Extended Evaluation of XZ-Shape Histogram for Human-Object Interaction Activity Recognition based on Kinect-like Depth Image

M.A.AS'ARI[1,2], U.U.SHEIKH[3], A.H. OMAR[1,2], N.A.ZAKARIA[3], N.H.MAHMOOD[3]
[1]Faculty of Biosciences and Medical Engineering, Universiti Teknologi Malaysia, Johor Bahru, Malaysia
[2]Sport Innovation and Technology Center (SITC), Institute of Human Centered Engineering (IHCE), Universiti Teknologi Malaysia
[3]Faculty of Electrical Engineering, Universiti Teknologi Malaysia, Johor Bahru
Malaysia
amir-asari@biomedical.utm.my, usman@fke.utm.my, aho@biomedical.utm.my,
norainiz@fke.utm.my, nasrul@fke.utm.my

Abstract: - In this paper, we extend our previous work in investigating the performance of XZ-shape histogram for recognizing human performing activities of daily living (ADLs) which focuses on human-object interaction activities based on Kinect-like depth image. The feasibility of XZ-shape histogram as well as general 3D shape descriptors namely; 1) shape distribution, 2) shape histogram, 3) global spin image and 4) local spin image, in recognizing human-object interaction was tested using RGBD-HOI dataset. Moreover, the proposed evaluation framework was formulated to infer the descriptors' performance. It was found that, the XZ-shape histogram outperformed other general 3D shape descriptors that compares the performance inferred by the area under receiver operating characteristic curve (AUC-ROC). The results of this study not only demonstrate the implementation of 3D shape descriptor in the dynamic of human activity recognition but also challenge the other general 3D shape descriptor in terms of providing low dimension descriptor that capable in improving the discrimination power of human-object interaction activity recognition.

Key-Words: - human-object interaction; activities of daily living (ADLs); RGBD image; shape distribution; spin image; shape histogram.

1 Introduction

Monitoring activities of daily living (ADLs) plays a major part in assessing the health status of a person suffering with either cognitive [1] or physical impairment[2, 3] which is commonly done by human caregiver or healthcare practitioner. Recently, there are many investigations emerged on developing automated system for monitoring the activities of daily living (ADLs) which can be divided into vision-based and non-vision based system[4]. However, a rapid growing of the vision-based ADLs monitoring system development in this few years has promised the practicality of this sensing modality [4] over the non-vision based ADLs monitoring system: 1) manage to track and sense gross and fine human movements that represent ADLs; 2) provide rich of information such as spatial information, patient characteristics and anomaly actions obtained using a single vision-based sensing agent; 3) easily set up according to the conditions and environments; and 4) has high user or patient acceptance due to the non-invasive modality.

Vision based ADLs monitoring system has been investigated widely within the computer vision community [5-7]. However, most of the previous studies emphasized more on the activities without the manipulation of objects such as walking, running and jumping; which are out of healthcare community's interest. This is because the community of healthcare focuses on monitoring the home and indoor ADLs such as drinking, reading or answering a phone which are categorized into the activities that involve object manipulation or human-object interaction.

Microsoft Kinect sensor was introduced initially for the purpose of gaming and uses a clever combination of RGB and depth camera. Due to its controller-free characteristics, its role was extended to several other fields such as automated sign language [8, 9], object recognition [10-12] and human detection [13]. However, many studies done in recent years focused on implementing the Kinect for human activity recognition [14] especially human-object interaction activities [15-17].

This present study involves with the ADLs recognition that focuses on object manipulation activities or human-object interaction based on Kinect-like depth image. The extensive evaluation on the performance of XZ-shape histogram (which was proposed in our previous study [18]) that compared with general 3D shape descriptors (according to [19, 20]) in recognizing human-object interaction based on Kinect-like depth image using the RGBD-HOI dataset [18]. This study has been carried out using evaluation framework from the previous study [18] to infer the descriptors' performance.

The coordination of this paper is as follows. We review the existing approach in vision based human activity recognition and 3D shape descriptor in next section. After that, the formulation of XZ-shape histogram and the general 3D shape descriptors using RGBD-HOI dataset (see figure 1) as well as the evaluation framework are presented in Section 3. The evaluation result is explained in Section 4 before we discuss and conclude in Section 5.

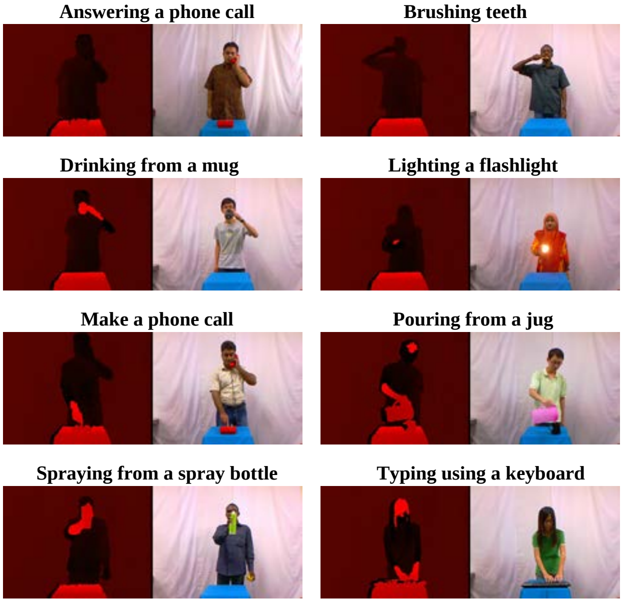


Fig.1: Example samples from RGBD-HOI dataset.

2 Related works on Human Activity Recognition and 3D Shape Descriptor

In this section, the existing vision-based human activity recognition is discussed as an overview of the current approach in extracting meaningful feature as well as general 3D shape descriptor that has been used for 3D object retrieval.

2.1 Human Activity Recognition

Human activity recognition has been widely investigated by the computer vision community. In general, the proposed approaches can be categorized into three level; 1) low-level; 2) middle-level; and 3) high-level [4] in parallel with the three levels of taxonomy activity. Since the previous studies were based on the RGB camera or video, there were several approaches developed according to the color or RGB image in extracting meaningful information to infer the human activity. The approaches has been reviewed in our previous study [4].

However, with the introduction of Kinect to the research community [21], many studies found exploring different perspectives like the depth information or combination between color (RGB) and depth information for object classification [10, 12, 22], human detection [13], automated sign language interpretation [8, 9] and human activity recognition [10, 12, 23]. A study done by Lang, [8] was the pioneer study in accessing depth information from Kinect in order to interpret the human activity. It was done by establishing a-bag-of-3D point from the depth image before inferring the human activity from action graph.

However, there were also a few studies done by combining the RGB and depth information in recognizing human activities [17, 24-27]. Since, spatio-temporal based descriptor generated based on the recent research interest in recognizing human activity by using the RGB camera or video, there were many studies found [24, 25, 27] that implement such descriptor in RGBD image for the similar interest. Study in [24] formulated hyper cuboid 4D from gradient which is taken from the interest point of RGB and depth image. Interest point was selected based on 2D Gaussian filter in spatial domain and 1D Gabor filter in temporal domain for both RGB and depth image. However, investigation in [25] recommended that, it is important to select the interest point solely on the RGB image before the bag-of-words was generated as descriptive histogram; while correspondence interest point of depth image was used to obtain depth information that separates the descriptive histogram into several depth channels. In line with this study, Zhao [27] performed the Histogram of Gradient (HOG) and Histogram of Flow (HOF) from the interest point of RGB image. However, local depth pattern which is adapted from local binary pattern (LBP) was generated from the correspondence depth interest point before classifying the human activity. Another approach was proposed by [26] which is modeling the probabilistic graphical model for human activities based on the joint of 3D point skeleton provided in Kinect. However, to our knowledge the only work which was focused on the human-object interaction activity was proposed by Koppula [17]. The study later was extended as in [26] introduced features of object manipulation as the contextual features to be used in improving the human activity recognition.

Therefore, this study put a highlight on the human-object interaction activity recognition. However, this study extend the evaluation of proposed XZ-shape histogram [18] for interpreting the human-object interaction activities and compared with other general 3D shape descriptors.

2.2 3D shape descriptor

3D shape descriptor can be categorized into four types: (1) Global based descriptor, (2) Local based descriptor, (3) View based descriptor, and (4) Graph-based descriptor. Originally, the 3D shape descriptor was designed for 3D object retrieval which is useful for the field of archeology, biology, anthropology and industrial part designing community. 3D object in a form of 3D mesh surface is commonly used for 3D object retrieval since such form has the capability to illustrate complex shape in a small memory capacity as compared to the 3D point cloud or 3D primitive form [28].

Global based descriptor describes the 3D object in terms of global shape or overall shape. The initial attempt of the descriptor was to generate the 3D object volume, moment and Fourier transform coefficients[29]. Other than that, a study in [30] suggested the convex-hull from the 3D object to be the 3D shape descriptor while several other studies, concentrated on extracting the shape [31] and shape distribution [32, 33]. However, there were also several investigations that demonstrate local based descriptor, which it describes the shape based on the geometric relation between local points in 3D object surface with neighbor points. The examples of local based descriptors are spin image [34] and curvature based descriptor [35, 36]. Graph-based descriptor interprets the 3D object shape in a form of simple informative skeleton such as Medial Scaffold [37] and Reeb Graph[38]. Meanwhile for view-based descriptor, the 3D object is illustrated into 2D view images first before determining the descriptor from the 2D view image. The example of approaches used in this category are Light-Field Descriptor [39], Characteristic view descriptor [40] and elevation descriptor [41].

3 General 3D Shape Descriptors' Extraction and XZ-Shape Histogram

In this section, the preprocessing formulated for depth frame in RGBD-HOI dataset is discussed (in Section 3.1) before the extraction of general 3D shape descriptors as well as XZ-Shape Histogram from the resultant image after preprocessing are presented in Section 3.2, 3.3, 3.4 and 3.5.

3.1 Preprocessing

Before extracting the 3D shape descriptors from depth frame in RGBD-HOI dataset, preprocessing was carried out on the depth frame as illustrated in Figure 2. The RGB frame in Figure 2a was not utilized in this work, as it was only for illustration purposes. During preprocessing, fixed bounding box (see Figure 2b) for each subject was defined manually per sample to highlight the region of interest for the purpose of avoiding unnecessary clutter. After that, multilevel thresholding was formulated on the region of interest in order to remove the background pixels as well as to retain pixels of interest. In this work, the minimum and maximum threshold values were set to 680 and 830 for the entire depth frame in the dataset as the subject performing the human-object interaction activities was within that range of depth value. The retained depth pixel was then converted into 3D points cloud (see Figure 2c) using the approach that was demonstrated in [12].

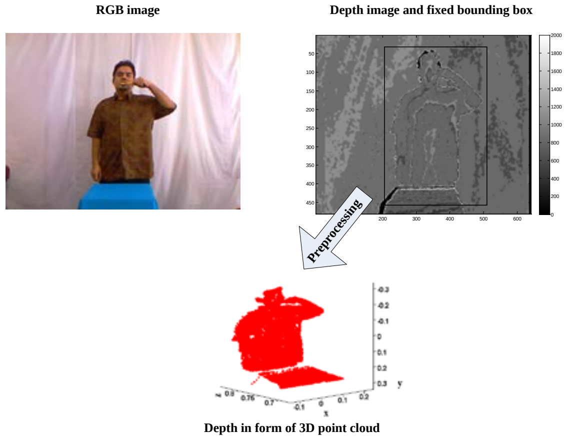
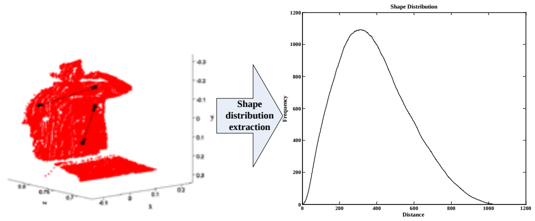


Fig.2: Preprocessing from depth image

3.2 Shape Distribution Extraction

Shape distribution [32, 33] is one of the common 3D shape descriptors that was designed for 3D object representation in a form of mesh triangulation surface. It computes distribution function based on geometric properties of 3D object surface such as angle, distance, areas and volumes from 3D points that lies within the 3D surface and are randomly selected; named as A3, D1, D2, D3 and D4. However, D2 was extracted as in our previous work [42] to obtain shape distribution directly from depth image since D2 require less computation demand. As illustrated in Figure 3a, a pair of 3D points were chosen randomly within 3D point cloud and Euclidean distance was measured between both 3D points. This mechanism was repeated in order to obtain a set of distances before the shape distribution is generated (see Figure 3b).



(a) (b)

Fig.3: Shape distribution generated from 3D point cloud

3.3 Shape Histogram Extraction

Another common 3D shape descriptor is the shape histogram which can model the 3D object's surface into: (1) histogram based on shell; (2) histogram based on sector; or (3) histogram based on fusion between shell and sector model [31]. In this work, histogram based on shell was utilized due to its invariance to rotation. As can be seen in Figure 4a, the extraction of shape histogram begins with first localizing the centroid O_{sh} for each 3D point cloud using the center value of obtained from the minimum and maximum values of x, y and z coordinates among the 3D points. From there, P number of shells is constructed with equal space between each other and with respect to the O_{sh}. The number of 3D points that resides within a shell is then calculated in order to obtain shape histogram as depicted in Figure 4b. In our work, $P=500$ and $P=1000$ were used to obtain two sets of shape histogram for the purpose of comparison.

shape histogram mechanism from 3D point cloud

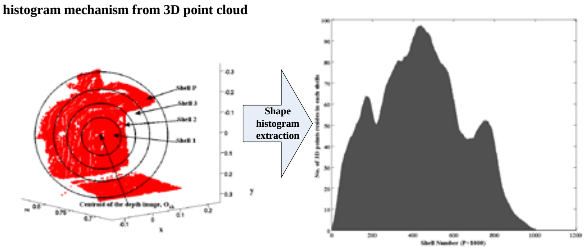


Fig.4: Shape histogram generated from 3D point cloud

3.4 Spin Image Extraction

Spin image [34] is different from the aforementioned 3D shape descriptors(shape distribution and shape histogram) as spin image is a local based shape descriptor. 2D histogram can be generated from point, O_{si} in 3D object surface in representing local shape relation between the O_{si} and its neighbouring points. This 2D histogram illustrates the number of neighbouring points, G that resides in each bin in the form of the cylindrical coordinates (a,b) with respect to the normal and tangent surface vectors of 3D points O_{si} (see Figure 5a and Figure 5b). The mapping of all neighbouring points into cylindrical axes with respect to 3D points O_{si} is formulated with the following equation,

$$S_o : R^3 \rightarrow R^2$$

$$S_o(x)\rightarrow(\alpha,\beta)=\left(\sqrt{\sqrt{\|x-p\|^2-\left(n\bullet(x-p)\right)^2}}, n\bullet(\right. \qquad (1)$$

where p and x are the 3D coordinates (in Cartesan) of points O_{si} and G while n is the normal surface vector with respect to points O_{si}. In this study, there are two types of spin image that are proposed:

1) Global spin image - a common centroid is determined from all 3D points as O_{si} and this centroid is used to generate spin image.
2) Local spin image - several local centroids from local regions were assigned as several O_{si} points before several 2D histograms were established as spin image. The local region is defined by separating region of interest into several equally sized sub regions.

Figure 6a illustrates the centroid used as O_{si} for global spin image while Figure 6b and Figure 6c depict the centroids used as several O_{si} points whereas the local spin image was fragmented into local spin image with; 1) two local centroids (see Figure 6b); and 2) four local centroids (see Figure 6c). These two types of local spin image were considered since simulating local spin image with more than four local centroids is rather time-consuming.

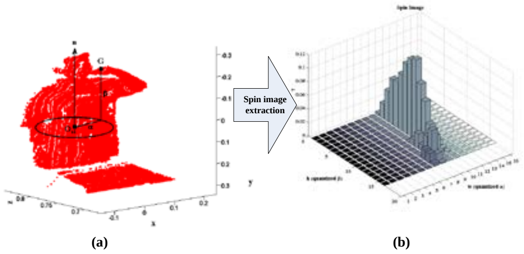


Fig.5: Shape image generated from 3D point cloud

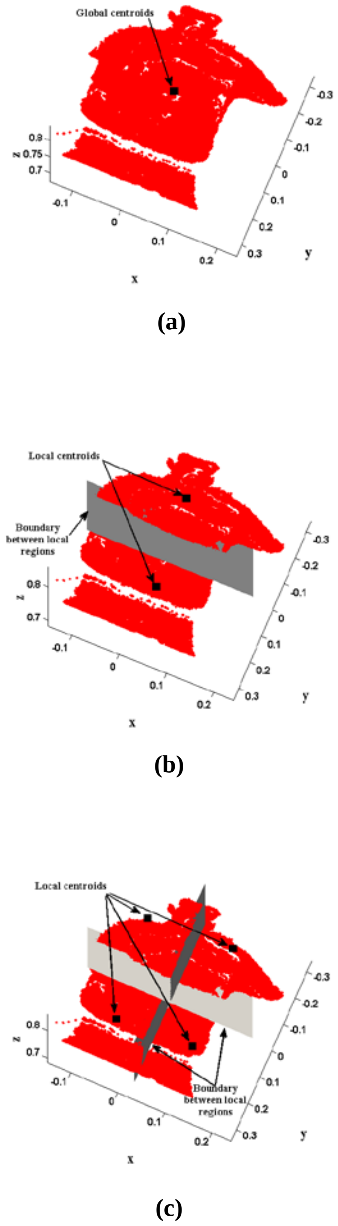


Fig.6: Centroid point assigned as O_{si} in extracting spin image for;

(a) Global spin image; (b) local spin image with two local centroids; and (c) local spin image with four local centroids

3.5 Proposed XZ-Shape Histogram Extraction

XZ-shape histogram was generated by merging the X-shape histogram and Z-shape histogram. X-shape histogram was formulated based on shell model which is in a form of plane with surface normal in x-axis direction (see Figure 7). The drawback of X-shape histogram is that the descriptor incapable in differentiating the left-handed or right-handed subject (see Figure 8(a)-(d)). Figure 8 (b) and (d) display the X-shape histogram generated from left-handed subject (see Figure 8(a)) and right-handed subject (see Figure 8(b)) which are dissimilar in shapes. In order to overcome this problem, each generated X-shape histogram was flipped on condition that the frequency of the first bin is more than the last bin. The output for flipping the histogram can be identified in Figure 8(e) that was formulated from X-shape histogram in Figure 8(d).This mechanism managed to correct the X-shape histogram for depth image in Figure 8 (b) which was similar to X-shape histogram generated from depth image in Figure 8(a). However, Z-shape histogram was prepared based on modeling the shell for several planes with surface normal in z-axis direction (see Figure. 9). Therefore, with the use of this approach, there was no issue of left-handed or right-handed subject.

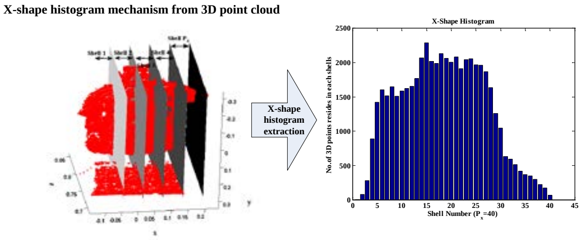
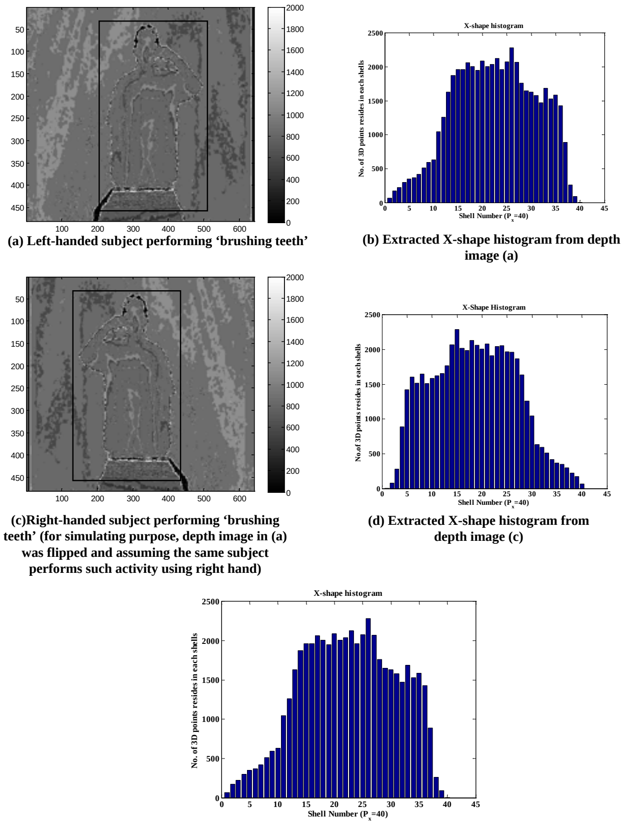


Fig.7: X-shape histogram extracted 3D point cloud



(a) Left-handed subject performing 'brushing teeth'

(b) Extracted X-shape histogram from depth image (a)

(c)Right-handed subject performing 'brushing teeth' (for simulating purpose, depth image in (a) was flipped and assuming the same subject performs such activity using right hand)

(d) Extracted X-shape histogram from depth image (c)

(e) X-shape histogram of (d) after flipping histogram process which is similar to histogram in (b)

Fig. 8: The flipping histogram process in order to avoid right-handed and left-handed subject occur in generating X-shape histogram

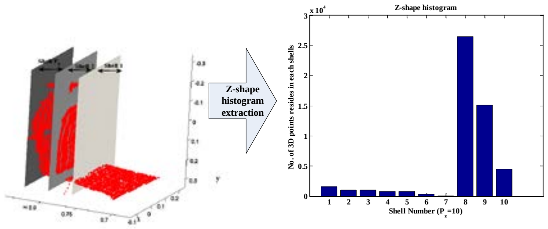


Fig. 9: Z-shape histogram extracted 3D point cloud2. Evaluation Framework

In our previous study [18], the performance of X-shape histogram with P_x =5, 10, 20, 30 and 40; and Z-shape histogram with P_z=3, 5 and 10 was evaluated. It was found that X-shape histogram with P_x =5 and Z-shape histogram with P_z=5 provide excellent performance as compare to other X and Z-shape histogram configuration. Thus, both aforementioned best configurations were used to formulate the XZ-shape histogram for this study (by concatenating the best X-shape histogram and the best Z-shape histogram).

3.6 Evaluation Framework

In order to evaluate the feasibility of the aforementioned 3D shape descriptors in recognizing the human-object interaction activity, the evaluation mechanism from our previous work in [42] was used by modifying the evaluation scheme. The performance of 3D shape descriptors were represented in terms of receiver operating characteristic (ROC) and area under ROC curve (AUC-ROC).

From a set of depth frame $T_i = \left\{ t_f \right\} \mid f = \left\{ 1, 2, 3 \ldots F \right\}$ representing a subject performing an activity class; f is the frame index in an activity class, the correspondence set of 3D shape descriptor, $T_i' = \left\{ t_f' \right\} \mid t_f' \in \mathbb{R}^d$ was extracted from T_i ; where i is the index number represents each sample of activity in the dataset. Each T_i' later was quantized using vector quantization approach in order to reduce the 3D shape descriptor from d dimension into one dimensional symbol; and produced the correspondence $T_i^q = \left\{ t_f^q \right\} \mid t_f^q \in \mathbb{R}^1$. This process was computed using K-means to cluster each 3D shape descriptor frame into K number of symbols. In our study, the evaluation was simulated based on several numbers of K before finding the best K which is appropriate to be incorporated with each type of general 3D shape descriptor and to be compared with other 3D shape descriptors. Then, the self-similarity matrix S was established and defined as follows,

$$S(i,j) = s\left(T_i^q, T_j^q\right) \text{ s.t } \quad i,j = \left\{ 1, 2, \ldots N \times C \right\} \qquad (2)$$

where N is set to 12 as the number of subjects per action while C is 8 as the total classes that exist in our dataset. $s\left(T_i^q, T_j^q\right)$ is defined as the similarity measurement function that is used for calculating the similarity value between two sequences of activities, T_i^q and T_j^q; as an element in matrix S.

In this investigation, edit distance [43] was implemented as the similarity measurement function due to its capability of comparing two entities that are represented in the form of features sequence. In our work, the edit distance weight constant $w = [w_{sub}\ w_{ins}\ w_{del}]$ that represents the weight for substitute, insert and delete operation were set equally to $w = [1\,1\,1]$.

In establishing self-similarity matrix S, the row or column expressed as T which are based on element T_i^q or element T_j^q were arranged and built according to the following equations,

$$T_c^q = \left\{T_n^q\right\} \mid n = \{1, 2, 3 \ldots.N\} \qquad (3)$$

$$T \ = \left\{T_c^q\right\} \mid c = \{1, 2, 3 \ldots.C\} \qquad (4)$$

An example of formulating (3) and (4) to arrange row and column elements of S can be observed in Figure 7a.

Since our intention is to generate the Receiver-Operator-Characteristic (ROC) curve for each type of 3D shape descriptor that incorporates different number of symbols K in vector quantization process, self-similarity matrix S was formulated with the correspondence ground-truth matrix S_{GT}. An example of the corresponding S_{GT} for matrix S in Figure 7a can be seen in Figure 7b. S_{GT} was built with binary numbers consisting of 1(white pixel) and 0(black pixel) as indicators for similar and dissimilar classes. It is arranged so that white blocks of $N \times N$ pixels are replicated in diagonally within the matrix S_{GT} of the size $\left(N \times C\right) \times \left(N \times C\right)$ pixels. Thus, true similar t_s, true dissimilar t_d, false similar f_s and false dissimilar f_d were calculated from S and S_{GT} using the following equations:

$$t_s = \sum_i^{NC}\sum_j^{NC}\{S_{GT}(i,j) \times S^{'}(i,j)\} \qquad (5)$$

$$t_d = \sum_i^{NC}\sum_j^{NC}\{(1 - S_{GT}\left(i,j\right)) \times (1 - S^{'}(i,j))\} \qquad (6)$$

$$f_s = \sum_i^{NC}\sum_j^{NC}\{(1 - S_{GT}\left(i,j\right)) \times S^{'}(i,j)\} \qquad (7)$$

$$f_d = \sum_i^{NC}\sum_j^{NC}\{S_{GT}(i,j) \times (1 - S^{'}(i,j)\} \qquad (8)$$

where $S^{'} \in \{1,0\}$ is computed by thresholding the matrix S using threshold value ε which is varied between 0 to 1. Finally, true-positive rate, $TPR(\varepsilon)$ and false-positive rate, $FPR(\varepsilon)$ are derived as,

$$TPR = \frac{t_s}{t_s + f_d} \qquad (9)$$

$$FPR = \frac{f_s}{f_s + t_d} \qquad (10)$$

before ROC curve for each case is obtained. In our work, ROC curve for each 3D shape descriptor with the same K number of symbols is generated three times and the average ROC was obtained and the area under ROC curve (AUC-ROC) was calculated to represent such test case to avoid any clustering error caused by K-means in vector quantization approach.

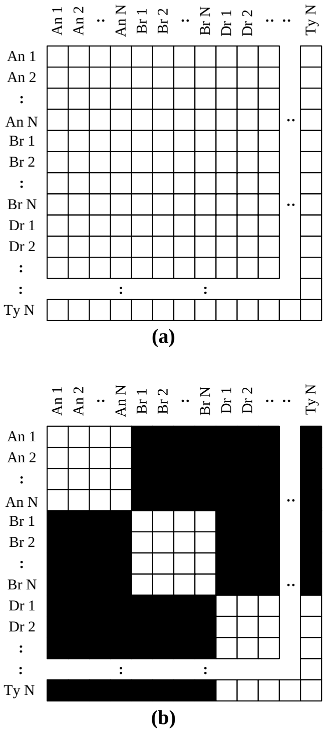


Fig. 7: (a) Self-matching matrix M (partially shown); both row and column were formulated using (2) and (3). (b) Ground-truth matching matrix, M_t which was built for self-matching matrix in (a).

4 Results and Discussion

The main purpose of this work is to carry out the analysis on the feasibility of XZ-shape histogram and general 3D shape descriptors (shape distribution, shape histogram, global and local spin image) for human-object interaction activity recognition based on depth images from our RGBD-HOI dataset.

The overall feasibility of shape distribution in terms of performance is reported in Table 1. Based on the AUC-ROC value, it is found that the performance of shape distribution strikingly increased once the K symbols used are more than 100. At $K = 20$, there were a lot of spatial information loss during the vector quantization process which resulted in different pose frame that represents different activity lies on the same cluster symbol. However, once K was increased to more than 100, the performance drastically increased but with slower increment in performance can be seen when K was set to above 100. The best performance of shape distribution descriptor in this work was at $K = 300$ (the grey highlighted cell). However, there was a trade-off between increasing the performance of shape distribution and the complexity of vector quantization if the K is increased.

Table 1. AUC-ROC for Several Shape Distribution based on Different K symbols

Number of K symbols used in vector quantization process	**AUC-ROC**
20	0.6345
100	0.6460
200	0.6458
300	0.6479

On the other hand, it seems that there are a variety conditions for testing the performance of the shape histogram using several pairs of P (as the number of shells used to formulate shape histogram) and K (as the number of symbols used in vector quantization process). The results are represented as AUC-ROC value which can be seen in Table 2 (the grey highlighted cells are the maximum AUC-ROC values within different K number of symbols). When P was set to 500, the trend of shape histogram's performance is almost similar to shape distribution. The performance increased significantly at $K = 100$ and marginally improved with K above 100. However, this behaviour is considerably different for shape histogram when $P = 1000$. As can be observed in Table 2, a remarkable performance increment can be observed at $K = 200$ and higher. This is due to the dependency between the number of shells P used in the extraction of shape histogram and the number of K symbols used in the vector quantization process. If the P used is too small, then it is insufficient to describe the shape of 3D surface. However, increasing P too much can cause poor discriminating descriptor as fewer number of 3D points reside within each shell. In deciding the K symbols to be used in the vector quantization process, the number of P shells used to extract shape histogram must be considered. It can be found in Table 2 that the K used to produce excellent shape histogram performance for $P = 500$ was different from shape histogram for $P = 1000$.

Spin image is one of the local based 3D shape descriptors that provides the local shape properties based on a single 3D point in 3D surface. In this study, two types of spin image descriptors were formulated (global and local spin image) for the purpose of evaluating the spin image capability. Table 3 displays the performance of these two types of spin image in terms of AUC-ROC value with different number of K symbols used during vector quantization process. The trend of global spin image performance is inline with the previous 3D shape descriptor as the performance significantly increased once K is increased to more than 100. However, both local spin image descriptors with two local centroids and four local centroids only performed exceptionally at $K = 20$ before insignificantly degrading when $K = 100$ or higher.

Table 2. AUC-ROC for Several Shape Histogram based on Different K symbols and P number of Shells

		Number of cells used to establish shape histogram, P	
		500	**1000**
Number of K symbols used in vector quantization process	**20**	0.6402	0.6339
	100	0.6457	0.6393
	200	0.6431	0.6452
	300	0.6436	0.6455

Overall, based on the maximum AUC-ROC value (grey cell in Table 3) in each type of spin image, local spin image manages to outperform global spin image since the local spin image is capable of extracting the meaningful local shape properties from depth frame that represents the human-object-interaction activity. However, an attempt to increase the local shape properties from local spin image by increasing the number of local centroids used from two to four points is unworkable. This is due to the active region in performing human-object-interaction is hand. Thus, only few local regions where hand is located will produce significant shape changes in completing human-object-interaction cycle while other local regions (passive regions) will remain with same shape properties. Increasing the number of local centroids formulated from local regions might reduce the performance of local spin image as the obtained local spin image is established by merging between descriptors from active and passive local regions. This might cause the sensitivity of active regions in representing the human-object-interaction activity to reduce as the passive region influences the descriptor.

Table 3. AUC-ROC for Several Types of Spin Image based on Different K symbols used.

		Type of spin image		
		Global spin image	**Local spin image with two local centroids**	**Local spin image with four local centroids**
Number of K symbols used in vector quantization process	**20**	0.6337	0.6479	0.6410
	100	0.6418	0.6369	0.6311
	200	0.6455	0.6379	0.6400
	300	0.6445	0.6398	0.6386

The overall performances of 3D shape descriptors are summarized in Table 4 whereas each type of 3D shape descriptor is represented by the one that showed excellent performance (based on AUC-ROC value) among the same type of 3D shape descriptor. This includes with proposed XZ-shape histogram which was formulated according to the previous study [18]. It is evident that by utilizing the local spin image, the meaningful local shape properties which are comparable to shape distribution that extracts global shape properties from the depth frame. However, the proposed XZ-shape histogram consists of less dimensional space achieved a remarkable performance compared to both shape distribution and local spin image in interpreting the human object interaction.

Table 4. AUC-ROC for different types of 3D shape descriptors

3D Shape Descriptors	**AUC-ROC**
Shape Histogram. P=500, K=100	0.6457
Shape Distribution, K=300	0.6479
Global Spin Image. K=200	0.6455
Local Spin Image, K=20	0.6479
XZ-shape histogram. K=300	0.6484

5 Conclusion

In summary, this study extend the evaluation of proposed XZ-shape histogram compared to other general 3D shape descriptors in recognizing the human-object interaction activity based on Kinect-like depth image using RGBD-HOI dataset. The proposed XZ-shape histogram achieved a remarkable performance as compared to other general 3D shape descriptors especially shape distribution and local spin image which are the close competitors. Shape distribution and local spin image suffer with high dimensionality and complexity while the proposed XZ-shape histogram manages to solve the problem and capable to improve the discrimination power of human-object interaction recognition.

This study also provides the framework for future studies especially in incorporating classifier mechanism with 3D shape descriptors to develop whole system for human activity recognition based on the Kinect camera. This can eventually lead to the embedment of Kinect as the vision-based sensor used in medical and rehabilitation for accessing subject performing activity of daily living (ADLs) in home and healthcare centers.

6 Acknowledgment

The author would like to express their gratitude to University Teknologi Malaysia (UTM) and the Minister of Higher Education (MOHE), Malaysia for supporting this research.

References:

[1] J. Hoey, P. Poupart, A. v. Bertoldi, T. Craig, C. Boutilier, and A. Mihailidis, Automated handwashing assistance for persons with dementia using video and a partially observable Markov decision process, Computer Vision and Image Understanding, Vol. 114, 2010,pp. 503-519.

[2] J. S. Brach and J. M. VanSwearingen, Research Report Physical Impairment and Disability : Relationship to Performance of Activities of Daily Living in Community-Dwelling Older Men, Physical Therapy, Vol. 8, 2002,pp. 752-761.

[3] A. Paraschiv-Ionescu, C. Perruchoud, E. Buchser, and K. Aminian, Barcoding Human Physical Activity to Assess Chronic Pain Conditions, PLoS ONE, Vol. 7, 2012,p. e32239.

[4] M. A. As'ari and U. U. Sheikh, Vision based assistive technology for people with dementia performing activities of daily living (ADLs): an overview, in Fourth International Conference on Digital Image Processing (ICDIP), Kuala Lumpur, 2012, pp. 83342T-83342T.

[5] T. B. Moeslund, A. Hilton, and V. Krger, A survey of advances in vision-based human motion capture and analysis, Comput. Vis. Image Underst., Vol. 104, 2006,pp. 90-126.

[6] Aaron F. Bobick, Movement, activity and action: the role of knowledge in the perception

of motion, Philosophical Transactions of the Royal Society B:Biological Sciences, Vol. 352, 1997,pp. 1257-1265.

[7] G. Lavee, E. Rivlin, and M. Rudzsky, Understanding Video Events: A Survey of Methods for Automatic Interpretation of Semantic Occurrences in Video, IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, Vol. 39, 2009,pp. 489-504.

[8] S. Lang, "Sign Language Recognition with Kinect," Bachelor, Institut für Informatik, Freie Universitä t Berlin, 2011.

[9] Z. Zafrulla, H. Brashear, T. Starner, H. Hamilton, and P. Presti, American sign language recognition with the kinect, in Proceedings of the 13th international conference on multimodal interfaces, Alicante, Spain, 2011, pp. 279-286.

[10] L. Bo, K. Lai, X. Ren, and D. Fox, Object Recognition with Hierarchical Kernel Descriptors, in Proc. of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), 2011.

[11] K. Lai, L. Bo, and X. Ren, Detection-based Object Labeling in 3D Scenes, 2012.

[12] K. Lai, L. Bo, X. Ren, and D. Fox, A Large-Scale Hierarchical Multi-View RGB-D Object Dataset, in Proc. of International Conference on Robotics and Automation (ICRA), 2011.

[13] L. Xia, C. Chen, and J. K. Aggarwal, Human Detection Using Depth Information by Kinect, in International Workshop on Human Activity Understanding from 3D Data in conjunction with CVPR (HAU3D), 2011.

[14] L. Wanqing, Z. Zhengyou, and L. Zicheng, Action recognition based on a bag of 3D points, in 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2010, pp. 9-14.

[15] M. Popa, A. K. Koc, L. J. M. Rothkrantz, C. Shan, and P. Wiggers, Kinect Sensing of Shopping related Actions, in Constructing Ambient Intelligence: AmI 2011 Workshops, 2011.

[16] J. Sung, C. Ponce, B. Selman, and A. Saxena, Human Activity Detection from RGBD Images, in AAAI workshop on Pattern, Activity and Intent Recognition (PAIR), 2011.

[17] H. S. Koppula, R. Gupta, and A. Saxena, Learning Human Activities and Object Affordances from RGB-D Videos, arXiv preprint arXiv:1210.1207, 2012.

[18] M. AS'ARI, U. SHEIKH, and E. SUPRIYANTO, XZ-Shape Histogram for Human-Object Interaction Activity Recognition based on Kinect-like Depth Image, WSEAS Transactions on Signal Processing, Vol. 10, 2014.

[19] P. Huang, A. Hilton, and J. Starck, Shape Similarity for 3D Video Sequences of People, Int. J. Comput. Vision, Vol. 89, 2010,pp. 362-381.

[20] M. A. As'ari, U. U. Sheikh, and E. Supriyanto, 3D shape descriptor for object recognition based on Kinect-like depth image, Image and Vision Computing, Vol. 32, 2014,pp. 260-269.

[21] Kinect. (2010, 04/02/2012). http://www.xbox.com/en-us/kinect.

[22] K. Lai, L. Bo, X. Ren, and D. Fox, A Scalable Tree-based Approach for Joint Object and Pose Recognition, in Twenty-Fifth Conference on Artificial Intelligence (AAAI), 2011.

[23] K. Lai, L. Bo, X. Ren, and D. Fox, Detection-based Object Labeling in 3D Scenes, in 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2012, pp. 2169–2178.

[24] H. Zhang and L. E. Parker, 4-dimensional local spatio-temporal features for human activity recognition, in 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2011, pp. 2044-2049.

[25] N. Bingbing, W. Gang, and P. Moulin, RGBD-HuDaAct: A color-depth video database for human daily activity recognition, in 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), 2011, pp. 1147-1153.

[26] J. Sung, C. Ponce, B. Selman, and A. Saxena, Unstructured human activity detection from rgbd images, in 2012 IEEE International Conference on Robotics and Automation (ICRA), 2012, pp. 842-849.

[27] Y. Zhao, Z. Liu, L. Yang, and H. Cheng, Combing RGB and Depth Map Features for human activity recognition, in 2012 Asia-Pacific Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC), 2012, pp. 1-4.

[28] F. Lafarge, R. Keriven, and M. Bredif, Insertion of 3-D-Primitives in Mesh-Based Representations: Towards Compact Models Preserving the Details, IEEE Transactions on Image Processing, Vol. 19, 2010,pp. 1683-1694.

[29] Z. Cha and C. Tsuhan, Efficient feature extraction for 2D/3D objects in mesh representation, in International Conference on Image Processing 2001, pp. 935-938 vol.3.

[30] J. Corney, H. Rea, D. Clark, J. Pritchard, M. Breaks, and R. Macleod, Coarse filters for shape matching, IEEE Computer Graphics and Applications, Vol. 22, 2002,pp. 65-74.

[31] M. Ankerst, G. Kastenmüller, H.-P. Kriegel, and T. Seidl, 3D shape histograms for similarity search and classification in spatial databases, SSD' 99, 1999,pp. 207-226.

[32] R. Osada, T. Funkhouser, B. Chazelle, and D. Dobkin, Matching 3D models with shape distributions, in SMI 2001 International Conference on Shape Modeling and Applications, 2001, pp. 154-166.

[33] R. Osada, T. Funkhouser, B. Chazelle, and D. Dobkin, Shape distributions, ACM Trans. Graph., Vol. 21, 2002,pp. 807-832.

[34] A. E. Johnson and M. Hebert, Using Spin Images for Efficient Object Recognition in Cluttered 3D Scenes, IEEE Trans. Pattern Anal. Mach. Intell., Vol. 21, 1999,pp. 433-449.

[35] S. Heung-Yeung, M. Hebert, and K. Ikeuchi, On 3D shape similarity, in Proceedings CVPR '96, 1996 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1996, 1996, pp. 526-531.

[36] T. Zaharia and F. Preteux, Three-dimensional shape-based retrieval within the MPEG-7 framework, in Proceedings SPIE Conference on Nonlinear Image Processing and Pattern Analysis XII, 2001, pp. 133-145.

[37] M.-C. Chang and B. B. Kimia, Measuring 3D shape similarity by graph-based matching of the medial scaffolds, Computer Vision and Image Understanding, Vol. 115, 2011,pp. 707-720.

[38] M. Hilaga, Y. Shinagawa, T. Kohmura, and T. L. Kunii, Topology matching for fully automatic similarity estimation of 3D shapes, presented at the Proceedings of the 28th annual conference on Computer graphics and interactive techniques, 2001.

[39] D.-Y. Chen, X.-P. Tian, Y.-t. Shen, and M. Ouhyoung, On Visual Similarity Based 3D Model Retrieval, presented at the Computer Graphics Forum (EUROGRAPHICS'03), 2003.

[40] S. Mahmoudi and M. Daoudi, 3D models retrieval by using characteristic views, in 16th International Conference on Pattern Recognition, 2002, pp. 457-460 vol.2.

[41] J.-L. Shih, C.-H. Lee, and J. T. Wang, A new 3D model retrieval approach based on the elevation descriptor, Pattern Recognition, Vol. 40, 2007,pp. 283-295.

[42] M. A. Asari, E. Supriyanto, and U. U. Sheikh, The Evaluation of Shape Distribution for Object Recognition Based on Kinect-Like Depth Image, in Computational Intelligence, Communication Systems and Networks (CICSyN), 2012 Fourth International Conference on, 2012, pp. 313-318.

[43] P. H. Sellers, The theory and computation of evolutionary distances: Pattern recognition, Journal of Algorithms, Vol. 1, 1980,pp. 359-373.