3D Shape Context Based Gesture Analysis Integrated with Tracking using Omni Video Array

Kohsia S. Huang and Mohan M. Trivedi
Computer Vision and Robotics Research (CVRR) Laboratory
University of California, San Diego
http://cvrr.ucsd.edu/

Abstract

In this paper we introduce a multilayer cylindrical histogram based feature space for representing 3D shape context of human body. Dynamics of gestures are analyzed using discrete hidden Markov models (DHMM) with quantized feature vectors. Extensive experimental trials with multiple subjects and a range of gesture classes are presented. Gesture recognition accuracies of over 85% (for nine gestures, and 9 subjects) and over 95% (for seven gestures) support the basic feasibility and robustness of the approach.

Keywords*: Movement analysis, 3D person tracking, activity analysis, multi-camera systems.*

1. Introduction

Human gestures provide an important and useful information source for intelligent ("smart") environments [1]. In this paper we address a multi-level voxel-based gesture recognition system with a 3D shape context descriptor. The 3D human tracking level gives a rough estimate of a person's location and height in the large space, then the next level of confined 3D voxelization provides more details on that location. Real-time voxel reconstruction of the human body is carried out by shape-from-silhouette on the networked omni video array (NOVA). On the gesture recognition level, we design the 3D shape context to capture human body configurations with noisy and low-resolution voxels obtained with wide field-of-view omni cameras in a large space. In this case, an articulated body model is very difficult to capture human gestures. With 3D shape context, the robustness and accuracy are enhanced by utilizing both spatial and temporal contexts of the 3D body dynamics. Spatially, the 3D shape context captures human body configuration for each frame. Temporally, discrete hidden Markov models (DHMMs) are trained for each gesture to accumulate the likelihood of the 3D cylindrical context features across frames to make the final gesture decisions.

Gesture analysis methods are typically either (a) appearance-based using single 2D image data, or (b) 2½D model-based deriving 3D articulated body models on single 2D image data with prior knowledge. They match or model gestures either on singleton image or continuous video sequences [2]-[10]. For 3D volumetric data, usually an articulated human body model is applied to match the data [2][3][5][11]. Some appearance-based approaches sample the occupancy of the body or body parts of a person in a 2D grid context which shifts with the person in 2D image sequences and models the periodicity of the motions to recognize gestures [7][12]. Another related gesture estimation work with 2D shape context [13] is with edge-based shape matching between a frame of a human image in a single-camera testing video and a huge set of human images of different poses with hand-labeled 3D human body models [8]. For camera modality, most of the work in the literature uses single or multiple rectilinear cameras. However, the only gesture recognition work with multiple omni cameras is still 2½D [10]. In [10], the 2D silhouettes of a human from multiple omni cameras are overlaid with the rectangular grids of 3×3 bins to capture the 2D body shape context on each aspect. Then the shape context data is combined to estimate gestures across frames.

In Table 1, we present a comparison of the related research efforts mentioned above with this work. The uniqueness of our approach is at using voxel reconstruction on omni video array, 3D cylindrical shape context of the voxel data, and spatial-temporal modeling of 3D shape context sequences by HMM. Using 3D analysis is more robust than 2D and 2½D because 3D not only provides richer and angle-independent information, but is also less subject to

Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)
1063-6919/05 $20.00 © 2005 **IEEE**

Table 1
Comparison of the closely related body gesture researches that use view-based shape context analysis

Research Efforts	Objective (Gesture/ Tracking)	Input Video	Shape Context	Gesture Classification	Simultaneous Gesture Analysis	Gesture Vocab.	# Subj.	Accuracy	Speed
Mori-Malik (2002) [8]	Body Modeling	Monocular 2½D	2D	N/A	N/A	N/A	N/A	N/A	Slow
Polana-Nelson (1994) [7]	Activity Analysis	Monocular 2D	2D	Spatial-temporal sequence matching by nearest centroid	Single person	7	1	Unclear (100% w/ small data set)	Medium
Yamato-Ohya-Ishii (1992) [12]	Activity Analysis	Monocular 2D	2D	HMM of 2D shape context sequence	Single person	6	3	96%, same training-testing static subjects 71%, mixed subjects	Fast
Ishiguro (2001) [10]	Activity Analysis & 2D Tracking	Multi-Perspective 2½D	2D	Spatial-temporal sequence matching by dynamic programming	Single person	10	Unknown	87%, moving person (16 cameras required)	Real-Time (~10 fps)
This Research	Activity Analysis & 3D Tracking	Multi-Perspective 3D	3D	HMM of 3D shape context sequence	Multi-people	7~9	9	95% for 7 gestures 86% for 9 gestures (Mixed subjects & motion with only 4 cameras)	Real-Time (~10 fps)

occlusions due to multiple cameras. In addition, 2D and 2½D involves iterations between the image and the 3D model. Our approach captures 3D data and directly performs 3D gesture analysis, which is a combination of the efficiency of the related works.

2. Real-Time Voxel Reconstruction Integrated with Multi-Person Tracker

The real-time omni array voxel reconstruction is built upon the 3D tracker to capture volumetric human forms [14] as shown in Figure 1. For multi-person tracking, the calibrated omni video array detects human blobs on the unwarped panoramas by background subtraction, measures the human locations and heights by horizontal and vertical triangulation procedures, and tracks the humans by Kalman filters [15]. The 3D track estimates provide a basis to integrate the subsequent human voxelization and gesture analysis efficiently.

Upon the 3D location, a confined voxel volume is applied to analyze the human object for a more detailed 3D shape as illustrated in Figure 2. The base of the volume is a square of 4m × 4m centered at the track's current horizontal location, and the height of the volume is 1.05 times the current height of the track. This size is able to cover people of different sizes, speeds, and gestures. Within this volume, cubic voxels of 8cm size fills up the volume. If the voxel size is smaller, more details of the human body can be represented. However, no further benefits are gained if the voxel size becomes smaller than the projected camera pixel size. In addition, smaller voxels require a heavier computation load. Therefore, we take such a voxel size for low resolution omni-cameras with higher computation speed but do not sacrifice the discerning ability for gestures. To voxelize a human body, the voxels are scanned one-by-one and projected onto the panoramas of the cameras in the array by finding the panoramic coordinate (x_P, y_P) of the voxel for each camera as shown in Figure 3,

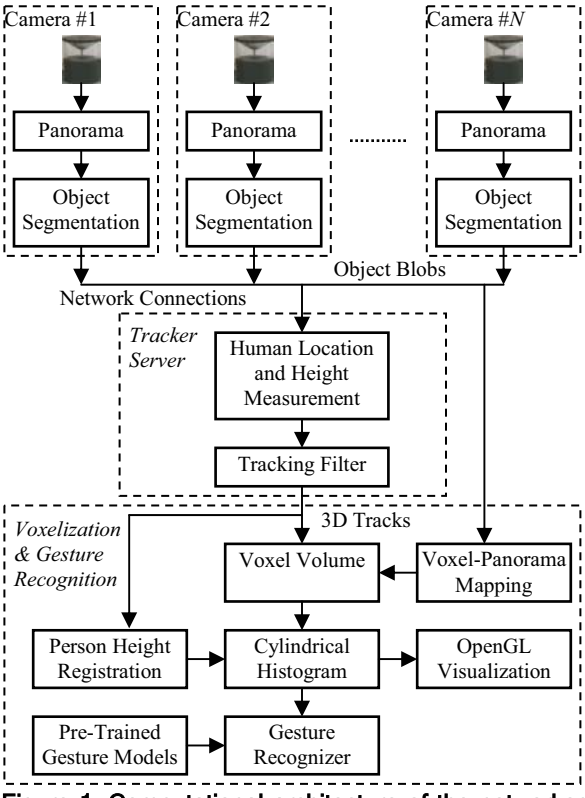

Figure 1: Computational architecture of the networked omni video array for real-time 3D person tracking, voxelization, and gesture recognition.

$$x_P = (\theta / 2\pi) \cdot W \ , \ \text{and} \ \theta = \arctan 2(y - Y, x - X) \quad (1)$$

$$y_P = \frac{z - Z}{\sqrt{(x - X)^2 + (y - Y)^2}} + y_H \quad (2)$$

where (x, y, z) is the world coordinate of the voxel, (X, Y, Z) is the world coordinate of the omni-camera optical center, W is the number of pixel columns of the panorama, and y_H is the panoramic row index that corresponds to the horizontal level. Thus, if the panorama pixel (x_P, y_P) is foreground, the count of the voxel is increased by one. After projecting to all the N cameras, the voxel is set if the count is greater than or equal to $N - 1$ since the person might be under one omni-camera and not seen by it. As compared to other shape-from-silhouette methods such as octree [16], this method is efficient due to its simplicity and the relatively small number of voxels in the volume [5][11]. An example of simultaneous multi-person voxelization is shown in Figure 4. Then, the voxels in the voxel volume are 3D-labeled and we only take the connected voxels of a person near the center of the voxel volume. This procedure rejects the errors caused by the invasive voxels of a nearby person.

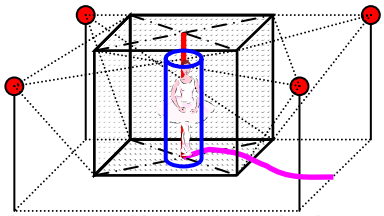

Figure 2: 3D human tracking and voxel volume on omni video array.

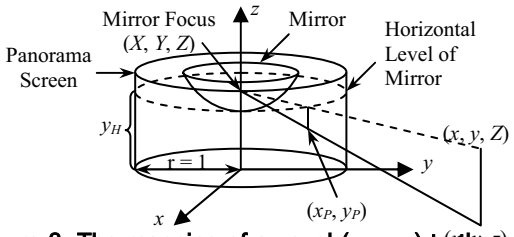

Figure 3: The mapping of a voxel (x, y, z) to the omni-camera panorama pixel (x_P, y_P).

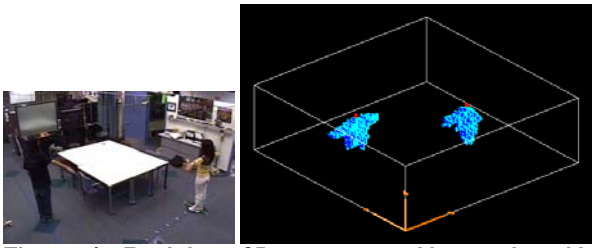

Figure 4: Real-time 3D person tracking and multi-person voxelization in an indoor environment. Note the raw image is taken at the point diagonally opposite to the origin of the voxel view.

3. 3D Shape Context for Gesture and Movement Analysis

After human voxelization, gesture analysis can be performed on each person respectively. Given the low-resolution human body voxel reconstruction on the omni array, it is obvious that gesture analysis with articulated model would be very unlikely to work reliably [2][1][3][11]. In this section, we present a novel view-based gesture recognition framework which utilizes both spatial and temporal contexts of the voxel reconstruction. The spatial context of the body voxels is captured by the proposed multilayered cylindrical histogram or 3D shape context, as shown in Figure 5. The temporal context of body gesture dynamics is captured by a HMM, which then makes the final gesture classification.

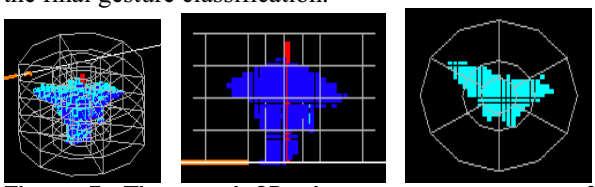

Figure 5: The novel 3D shape context capture of human body voxels by multilayered cylindrical histogram.

3.1. Voxel-Based Human Shape Context

In terms of *spatial context*, human body configuration is captured by the 3D shape context in all directions, as shown in Figure 5. Similar inspirations are also found in the works of solving correspondence problems for shape matching [8][13], where the 2D shape context is designed to compare the shape configurations between different subjects. The fundamental principle in these methods is to resolve ambiguity by referencing not only the point under consideration but also some neighborhood around it. In our method, the 3D shape context is invariant to human scale and translation. When a person walks into the space, his/her height is registered by a moving average of the Kalman height estimates in the first 3 seconds. Then the 3D shape context is scaled to the human size by keeping the cylinder height H 1.15 times the registered human height, and the cylinder radius is also scaled proportionally. For translation invariance, the central axis of the cylinder is anchored to the centroid of the human voxels with the base on the ground. Thus, as the person gestures, voxels of different body parts fall into different cylindrical bins, B_{ij}'s, and accumulate the bin counts, as shown in Figure 5, where the horizontal index i and the vertical index j are

$$i = \text{int}(a/(2\pi/F)) + 1 + (r-1)F,\ \ j = \text{int}(h/(H/L)),$$
and $i = 0$ if $r = 0$ (the central bin) (3)

Note here a is the horizontal angle of a voxel with respect to the centroid of voxels, F=6 is the number of cylindrical fans, r is the ring index increasing from center to the outer ring, h is the elevation of the voxel, and L=4 is the number of vertical layers. The central cylinder is designed to capture the torso, and the outer rings of the cylinder capture the limb gestures in all directions. Two outer rings (R=2) are needed to discern different limb gestures such as pointing and hand-waving. Also, the vertical layers are designed to capture upper and lower body gestures. Thus, this 3D shape context produces a 52 dimensional feature vector $\mathbf{v}_k$ for each frame by stretching the bins from the center to the outer rings counterclockwise and from bottom to top in the following order

$$k = i \cdot (RF + 1) + j \quad (4)$$

Note that R, F, and L can be reconfigured for differentiating various complex gestures. Next, the number of voxels in each bin is normalized by the total number of voxels for that person so that the entries of the feature vector $\mathbf{v}_k$ sum up to 1. Thus, 3D shape perturbations introduced by noisy voxels can be absorbed and alleviated by this normalization. This 3D shape context is able to work reliably with noisy and low-resolution voxels for gesture modeling.

Concerning rotational invariance, we note that it is possible to rotate the 3D shape context to equalize the orientation of the subject. However, since the human voxels are noisy due to sensor noise, it would be very difficult to equalize the subject orientation robustly. The noisy cylinder rotation would cause even more severe disturbances on the shape context features and gesture recognition. Therefore, we chose not to apply this rotational invariance and ask the subjects to rotate while training the gesture models.

3.2. Spatial-Temporal Context Dynamics Modeling for Gesture Recognition

In terms of *temporal context*, the dynamics of the spatial configuration of body gestures is tracked by hidden Markov models. Similar to spatial context, temporal context resolves single-frame ambiguities by accumulating the likelihoods and interpolating the features across frames [17][18], which also significantly boosts accuracy.

In our method, the cylindrical histogram feature space $\{\mathbf{v}_k\}$ is first partitioned into M regions by vector quantization (VQ). The vector quantizer is trained by the Linde-Buzo-Gray algorithm using Euclidean distance [19]. Then for each gesture, a discrete hidden Markov model (DHMM) is designed to model the gesture dynamics. As shown in Figure 6, we evaluate nine gesture classes: kick, stretch, walk, bow, pick, point, wave-hand, sit-down, and stand-up. It is noted that except for the sit-down and stand-up gestures, the other seven all involve a cyclic process. For example, a waving hand process is described as standing ↔ lifted arm ↔ hand near head ↔ hand away from head ↔ lowered arm ↔ standing, and hand near head ↔ hand away from head, and may repeat several times. For stand-up and sit-down gestures, it is obvious that the process is left-to-right, e.g. sitting → half way standing → standing. Figure 7 summarizes the structure of the Markov chains. It is obvious that the states of the Markov chain represent certain intermediate body poses for each gesture. Each gesture should have N different sets of VQ steps associated to the N states since the states represent different intermediate body poses. The VQ steps of each state represent similar and noisy instantaneous body poses.

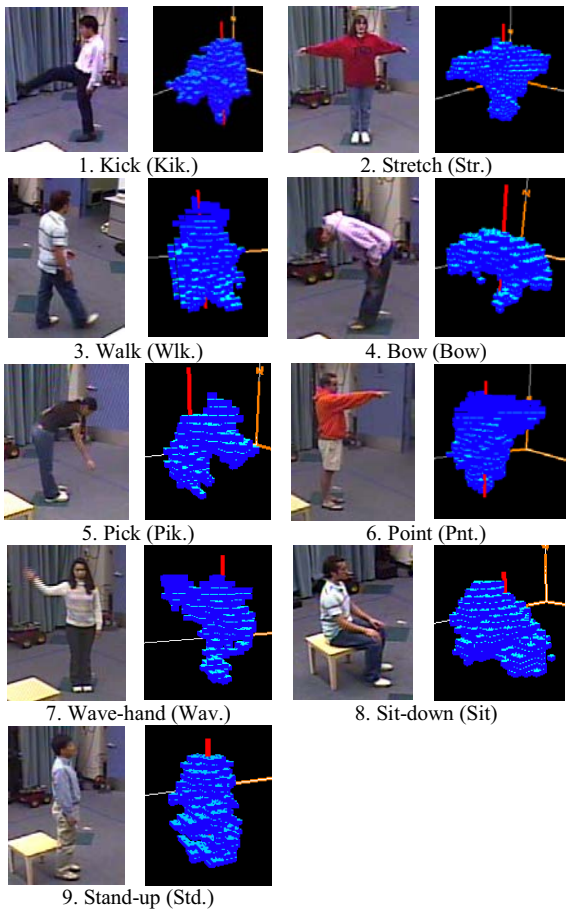

Figure 6: Examples of human body voxelization of the 9 gestures analyzed in the experiments.

DHMMs are parameterized by the number of hidden Markov chain states N, the number of observation symbols M, and the probability set $\lambda=(\boldsymbol{\pi}, \mathbf{A}, \mathbf{B})$, where $\boldsymbol{\pi}$ is the $1\times N$ vector of initial probabilities of the N states, $\mathbf{A}$ is the $N\times N$ state transition matrix whose rows sum up to 1, and $\mathbf{B}$ is the $N\times M$ matrix of the M observation probabilities of the N states, where the rows of $\mathbf{B}$ sum up to 1 [20]. Note that the M observation symbols are the M vector quantization steps. To train the DHMMs, the initial probability $\boldsymbol{\pi}$ is initialized randomly for the cyclic DHMMs and initialized to start from the very left state for the left-to-right DHMMs. The state transition matrix $\mathbf{A}$ is initialized according to Figure 7 with minor randomness on the diagonal elements. We also define a parameter TrL to indicate the number of states that a state can transit to among itself and its neighbors, for example, TrL=2 means that state i can only transit to i and i+1 for left-right Markov chain and to i–1, i, and i+1 for left-right-left Markov chain. The rest of the state transition probabilities in the rows of $\mathbf{A}$ are initialized as zeros and will be kept as zeros during the Baum-Welch training iterations [20]. The observation matrix $\mathbf{B}$ is first initialized as the distribution of the VQ outputs (from 1 to M) of the training sequences plus minor noises. Then we apply the Viterbi algorithm once on $(\boldsymbol{\pi}, \mathbf{A}, \mathbf{B})$ to back-track the state sequence and find the new distribution of (1,…,M) for each state to fill in the rows of the new $\mathbf{B}$ [20]. Then the standard Baum-Welch algorithm is applied to re-estimate $\lambda=(\boldsymbol{\pi}, \mathbf{A}, \mathbf{B})$ for each gesture.

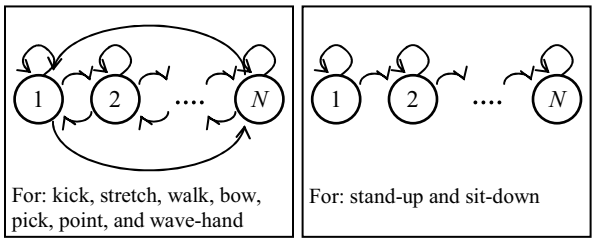

Figure 7: Two types of Markov chains of the DHMMs that model the gesture classes defined in this section.

In the testing phase, the testing sequence of 3D shape context features $\mathbf{v}_k$ passes through the trained vector quantizer to be the observation sequence of (1,…,M). Then the accumulated likelihoods of the sequence of K frames $L = \pi_i a_{ij} b_{j1} a_{jk} b_{k2} \cdots a_{mn} b_{nK}$, $1 \le i, j, k, m, n \le N$ on the DHMMs are computed by forward procedure, as shown in Figure 8. As a classification problem [20], the final gesture decision of the pose dynamics in the observation sequence is determined by maximum likelihood. We note that the likelihood L is a spatial-temporal accumulation across K frames, since the observation probability b is related to the 3D shape context of the body poses and L by itself is a temporal accumulation. This mechanism would enhance accuracy and robustness of gesture recognition under challenging environmental variations across people and gestures.

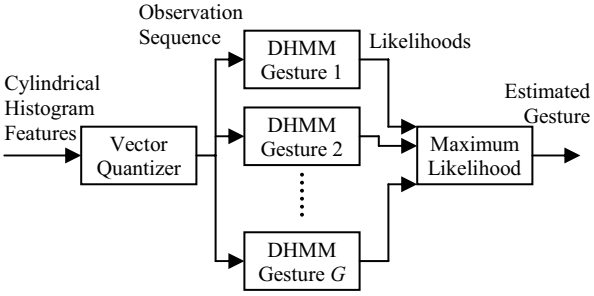

Figure 8: Gesture recognition scenario with spatial-temporal context accumulation.

4. Experimental Validation

In this section we describe the experimental setup and results to evaluate the feasibility of the proposed gesture recognition system. We conducted the experiments in an indoor environment of 6.6m × 6.7m. Four omni cameras are installed on the ceiling of the room at a height of 2.8m, each located near the center of a quadrant. As in Figure 1, the omni cameras are connected to four network-linked dual 1 GHz Pentium III computers. This configuration distributes the most time-consuming pixel processing to four processors in parallel. The 3D tracker, voxelization, and gesture analysis are running on a dual 2 GHz Xeon server. The camera videos are acquired in 640×480 and unwarped to panoramas of 720×170. With proper multi-thread timing, the system runs at 23 fps when there is no one in the room, while it runs at 6~10 fps when people are in the room. The bottleneck is at the processing speed of the four omni video processors which perform low-level pixel processing. Figure 9 shows an example of real-time tracking, voxelization, and 3D shape context capture of multiple people.

To train the VQ and DHMMs, 9 subjects participated in the data collection. 10 clips of 35~135 frames were collected as training sequences on 9 gestures (kick, stretch, walk, bow, pick, point, wave-hand, sit-down, and stand-up) for each person. In the training clips, the subject gestured and rotated about 40 degrees each time to cover 360 degrees in the 10 clips. Then 5 testing sequences of arbitrary directions were collected on the gestures of each subject. There were a total 53901 training frames and 25551 testing frames collected. Occlusions were avoided by removing some furniture. As shown in Figure 5, the cylindrical histogram has 4 layers, and each layer has one central bin and two rings of 6 bins each, totaling 52 bins. So each frame is a 52 dimensional vector. These data

were later analyzed offline with Matlab. Since the VQ and HMMs are initialized randomly on training, we took 5 trials for each set of parameters, (N, M, TrL). Then we computed the average and unbiased standard deviation of the 5 trials for accuracy estimation.

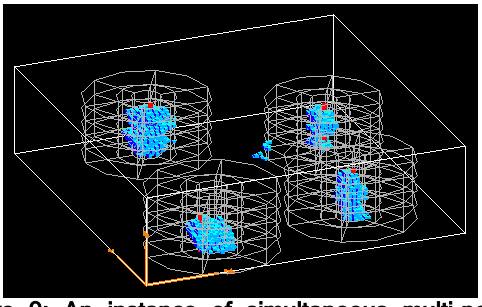

Figure 9: An instance of simultaneous multi-person tracking, voxelization, and gesture acquisition.

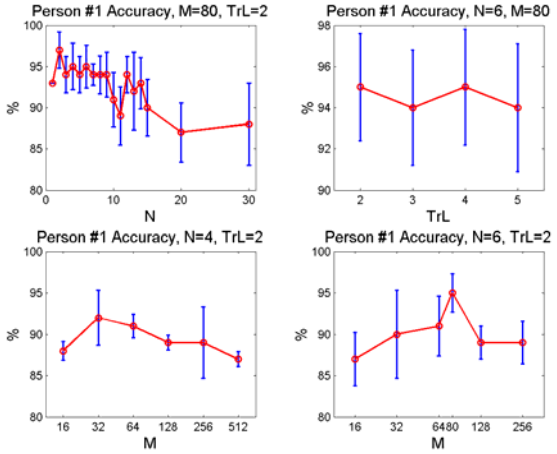

Figure 10: DHMM single-person training-testing performances on (N, M, TrL). 9 gestures were tested collectively. The circles denote the mean accuracy and the error bars denote the standard deviation in 5 trials.

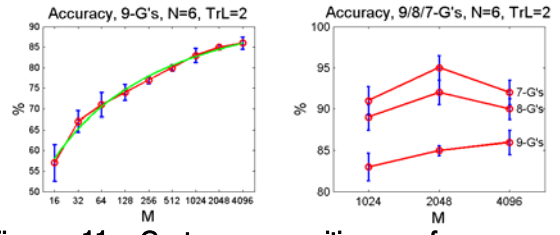

Figure 11: Gesture recognition performance of collective subjects on *M*. The left plot is on 9 gestures, and the right plot is on 9, 8, and 7 gestures by excluding the most ambiguous gestures.

We first evaluated the performance characteristics versus DHMM parameters on a single subject, as shown in Figure 10. Here VQ and DHMM were trained and tested on single person data. The N-dependence trial is designed to prove the interpretation of the HMM states as intermediate poses, as mentioned in section 3.2. From this trial we found the appropriate range of N. Thus, the main focus of the characteristics trial for the best model parameters falls on M and TrL. From Figure 10, it indicated that the performance was more dependent on M. Then we evaluated the single-person accuracy with the selected parameter set. Single-person accuracy was measured when both training and testing data are from one person only. The accuracy of the 9 subjects are shown in Table 2. The numbers in Table 2 reveal that the performance is quite dependent on a different person, so we evaluated the accuracy on a collection of people. Figure 11 shows the accuracy curve versus M collectively on 9 gestures and 9 people. It also shows the accuracy after excluding the most ambiguous gestures. The accuracy confusion matrices of 9 and 7 gestures of all subjects are shown in Table 3 and Table 4 respectively. Table 5 presents the accuracy in terms of people and number of gestures. The average accuracy of all subjects also indicate the feasibility of the proposed method. These results will be explained in the next section.

Table 2
Single-person accuracy on the 9 subjects. N=6, M=80, TrL=2, 9 gestures in both training and testing sequences.

Person	#1	#2	#3	#4	#5	#6	#7	#8	#9
μ (%)	95	87	91	79	83	85	88	80	80
σ (%)	2.3	3.3	1.4	1.8	3.0	3.1	1.1	3.0	3.7

Table 3
Accuracy confusion matrix of 9 gestures for all 9 subjects. *N*=6, *M*=2048, *TrL*=2. Each row sums up to 100%.

%	Kik.	Str.	Wlk.	Bow	Pik.	Pnt.	Wav.	Sit	Std.
Kik.	77.8	4.4	6.7	0	11.1	0	0	0	0
Str.	0	95.6	0	0	0	2.2	2.2	0	0
Wlk.	0	0	100.0	0	0	0	0	0	0
Bow	0	0	0	86.7	13.3	0	0	0	0
Pik.	2.2	0	2.2	8.9	86.7	0	0	0	0
Pnt.	4.4	4.4	2.2	0	0	51.1	37.8	0	0
Wav.	0	0	0	0	2.2	20.0	77.8	0	0
Sit	0	0	0	0	0	0	0	95.6	4.4
Std.	0	0	0	0	0	0	0	0	100.0

Table 4
Accuracy confusion matrix of 7 gestures (excluding pointing and kicking) for all 9 subjects. *N*=6, *M*=2048, *TrL*=2.

%	Str.	Wlk.	Bow	Pik.	Wav.	Sit	Std.
Str.	97.8	0	0	0	2.2	0	0
Wlk.	0	100.0	0	0	0	0	0
Bow	0	0	86.7	13.3	0	0	0
Pik.	0	2.2	8.9	88.9	0	0	0
Wav.	0	0	0	2.2	97.8	0	0
Sit	0	0	0	0	0	95.6	4.4
Std.	0	0	0	0	0	0	100.0

Table 5
Accuracies on 9, 8, and 7 gestures of the 9 subjects. N=6, M=2048, TrL=2.

#G's.	P-1	P-2	P-3	P-4	P-5	P-6	P-7	P-8	P-9	**Avg.**
9	84.4	80.0	93.3	82.2	82.2	84.4	95.6	82.2	86.7	**85.7**
8	92.5	92.5	100.0	90.0	90.0	87.5	100.0	92.5	90.0	**92.8**
7	97.1	94.3	100.0	100.0	91.4	88.6	100.0	94.3	91.4	**95.2**

5. Analysis of Results

In this section we discuss the significance of the experimental results and point out the future work. Figure 10 reveals performance characteristics of the VQ-DHMM setup. Since the states of the Markov chain represent the substantial truth of the intermediate body poses, one can imply that not too many numbers of states are needed. In the first plot, the accuracy trend seems level before N=9, then decreases gradually with N. Also the standard deviation gets larger with N. It indicates that as N increases, VQ-DHMM models finer and finer intermediate poses. It eventually causes the observed over-fitting phenomena because some of the intermediate poses might be very similar between some gestures like point and wave-hand. For the second plot, the accuracy does not change much with different TrL's and the state transition matrix A is similar among them. Hence TrL=2 should cover the needs. On the third and the fourth plots in Figure 10, it indicates that the number of VQ codebook M matters. Since VQ codebook relates to the discerning of feature vectors, a too small or a too large M would either under-represent the gesture poses or over-fit them. These two plots also indicate that the optimum M varies with N. Due to the length of time, we did not evaluate all the combinations of M and N. Table 2 shows that there is large variation in the single-person accuracy among different people. This is because every person has a different gesturing style. Therefore, a subject-independent training is necessary.

In Figure 11, we trained the model with data from all subjects with N=6 and a different M. The required M is much larger for the same accuracy as compared to single-person cases. That is because every person has a different gesture styles, and VQ-DHMM needs to discern the intermediate poses in more detail. The fitting curve in the first plot is a second order polynomial along the $(\log)^2(M)$ axis. On this curve, the accuracy becomes 80% when M=442, and 90% when M=56790. With the limitations of Matlab, we tried up to M=4096. On the second plot of Figure 11, we tried to see the accuracy versus M when excluding some ambiguous gestures. It appears that M=2048 is a good choice since the accuracy is better with less gestures. From the confusion matrix in Table 3, pointing is the most ambiguous gesture with hand-waving, and kicking is the second most ambiguous gesture with picking. The ambiguities could be better resolved if the voxel quality were improved by better foreground segmentation quality on the ODVP panoramas, smaller voxel size, and increased number of cameras. Also the geometry of the cylindrical histogram can be finer, e.g., assigning more vertical layers, to resolve the ambiguity between kicking and picking. However, the improvements would not be limitless, since by nature these gestures have certain similarities or overlaps among them, especially with different people. Thus, as these ambiguous gestures are excluded, the overall accuracy increases as indicated in Table 4. Table 5 gives another aspect of accuracy in terms of the subjects, showing the accuracy improvement on excluding ambiguities.

These results prove the basic feasibility of the proposed framework of 3D shape context and VQ-DHMM classification. We have verified the results with video data of 9 gestures from 9 people. As compared in Table 1, some other works do not test on this amount of gestures and human subjects, e.g., [12] verifies 6 tennis strokes of 3 people on 625-dimensional feature vectors. Our unique 3D framework achieved similar accuracy with only 52-dimensional feature vectors. For improvements, we can include a likelihood threshold test in Figure 8 in order to classify known and unknown gestures. We can also add a likelihood ratio test among the likelihoods in Figure 8 in order to assess the recognition confidence. We also need to expand the training and testing sets to include more gestures and subjects so that the system is more person-independent. Also, a hierarchical gesture recognition scheme as in [21] can be considered to further enhance accuracy and/or broaden the range of gestures.

We are currently investigating the online implementation issue. For best and most unobtrusive gesture recognition, the system needs to automatically detect the start and end of a gesture [22][23]. For future research directions, first as suggested by Table 2, person identification is possible by differentiating the intermediate pose processes of the gestures among people. Some feature analysis algorithms such as independent component analysis may also be applied to process the cylindrical histogram feature vectors for each gesture to extract motion composition elements ("movemes" [5]) for different people. This would further enhance gait analysis and identification discernment. Finally, interactions among a group of people can be analyzed. A dynamic Bayesian network with input from multiple cylindrical contexts of the group can be utilized for this purpose.

6. Concluding Remarks

In this paper we proposed a novel 3D view-based algorithm for human form and gesture analysis. The real-time 3D tracking system gives rough estimates of the locations and heights of people, and the voxel reconstruction on omni video array renders more detailed body forms. A novel resizable and multilayered cylindrical histogram is then latched to the voxels to capture the 3D shape context of the human body poses. The spatial-temporal dynamics of the captured context feature vectors over frames is modeled by the VQ-DHMMs for different gestures. For an unknown sequence of cylindrical feature vectors, the gesture is classified by maximum likelihood among the trained VQ-DHMMs. Experimental results support the feasibility of the proposed scheme.

Acknowledgements

This work was supported by US DoD Technical Support Work Group (TSWG) and UC Discovery Grant. The authors are thankful to the contributions of our colleagues in the UCSD CVRR Laboratory.

References

[1] M. M. Trivedi, K. S. Huang, I. Mikić, "Dynamic Context Capture and Distributed Video Arrays for Intelligent Spaces," *IEEE Trans. Systems, Man, and Cybernetics, Part A*, vol. 35, no. 1, pp. 145-163, Jan. 2005.

[2] T. B. Moeslund, E. Granum, "A Survey of Computer Vision-Based Human Motion Capture," Computer Vision and Image Understanding, vol. 81, pp. 231-268, 2001.

[3] D. M. Gavrila, "The Visual Analysis of Human Movement: A Survey," *Computer Vision and Image Understanding*, vol. 73, no. 1, pp. 82-98, Jan. 1999.

[4] J. K. Aggarwal, Q. Cai, "Human Motion Analysis: A Review," *Computer Vision and Image Understanding*, vol. 73, no. 3, pp. 428-440, 1999.

[5] G. K. M. Cheung, T. Kanade, J.-Y. Bouguet, M. Holler, "A Real Time System for Robust 3D Voxel Reconstruction of Human Motions," *Proc. IEEE Conf. on Comp. Vis. Patt. Recog.*, vol. 2, pp. 714-720, Jun. 2000.

[6] I. Haritaoglu, D. Harwood, L. S. Davis, "W4: Real-Time Surveillance of People and Their Activities," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 809-830, Aug. 2000.

[7] R. Polana, R. Nelson, "Recognizing Activities," *Proc. Int'l. Conf. Patt. Recog.*, vol. 1, pp. 815-818, Oct. 1994.

[8] G. Mori, J. Malik, "Estimating Human Body Configurations using Shape Context Matching," *Proc. Euro. Conf. Comp. Vis.*, pp. 666-680, May 2002.

[9] K. Grauman, G. Shakhnarovich, T. Darrell, "Inferring 3D Structure with a Statistical Image-Based Shape Model," *Proc. IEEE Int'l. Conf. on Computer Vision*, pp. 641-648, Oct. 2003.

[10] M. Furukawa, Y. Kanbara, T. Minato, H. Ishiguro, "Human Behavior Interpretation Systems Based on View and Motion-Based Aspect Models," *Proc. IEEE Int'l. Conf. on Robotics and Automation*, pp. 4160-4165, Sep. 2003.

[11] I. Mikić, M. Trivedi, E. Hunter, P. Cosman, "Human Body Model Acquisition and Tracking Using Voxel Data," *Int'l. J. of Computer Vision*, vol. 53, no. 3, pp. 199-223, 2003.

[12] J. Yamato, J. Ohya, K. Ishii, "Recognizing Human Actions in Time Sequential Images using Hidden Markov Models," *Proc. IEEE Int'l. Conf. on Computer Vision and Pattern Recognition*, pp. 379-385, Jun. 1992.

[13] S. Belongie, J. Malik, J. Puzicha, "Shape Matching and Object Recognition using Shape Contexts," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 24, no. 4, pp. 509-522, Apr. 2002.

[14] A. Laurentini, "The Visual Hull Concept for Silhouette-Based Image Understanding," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 16, no. 2, pp. 150-162, Feb. 1994.

[15] K. S. Huang, M. M. Trivedi, "Video Arrays for Real-Time Tracking of Persons, Head and Face in an Intelligent Room," *Machine Vision and Applications*, vol. 14, no. 2, pp. 103-111, Jun. 2003.

[16] R. Szeliski, "Rapid Octree Construction from Image Sequences," *CVGIP: Image Understanding*, vol. 58, no. 1, pp. 23–32, 1993.

[17] K. S. Huang, M. M. Trivedi, "Robust Real-Time Detection, Tracking, and Pose Estimation of Faces in Video Streams," *Proc. Int'l. Conf. on Pattern Recognition*, vol. 3, pp. 965-968, Aug. 2004.

[18] K. S. Huang, M. M. Trivedi, "Streaming Face Recognition using Multicamera Video Arrays," *Proc. Int'l. Conf. on Pattern Recognition*, vol. 4, pp. 213-216, Aug. 2002.

[19] K. Sayood, *Introduction to Data Compression*, Morgan Kaufmann Publishers, 1996.

[20] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257-286, Feb. 1989.

[21] T. Mori, Y. Segawa, M. Shimosaka, T. Sato, "Hierarchical Recognition of Daily Human Actions Based on Continuous Hidden Markov Models," *Proc. IEEE Int'l. Conf. on Auto. Face Gest. Recog.*, pp. 779-784, May 2004.

[22] A. Ali, J. K. Aggarwal, "Segmentation and Recognition of Continuous Human Activity," *Proc. IEEE Wksp. on Detection and Recognition of Events in Video*, pp. 28-35, July 2001.

[23] Y. Rui, P. Anandan, "Segmenting Visual Actions based on Spatio-Temporal Motion Patterns," *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, vol. 1, pp. 111-118, Jun. 2000.