, velocity, orientation, etc, in three-dimensional space are not camera view dependent. Object segmentation can also be improved using three-dimensional information about the environment, such as floors, walls, and objects for the removal of additional erroneous artefacts, such as shadows and specular highlights [1][2]. At each step in computer vision, be it construction, shape refinement, etc, various forms of uncertainty are present and should be utilized.

We previously demonstrated a method for constructing a three-dimensional human model, voxel person, in voxel (volume element) space (figure 1) through the use of privacy protected images of the human called silhouettes [1][2].

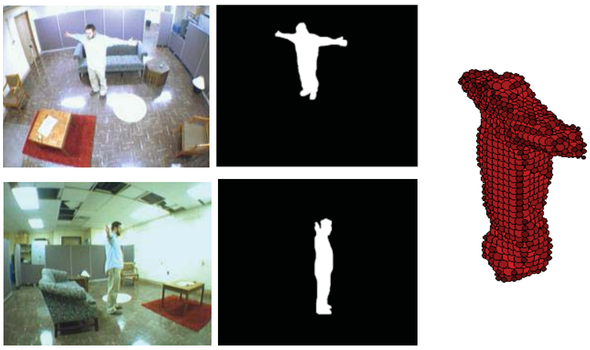


Fig. 1. Crisp voxel person created at a voxel resolution of 1.25"x1.25"x1.25", using 2 cameras and image silhouettes.

A silhouette is a binary map that distinguishes an individual (foreground) from his or her background. Well-known algorithms for computing the silhouette include Mixtures of Gaussians [3], Eigen Backgrounds [4], and Wallflower [5]. Silhouettes are used and we extract linguistic summarizations of human state based on crisp voxel person and fuzzy logic [1]. The resulting information is in a natural format for humans (linguistic) and it also yields succinct descriptions for managing complexity. We extended the work in [1] using a hierarchy of fuzzy logic and linguistic summarization for the inference of activity [2]. The knowledge (rules) is designed under the supervision of nurses for the recognition of falls in elderly populations.

Construction of three-dimensional objects, both solid representations and hulls, from two-dimensional images has been studied intensely in computer graphics, computer vision, biomedicine, and even in the activity analysis domain [6][7][8][9][10]. What separates our object construction related work from most, besides the use of silhouettes for back-projection, is the way in which a small number of low cost cameras, typically two per environment, are used to build a low, but rich for tracking, resolution voxel object; how its shape is refined using environment knowledge; and how features are extracted and used for fuzzy-based activity recognition [1][2]. Segmentation of the object into body parts (torso, head, etc) is not attempted, such as in [11][12]. While advances have been made in body segmentation, no approach to date is either real-time or mature enough to be included in a real-world eldercare system that runs unsupervised for long time periods.

This paper is divided as such. Section 2 describes the crisp voxel model, section 3 is the fuzzy model, and section 4 is the assesment of the quality of construction. Feature extraction is described next, followed by an alpha-cut procedure for an improved crisp voxel person, and section 7 is experiments and qualitative and quantitative results.

2 Crisp Voxel Object

An environment used for tracking is first converted to a voxel representation. Voxels are non-overlapping cubes that discretize a volume. The voxel centered at position $< i, j, l >$ is $v_{(i,j,l)}$. For each pixel in each camera, a list of intersected voxels is pre-computed, $L_{c,(n,m)}$, where c is the camera index, $1 \le c \le C$, and (n,m) is the pixel index [1]. The lists are sorted according to their distance (Euclidean) to the camera. This way, voxels that correspond to regions of change (the foreground, e.g. the human) from the viewpoint of a given camera can be quickly identified. For camera c at time t, the human voxel object is the union of all voxel lists

ISBN: 978-989-95079-6-8

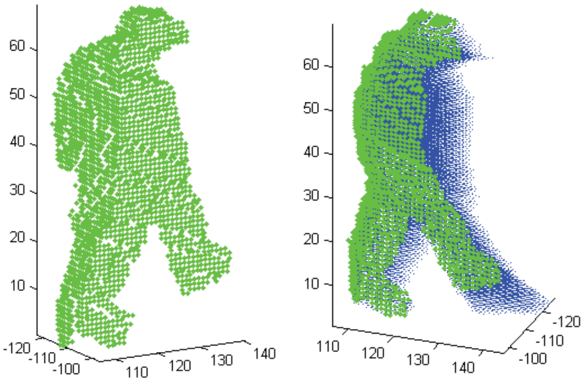
for the foreground pixel set, V_t^c. The object's shape is refined by intersecting the voxel objects acquired from each camera in the scene, i.e. $\{V_t^1,\ldots,V_t^C\}$, thus

$$V'_t = \wedge_{c=1}^{C} V_t^c . \qquad (1)$$

The object is then subject to additional morphological operators for noise removal and voxel space processing for removing additional shadows and specular highlights [1].

Another useful object of interest is the visible shell, S_t, which is the set of all voxels that are directly visible according to the C cameras. The algorithm for computing the shell (shown in figure 2) at time step t is

```
Initialize the set $S_t$ to empty
Compute voxel person, i.e. $V'_t = \wedge_{c=1}^{C} V_t^c$
For each camera $(1 \le c \le C)$
    For each pixel in the foreground set
        Find the closest voxel, $v_{(i,j,l)}$, from $L_{c,(n,m)}$ in $V'_t$
        Add $v_{(i,j,l)}$ to $S_t$.
```



(a) (b)

Fig. 2. (a) Shell (shown in green) and (b) shell along with the rest of voxel object (shown in blue) for a two camera setup.

3 Fuzzy Voxel Object

The quality of voxel person construction varies over the space depending on the object's position and the location of the cameras. As a result, information gathered from the crisp model can be inaccurate. The crisp voxel object is a good real-time initial technique to produce three-dimensional models for tracking. The technique works quite well when there are several cameras viewing a space with overlapping fields of view. Unfortunately, it is rare to have more than two cameras in a given area, due to factors such as cost, processing time of subsequent algorithms, data transmission and storage, and in the case of eldercare, limited space and installation locations, for the seamless integration of a passive video sensor network into the home.

Fuzzy voxel person is an attempt to model and use the different types of uncertainties related to the construction of the object. Each voxel is assigned a membership value that reflects how much it belongs to the actual object. The two types of uncertainty identified and fused in this paper are (a) how reasonable it is to infer that a voxel is part of the object based on the shell and (b) where a voxel is located relative to the distribution of mass across the object.

The first measure considers the minimum distance one must "step" in voxel space until a voxel on the shell is reached. More importance is placed on voxels near the shell. However, this measure has a natural tendency to favor the shell and pull confidence away from the central mass of the object. The second measure determines how confident one is in a voxel, based on how close it is to the central mass. This measure obviously favors the central mass and in return pulls confidence away from the shell. When these two sources of information are combined, the final value is a measure of how dense the region is around the voxel and how reasonable it is to infer that location given its distance to the observed shell. This combined confidence is low for voxels in the tail of the object (the furthest part from the shell), high in the dense close to the shell regions, and relatively higher (than the tail) for voxels in non-dense volumes close to the shell.

The first value is inversely related to a voxel's distance from the shell. The greatest certainty of object intersection occurs at the shell and decreases outwardly. The distance value is found using mathematical morphology, specifically dilation, $\oplus$. The kernel used is K, and $V'_t \oplus K$ is the dilation of the model V'_t by kernel K. A K of 3x3x3 of all ones is used (i.e. the immediate 26 adjacent voxel neighborhood). The distance value is computed quickly by repeatedly dilating the voxels in the shell, then subtracting that set from the set of remaining voxels in the intersected object. All surviving voxels in the intersected set have a related value, $m_{(i,j,l)}$ for $v_{(i,j,l)}$, that is incremented each iteration it survives. Formally, the algorithm is

```
Initialize all $m_{(i,j,l)}$ to 0
$V' = V'_t$
$S' = S_t$
while $|V'| > 0$
    for each $v'_{(i,j,l)} \in V'$
        add one to $m_{(i,j,l)}$
    end
    $V' = V' - S'$
    $S' = S' \oplus K$
end
```

The confidence for each voxel is

$$m'_{(i,j,l)} = 1 - \left(\frac{m_{(i,j,l)}}{\max\limits_{m_{(a,b,c)}} m_{(a,b,c)}} \right). \qquad (2)$$

The density value is computed using erosion, Θ. Again, a K of 3x3x3 of all ones is used. The voxel object is continually eroded until no voxels remain in the set. For each erosion step that a voxel remains in the set, a separate value, $e_{(i,j,l)}$ for $v_{(i,j,l)}$, is incremented. The algorithm is

ISBN: 978-989-95079-6-8

```
Initialize all $e_{(i,j,l)}$ to 0
$V' = V'_t$
while $|V'| > 0$
    for each $v'_{(i,j,l)} \in V'$
        add one to $e_{(i,j,l)}$
    end
    $V' = V' \ominus K$
end.
```

The most direct method for calculating the membership value per voxel is

$$e'_{(i,j,l)} = \frac{e_{(i,j,l)}}{\max\limits_{e_{(a,b,c)}} e_{(a,b,c)}} . \qquad (3)$$

However, this assigns low confidence to the object's entire outer shell. Up to this point, no a priori knowledge about the object being tracked has been assumed. For some objects, such as a human with appendages (arms, legs, chest, etc), this calculation might be too drastic, resulting in too little of importance assigned to the appendages (which are less dense than the chest). In order to accommodate such cases, the confidence values can instead be calculated as

$$e'_{(i,j,l)} = (1-\beta)\frac{e_{(i,j,l)}}{\max\limits_{e_{(a,b,c)}} e_{(a,b,c)}} + \beta . \qquad (4)$$

This maps the value into [β , 1], where ($0 \le \beta \le 1$) is a free parameter, either user defined or learned. The value β can be thought of as the minimum support one is willing to assign to voxels on, and subsequently near, the shell.

The final value per voxel, $\mu_{(i,j,l)} \in [0,1]$, is found using a t-norm (we use the product),

$$\mu_{(i,j,l)} = m'_{(i,j,l)} \wedge e'_{(i,j,l)} . \qquad (5)$$

4 Object Construction Quality

As alluded to earlier, the quality of construction varies over the space with respect to the object's location, the installation locations of the cameras, and their respective orientations. Quantitatively representing the construction quality is of importance for two reasons. First, the value informs us about the reliability of the lower level results. For example, knowledge about the ability to properly construct the human affects the features calculated, and subsequently our decisions regarding activity inferred using those features. Secondly, there may be many locations in a space for which it is impossible to properly construct a voxel object given a particular camera configuration. In many instances, suppression of any object construction in those particular areas is of value.

The best construction generally occurs when the intersecting view vectors are orthogonal. This means that each individual voxel has a different construction quality. The quality of construction for voxel $v_{(i,j,l)}$ is

$$Q_{(i,j,l)} = 1 - \max_{\substack{C_g, C_h \\ g \neq h}} \left| (v_{(i,j,l)} - C_g)^T (v_{(i,j,l)} - C_h) \right| , \qquad (6)$$

where C_g and C_h are the center locations for two cameras. Thus, the best case per voxel per camera pair with respect to orthogonality is considered. If the intrinsic camera parameters have been estimated, for example using [13], then the pixel ray in camera c that intersects voxel $v_{(i,j,l)}$ can be used instead. The environment is again converted into a voxel representation, but this time the resolution is lower. We empirically determined that a sampling of once every foot was a good resolution for our application. Figure 3 shows 9 horizontal slices (x-y planes for a varying z) of the sampled voxel space and their respective qualities.

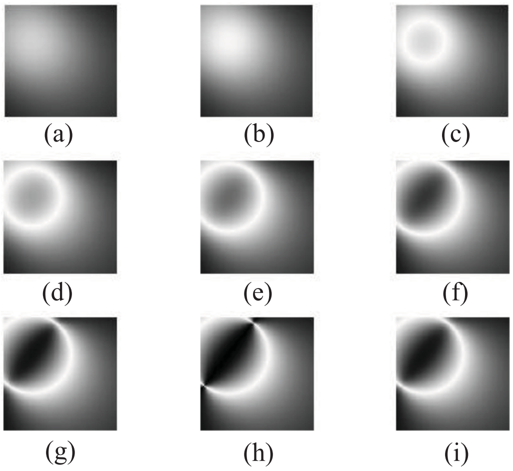
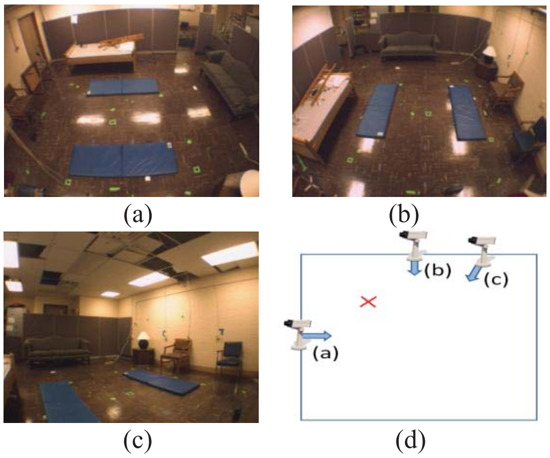

(a) (b) (c)
(d) (e) (f)
(g) (h) (i)

Fig. 3. Nine horizontal (x-y plane) slices of voxel object construction quality for the camera configuration {(a),(b)} shown in figure 4. Brighter values represent higher quality. Map (a) above is at a height of 1 foot, and each consecutive map is 1 foot higher in the z dimension (world up direction).



(a) (b)
(c) (d)

Fig. 4. Camera installation locations (d) and images showing their respective views of the monitored space (a-c).

The next matter is the determination of the quality of construction for an object. One can compute the mentioned quality measure for each voxel and make a decision based on the set of all memberships, or the object can be summarized, according to its centroid, height, or some other measurement, and that point estimate can be used to make a decision for the entire object. We perform the latter. We convert the monitored space into a low resolution voxel configuration (such as demonstrated in figure 3), then the height domain is collapsed and a single voxel plane in the x-y dimension is produced. The voxel qualities are combined using a t-norm (we used the min),

$$Q_{(i,j)} = \min_{k} Q_{(i,j,k)} . \qquad (7)$$

This pessimistically selects the worst construction case per element in the final x-y plane voxel set. Figure 5 shows the final x-y voxel plane of combined confidences for figure 3.

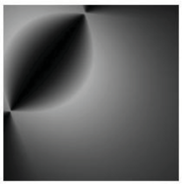
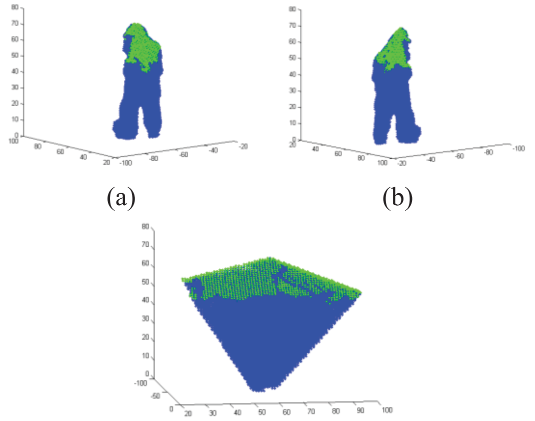


Fig. 5. T-norm produced x-y voxel plane of combined object construction confidences for the nine slices in figure 3.

The voxel object's centroid is mapped to the closest (Euclidean) voxel in the quality map (according to its x-y plane distance), and the respective quality value, ϕ_t, is retrieved. If $\phi_t < \delta_1$ (we empirically determined $\delta_1 = 0.2$), then the object is not constructed and ultimately not tracked. The parameter δ_1 is not specific to any one single tracking case, but it is rather a general parameter related to ray-based back-projection object construction. Figure 6 shows an example construction in which the quality value is 0.05.



(a) (b)

(c)

Fig. 6. Example low quality construction (value of 0.05) using camera configuration {(a),(b)} (the human's location is marked by an X in figure 4). The object appears to be constructed correctly when viewed independently with respect to the two cameras, images (a) and (b). However, its actual shape from another position (c) in three-dimensional space shows that the object was not built correctly.

5 Fuzzy Feature Extraction

After fuzzy voxel person is created, and its quality is assessed and determined to be adequate, one of two possible decisions can be made. The object can remain fuzzy and fuzzy feature extraction can be performed for activity analysis, or the fuzzy object can be used to acquire a better crisp voxel person and standard crisp feature extraction can be performed. This section details the prior.

The fuzzy feature selected for analysis in this paper is the centroid, which is most often used for determining object interaction, velocity, direction, position, and more generally, inferring the overall state of the tracked person (standing, kneeling, on the ground, etc) [1]. The fuzzy centroid is

$$M_{fuzzy} = \frac{\sum_{v'_{(i,j,l)} \in V'_t} \mu_{(i,j,l)} v'_{(i,j,l)}}{\sum_{v'_{(i,j,l)} \in V'_t} \mu_{(i,j,l)}} . \qquad (8)$$

6 Crisp Voxel Object Improvement

An alternative to directly using fuzzy voxel person for tracking is if the volume of the object is known ahead of time, or if it can be sufficiently approximated online, fuzzy voxel person can be used to acquire an improved crisp voxel object. Once the crisp object is obtained, standard crisp tracking and feature extraction can be used. The concept here is that fuzzy voxel person is useful for obtaining the global relative set of memberships, and a search for a better crisp object is guided by the evidence contained in fuzzy voxel person. By using the known or approximated volume, P_{real}, the highest confidence areas can be identified and kept. One possibility is to look for an alpha cut such that the crisp volume (cardinality) of the resultant set is closest to P_{real}. Formally, the search is for an α such that

$$\arg\min_{0 \le \alpha \le 1} \left| \left| {}^{\alpha}V' \right| - P_{real} \right| , \qquad (9)$$

where ${}^{\alpha}V'$ is

$${}^{\alpha}V' = \left\{ v'_{(i,j,l)} \mid \mu_{(i,j,l)} \ge \alpha \right\} . \qquad (10)$$

One can easily solve this by sorting all voxel memberships in descending order and then selecting the first P_{real} voxels (in the case of a tie, include all ties).

In the case that P_{real} must be approximated, the procedure is as follows. For a user defined time window, T frames, compute crisp voxel person and its associated quality value at each time step. Find the maximum ϕ_t value for the T time steps. If this value is above some threshold δ_2, again this is a global parameter not specific to any one camera configuration, use the crisp voxel-based volume to approximate P_{real}. Figure 7 shows a human built using an approximated P_{real}.

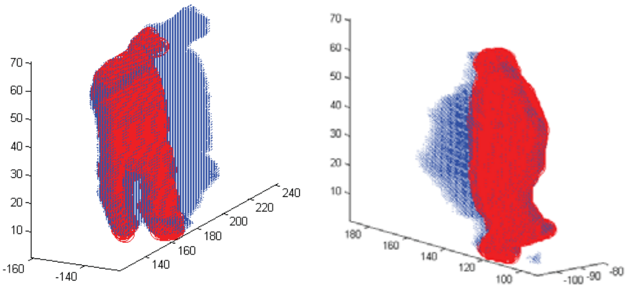


Fig. 7. The proposed alpha cut-based procedure for fuzzy voxel person to obtain an improved crisp object. Red areas are the improved voxel person and the blue areas are the rest of the original crisp voxel person.

In order to quantitatively demonstrate the improvement in the general shape and object's orientation, the following metrics for tracking humans are presented. The voxel covariance matrix is calculated, and its major orientation, the eigenvector with the largest respective eigenvalue, is compared to the world up vector ($<0,0,1>$). For a human standing and for a file cabinet used in the results section the major orientation should be very similar to the world up vector. Next, the difference, according to the variation associated with each eigenvector, between the new and real object is compared. These values indicate how the mass was distributed across the object according to its primary orthonormal basis with respect to variation. The eigen-information is extracted from the covariance matrix of voxel person, and the eigenvalues (φ) at 3 standard deviation are calculated, $\omega = \sqrt{\varphi} * 3$. The eigenvalues are sorted in decreasing order, and the difference between the real object (ω_{real}) and the crisp voxel (ω_{voxel}) is computed, $\omega^{'} = |\omega_{real} - \omega_{voxel}|$. The feature used in the results section is the sum of the three components in $\omega^{'}$, denoted by ω_{diff}.

7 Experiments and Quantitative Results

While figure 7 qualitatively demonstrated the benefits of the techniques presented in this paper, comparison between the (a) fuzzy and crisp centroids, (b) eigen-based analysis of the voxel spread (variance), and (c) comparison between the expected and the observed major orientations using eigen-analysis is performed. The subject in this case is a file cabinet that was placed in unique locations and orientations throughout the scene. A file cabinet has easy-to-recognize shape, orientation, known volume, and centroid.

For the file cabinet and the density-based measure, a value of $\beta = 0.5$ is used. Combinations of camera pairs, $\{(a),(b)\}$ and $\{(b),(c)\}$ were used. Camera combination $\{(a),(b)\}$ is an ideal situation, given the near orthogonal installation, while $\{(b),(c)\}$ is less than ideal. The later configuration results in much larger inaccuracies due to the trailing tail of intersected voxels. Tables 1-6 numerically display the accuracy of the fuzzy-based approaches, figures 8 and 9 show top-down view of the file cabinet at several locations, and figure 10 shows a few perspective views of the constructed object to give a better understanding of camera view induced warping and the resultant object shape.

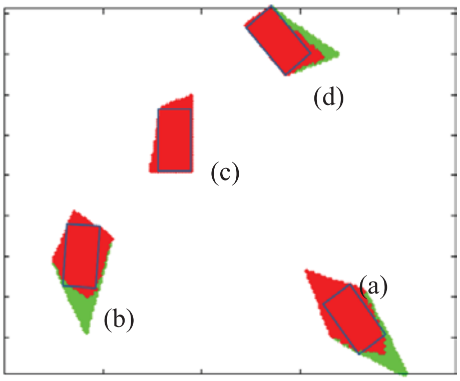


Fig. 8. Crisp (red and green areas) and alpha-cut based crisp objects (red areas) for camera configurations $\{(a),(b)\}$. The real object's approximate shape and orientation are overlaid using a rectangular outline.

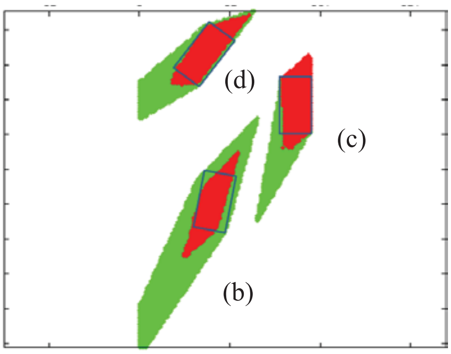


Fig. 9. Crisp (red and green areas) and alpha-cut based crisp objects (red areas) for camera configurations $\{(b),(c)\}$. The real object's approximate shape and orientation are overlaid using a rectangular outline.

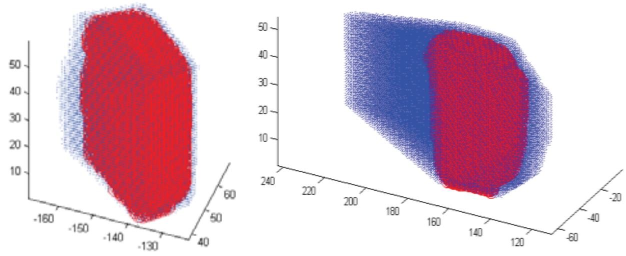


Fig. 10. Improved construction. Red areas are the improved voxel object and the blue areas are the rest of the original crisp voxel object. The left image is for (d) in figure 8, i.e. cameras $\{(a),(b)\}$, and the right image is for (b) in figure 9, an extreme case, using cameras $\{(b),(c)\}$.

Table 1: Difference between real and computed object centroid. Camera configuration $\{(a),(b)\}$. Units are specified in feet. Smaller numbers are better.

	a	b	c	d
Crisp	0.46	0.51	0.49	0.21
Fuzzy	0.19	0.29	0.47	0.27

ISBN: 978-989-95079-6-8

Table 2: Difference between real and computed object centroid. Camera configuration {(b),(c)}. Units are specified in feet. Smaller numbers are better.

	b	c	d
Crisp	2.24	1.18	1.13
Fuzzy	1.47	0.56	0.79

Table 3: Camera configuration {(a),(b)}. Value of ω_{diff}. Smaller numbers are better.

	a	b	c	d
Crisp	87.79	58.60	16.72	42.78
α -cut	4.60	1.72	0.82	0.07

Table 4: Camera configuration {(b),(c)}. Value of ω_{diff}. Smaller numbers are better.

	b	c	d
Crisp	3776.99	411.02	658.13
α -cut	807.52	55.71	17.11

Table 5: Camera configuration {(a),(b)}. Values are the dot product between the object's major orientation (eigenvector with the largest eigenvalue) with the world up vector. Larger numbers are better.

	a	b	c	d
Crisp	0.91	0.96	1.00	0.99
α -cut	1.00	1.00	1.00	1.00

Table 6: Camera configuration {(b),(c)}. Values are the dot product between the object's major orientation (eigenvector with the largest eigenvalue) with the world up vector. Larger numbers are better.

	b	c	d
Crisp	0.15	0.32	0.31
α -cut	0.93	0.96	0.79

The fuzzy centroid is more accurate than the crisp centroid in every case for configuration {(b),(c)} (i.e. extreme viewing condition). For configuration {(a),(b)}, very ideal condition, the centroids are very close. Fuzzy is better overall, however, in one instance the fuzzy centroid is slightly worse (Table 2.d). In that situation, camera (b) was looking at the object from a top down view, and the visible shell resulted in a greater vertical (z direction) pull. Tables 3 and 4 show that the objects shape (according to ω_{diff}) is closer to the real object. Lastly, tables 5 and 6 show that the objects major direction is always equal or more accurate (significant in Table 6) to the world up direction.

8 Conclusion and Future Work

In this paper, a fuzzy-based computer vision approach to the robust construction of three-dimensional objects for human activity analysis using only a few cameras and minimal a priori knowledge of the object is introduced. Extreme joint viewing conditions were considered and the fuzzy acquired results are better than the crisp counterpart in both the quantitative and qualitative regard.

The shell is currently approximated using the visible parts of the voxel object, created using back-projection. As the figures illustrate, this results in some warping in the objects shape. Stereo vision will be used to improve the estimation of the visible shell from depth maps (which will also require the t-norm operator in (5) to be revisited). Additionally, the use of a priori knowledge about the object being tracked will be used, for example, attempting region segmentation, and using that information for improved membership calculation.

Acknowledgment

Robert Luke and Derek Anderson are pre-doctoral biomedical informatics research fellows funded by the National Library of Medicine (T15 LM07089). This work is also supported by the National Science Foundation (ITR award llS-0428420) and the Administration on Aging (90AM3013).

References

[1] 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*, doi:10.1016/j.cviu.2008.07.006, Vol. 113, pp. 80-89, 2008.

[2] 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*, Accepted July 20th 2008.

[3] C. Stauffer, W.E.L. Grimson, Adaptive Background mixture Models for Real-time Tracking, *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, (2), 243-252, 1999

[4] N. Oliver, B. Rosario, A. Pentland, A bayesian computer vision system for modeling human interactions. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(8), 831-843, 2000

[5] K. Toyama, J. Krumm, B. Brumitt, B. Meyers, Wallflower: Principles and Practice of Background Maintenance. *In Proceedings of ICCV'1999*, 255-261 1999

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

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

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

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

[10] M. Pardas and J. Landabaso, "Foreground regions extraction and characterization towards real-time object tracking," in *Proc. of Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms*, Lecture Notes in Computer Science. Springer, 2005.

[11] G. Cheung, T. Kanade, J. Bouguet, and M. Holler, "A Real Time System for Robust 3D Voxel Reconstruction of Human Motions", in *Proc. of Computer Vision and Pattern Recognition*, Vol. 2, pp. 714-720, 2000.

[12] N. Thome, D. Merad, and S. Miguet, "Human Body Part Labeling and Tracking Using Graph Matching Theory", in *Proc. of IEEE Intl. Conf. on Video and Signal Based Surveillance*, 2006.

[13] Jean-Yves Bouguet, "Camera Calibration Toolbox for Matlab," Intel Corp, 2004, http://www.vision.caltech.edu/bouguetj/calib_doc/.