Real-time Gesture Recognition for Interactive Presentation with Depth Sensor

Anonymous CVPR submission

Paper ID 1106

Abstract

In this paper, we propose a method to enable real-time interaction between the projection contents and speaker through detecting and recognizing meaningful human gestures from depth maps captured by depth sensor, making projection screen as a kind of touch screen. Considering that depth noise and serious occlusion may ruin the construction of skeleton, our hand trajectory is derived from Potential Active Region. To cope with their inter-class and intra-class variations, hand trajectory is temporally segmented into movements, which are represented as Motion History Images. A novel set-based soft discriminative model is learned to recognize gestures from these movements. In addition, as it is a real-time system, a complexity reduction method is employed. The proposed approach is evaluated on our dataset and performs efficiently and robustly with 90% correct recognition rate.

1. Introduction

Recently, gesture recognition has been attracting a great deal of attention as a natural human computer interface, since it allows users to control or manipulate devices in a more natural manner through intentional physical movements of figures, hands, arms, face, head, or body. So far, numerous studies [10][16] have been conducted on gesture recognition for human computer interaction, especially for hand and arm gesture recognition. Based on these technologies, a number of recognition applications are developed, including sign language recognition, game controlling, navigating in virtual environment, etc. In this paper, we present an effective and efficient arm gesture recognition algorithm that enable natural interaction between speaker and presentation contents. During presentation, speakers often stand in front of the projection screen at a distance from the machine with projection contents. It is more natural for speakers to remote control the page flipping, scrolling and clicking by arm gestures. Considering the potential light influences caused by projector on the color images, depth maps captured during presentation are employed for gesture recognition in our work. Fig. 1 shows the framework of proposed approach. In the framework, human body is segmented from noisy depth maps, and then potential active regions (PAR) are derived from head position for meaningful gesture detection. Once the hand is observed in the potential active region, its trajectory will be recorded and decomposed to a series of movements. (see green box in dash line). These movements are represented as Motion History Images (MHI) and assembled to a labeled gesture by utilizing proposed set-based soft discriminative model.

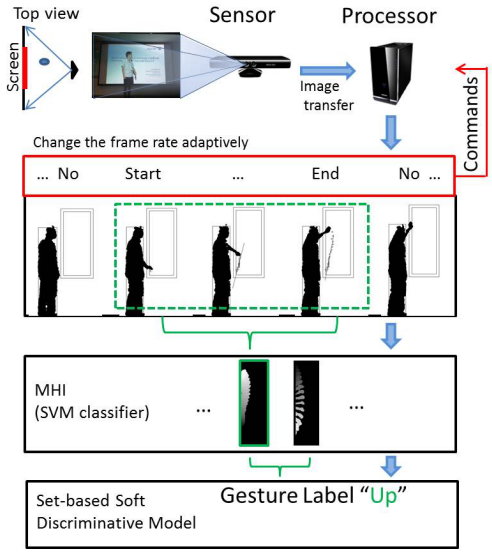

Figure 1. Framework of proposed approach.

As depth data is noisy, background suppression technique is used to segment human body from background. The location and size of human body are determined by searching a proper bounding box in the generated human body's depth map. Considering that human body may be incomplete in that bounding box because of occlusion or limited view angle of depth sensor, head detection is applied for estimating the size of the complete human body. However, normal method using face to detect head is impossible in presentation since speakers may turn their faces to the

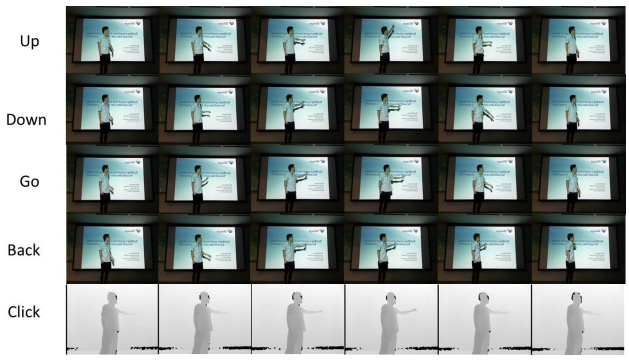

Figure 2. Five gestures ("click" is shown in depth since it is a gesture vertical to the plane).

side (not facing the sensor) and skin color is also changed due to the strong light from the projector. As a result, the only useful cue to detect head is the depth map.

Potential Active Regions of human arms (see the boxes beside body in Fig. 1) are adaptively determined by the location and size of body. Intuitively, PARs are the most discriminative regions. Therefore arm is only detected when reaching in PARs, which is shown in the framework of our system in Fig. 1. If no arm is detected in PARs, it is unnecessary to perform recognition step in this frame or even unnecessary to detect arm in the next few frames. That vastly reduces the computational complexity. Once arm is observed in a PAR, the trajectory of hand will be recorded and decomposed to a series of movements (see the green box in dash line), which are classified by Support Vector Machine (SVM). Offline training is not required to decompose the trajectory. These movements are assembled into a set, which is then labeled and understood as a specific gesture. Considering that one misclassified movement in the set may influence the result of gesture recognition. Original proposal of a set-based soft discriminative model can correct the misclassified movement and designate a most likely gesture label to the set. Experimental results show that this model has a better performance than traditional one.

Our main contribution lies in three folds. First, the noise and occlusion problems caused by depth sensor are solved by efficient preprocessing. Second, with PAR, the computational complexity is dramatically reduced by selecting active frames. Third, a set-based soft discriminate model is originally proposed, which has the ability to correct misclassified movements. To test our method, we capture a dataset including 5 gestures: "up", "down", "go", "back" and "click" (see Fig. 2).

The rest of this paper is organized as follows. Related work is introduced in section 2. Section 3 presents the proposed method and the experimental results in section 4 demonstrate the efficiency of the proposed method. Section 5 concludes this paper and gives the future directions.

2. Related Work

In the last decade, most of recognition algorithms are designed for the color images captured by monocular camera sensors. One challenge of the color image based recognition is how to efficiently segment the object from the background. In order to obtain foreground silhouettes of objects, most of the recognition algorithms are restricted to pure and static backgrounds. For example, public dataset KTH human motion dataset [12] and Weizmann human action dataset [1] both record human actions under relatively static backgrounds. The rapid development of depth sensors open up the possibilities of dealing with cluttered background by providing depth information. Even though, depth sensors like Kinect [14], Time of Flight camera [5] or stereo camera [15] still present two challenges: noise and occlusion. Noise decreases the quality of background subtraction mainly because of the limited measurement accuracy of the depth sensor. Occlusion occurs when there is an object (e.g. desk) in front of human body. It will ruin body detection because part of the human body is blocked. This is also a main problem when adapting monocular camera. What is worse, Kinect sensor regards black objects, such as black trousers or black hair, as occlusions due to its generation principle. To overcome the occlusion problem, some approaches employ the location of human face to indicate the location of human body. Face detection is usually implemented as the first step to detect human body. For example, Wang et al. [18] locate the face before obtaining skin model from face and use it to detect human hand. As stated before, they cannot work well in presentation due to the variant directions of face and abnormal skin color, which is influenced by projector light.

When equipped with depth sensor, many researchers make effort to compute 3D joint positions of human skeleton. Shotton et al. [14] provide a rather powerful human motion capturing technique. There are also many action recognition works start directly from human skeleton. For example, Jiang et al. [17] track 20 joint positions by the skeleton tracker proposed by [14] and use local occupancy pattern to represent the interaction. Sadeghipour et al. [11] also directly use the 3D joint positions of human skeleton for gesture-based object recognition. Both of them use Kinect as the depth sensor. When referred to the other depth sensors, robust and fast method has not been presented yet. That means, methods in [17] and [11] are restricted to Kinect sensor. Besides, distance from human and Kinect in both of their datasets are fixed while in a normal presentation, speakers walk around. Therefore, to make our system more natural for interaction and more general for other stereo sensors, skeleton are replaced with silhouette. By accumulating the silhouettes in PARs, MHIs are generated.

PAR is a spatial region (where), the generation of MHI still asks for a temporal region (when). In another word, the

start and end of a gesture in a video sequence should also be determined. Videos in most published 3D datasets are readily segmented into sequences that contain one instance of a known set of action labels. For example, Sadeghipour et al. [11] capture the 3D Iconic Gesture dataset and segment the video by the moment when subject retracting their hands back to the rest position, so does NATOPS Aircraft Handling Signals Database captured by Song et al. [15]. Different from their works, trajectory decomposing is employed to automatically segment the video in our approach to provide a more natural interaction.

As we know, representation of suitable feature and modeling of dynamic patterns are the two main important issues. In our work, MHI serves as the representation of the movement in the action recognition framework according to the taxonomy summarized by Bobick [2] in an early survey. Similarly, 3D low-level features are deeply studied in recent years. Most of them are extended from normal 2D features such as [9] (3D Harris corner detector), [20] (3D SURF descriptor), [8] (3D HOG descriptor) and [13](3D SIFT descriptor). However, these local featurLes are not discriminative features in textureless depth maps. To assemble low-level movements into a gesture in the proposed framework, generative model and discriminative model are the main two temporal state-space models. Generative model learns to model each class individually and always assume that the observations in time are independent, e.g., Feng et al. [4] and Weinland et al. [19]. Discriminative model is trained to discriminate between action classes and model a conditional distribution over action labels given the observation. Jordan et al. [7] compare discriminative and generative learning as typified by logistic regression and naive Bayes. Our gesture recognition model belongs to the discriminative model.

3. Method

In our framework, the background is firstly suppressed from noisy depth maps, and then PARs are derived from head position for meaningful gesture detection. The generation of PARs and the segmentation of the hand trajectory is to determine where and when to generate MHI in the video sequence. Then, a soft discriminative model is originally proposed to assemble them into one gesture. In addition, with hand detecting in the PARs, the complexity of the system is vastly reduced.

3.1. Background Suppression

In presentation, touching the screen while performing gestures is a natural way to control the projection contents. However, measurement accuracy of the depth sensor is limited. It is difficult to segment objects from the background when the depth of objects is too close to that of background. As the result, even if the depth of background has been captured, segmenting arms from screen is impossible by background subtraction techniques. It is worth noticing that the depth value of each pixel in the background is approximately Gaussian distributed with time going on, and the depth value at a fixed position lies in a certain range with a high probability. When the probability of the depth value is below a threshold, the corresponding point is processed as human body. Fig. 3 (a) shows the improvement.

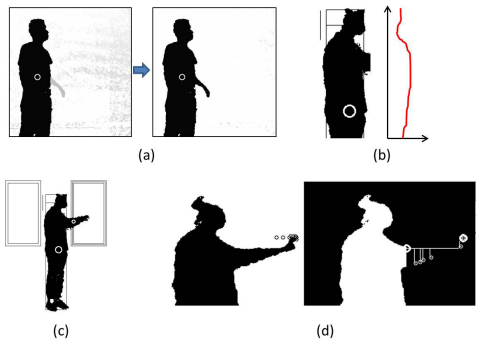

Figure 3. (a) Results of improved background suppression method. (b) Method to obtain the size and location of head. (c) Potential active regions for two arms. (d) Distance map using chamfer distance.

3.2. PAR Generation and Hand Trajectory Decomposition

The size and location of the whole human body is necessary for PAR generation. Therefore, a bounding box containing the whole body is constructed. The segmented depth map is scanned along vertical lines from left to right while recording the proportion of body pixels in the line. Location of the left border of bounding box is determined when the first time the proportion reaches a threshold. Locations of other three borders are searched in a similar way.

Notice that human body in such bounding box may be incomplete because of the occlusion or the limited view angle of sensor. Size of head is thus employed to estimate the size of the whole body according to the normal proportion of human figure. As we know, physical width of human body varies with height. In generally, the width of neck is the smallest while that of shoulder is the largest. Inspired by this observation, pixels' values are summed along horizontal direction in the body bounding box. A curve is drawn to show the proportions of human pixel in each horizontal line (see Fig. 3 (b)). After smoothing the curve by media filter, location and size of head are obtained by detecting the largest width variation. Two PARs are then constructed for left and right arms. Since PARs cover the most likely arm movement region, they can serve as the constraint for hand position tracking and adaptive detection mode transfer. The detection mode transition will be introduced in details in the

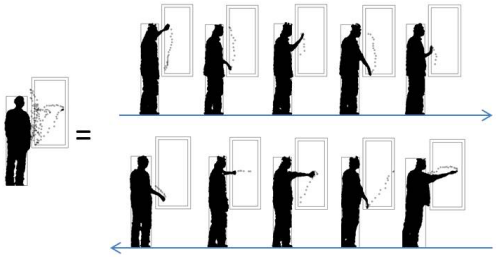

Figure 4. Decomposing the movement trajectories. The start and end points in each segmentation are interest points. 10 movements are detected in this example.

section 3.3.

A chamfer distance map is derived when hand enters the PAR. As the farthest endpoint from the edge of PAR, hand is detected and tracked in that region before leaving (see Fig. 3 (d)). However, it is difficult to recognize gestures directly from the trajectories of hand since our dataset has the following two challenges. First, the dataset has small inter-class variances. Gesture "up" is very similar to "down" while the trajectories of "go" and "back" are similar if not considering the order of frame. Second, it also has large intra-class variances. There are variant ways to act the same gesture by different subjects. Even the same subject performs the same gesture differently each time. To recognize gestures from the complex trajectory, a method is proposed to decompose the trajectory into movements, classify the movements and then assemble them to form a meaningful gesture. See from Fig. 4, the subject successively performs four actions in one video. Among the complex trajectory, interest points are required to help segmenting it into movements. As we know, interest point is the sudden change of trajectories in video sequence. Based on that common sense, the method to search the interest points is introduced as follows.

Method of Least Squares (MLS) is used to detect the interest points on the trajectories. The MLS assumes that the best-fit curve of a given type is the curve that has the minimal sum of the deviations squared (least square error) from a given set of data. The type of straight line $y = ax + b$ is employed to approximate a given set of points. A new point is determined as an interest point when deviation d_{n+1} of a new point (x_{n+1}, y_{n+1}) is larger than a threshold d_{thre}, otherwise not. Our method avoids complex off-line training. So long as an interest point is found, the trajectory is segmented, which indicates that a movement is detected (see Fig. 4).

3.3. Detection Mode Transfer

Except defined gestures, most of the arm movements of the speaker are meaningless to the system. If hand trajectories are far away from screen or not in the PARs, it is unnecessary to record the hand trajectory or recognize the gesture. The potential hand position can be predicted from the hand position in the previous frame. The detecting frequency can be reduced if the hand position is far from the PAR or its depth out of defined depth range. To adjust the detection frequency to arm movements, three hand detection modes are defined as "inactive", "semi-active" and "active", and each mode has different detection intervals k. In an inactive mode, hands will be re-detected after K frames, that is, $k = K - 1$. In the active mode, hands will be re-detected in the next frame, i.e., $k = 0$. If the detection mode keeps semi-active, the interval k is linearly increases from 1 to $(K - 1)$.

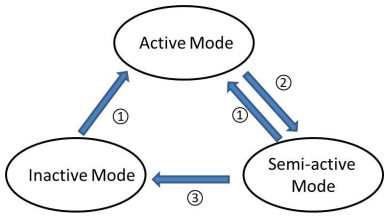

Figure 5. (1) Hand is detected in PARs. (2) Hand leaves PARs. (3) Hand leaves PARs for a long time.

When a hand is detected in PARs and is close enough to screen, the mode is immediately switched to "active" from "inactive" or "semi-active" mode. Otherwise, the mode is switched to "semi-active" mode. Since the number of frames being detected decreases in "semi-active mode". It will finally become "inactive mode" (see Fig. 5). It should be noted that "active" mode is not directly changed to "inactive" mode. Because the detecting module may generate a false reject error, i.e., missing a hand in one frame. In "Semi-active" mode, hand can be re-detected after less frames than in "inactive" mode. Notice that changing mode from "inactive" to "active" may require a relatively long time, during which hand may have already entered the PARs in those skipped frames. To avoid this phenomenon, an extension of PAR is generated for pre-detecting hand to ensure the completeness of trajectory. The hand trajectory does not include the hand detected in extension of PAR.

3.4. Feature extraction and classification

Before feature extraction, the size of PAR is normalized to facilitate the generation of the uniform MHI, which is used to represent decomposed movements. We adopt two stages classification pipeline to extract global feature, i.e., coding and pooling. The model is trained and tested with this feature by SVM.

Coding. The silhouettes of human arm belonging to the same movement in PAR are accumulated to generate the MHI, which is originally proposed by Bobick et al. [3]. In a MHI, pixel intensity records the temporal history of motion

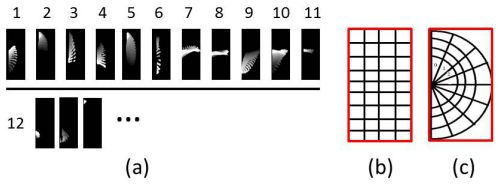

Figure 6. (a) Samples of 12 classes of movements. Class 12 includes meaningless movements. Three of them are listed. (b) Uniform rectangle spatial region. (c) Uniform semi-circle spatial region.

at each position. The MHI of decomposed movements are shown in Fig.6 (a).

Pooling. Designing a proper spatial region for pooling has a significant impact on feature's distinguishing capacity [6]. Since arms rotate along shoulder joints, the pixel intensity distributes in a semi-circle manner. Feature pooled from the uniform spatial region like that in Fig. 6 (b) is improper for further classification. For more effective representation, we choose the spatial regions shown in Fig. 6 (c). The blocks in the semi-circle region are represented as $R_1, R_2, ..., R_N$. The final global feature is denoted as $V = v_1, v_2, ..., v_N$ and $I_{m,n}$ is the intensity of pixels at the position (m, n) in the MHI. v_i is computed as follows,

$$v_i = \frac{1}{|R_i|} \sum_{(m,n)\in R_i} I_{m,n}, i = 1, 2, \ldots, N \tag{1}$$

where $|R_i|$ represents the pixel number in this block.

Classification. SVM is applied for classifying movements. The standard SVM divides two classes by clear gap that is as wide as possible. The classifier predicts new examples to a category based on which side of the gap they fall on. 12 classes of movements are designed and shown in Fig. 6 (a). The formal 11 classes are meaningful movements while class 12 is meaningless. A multi-class SVM model is trained when given the manually labeled training data. In practice, "one against one" strategy is applied for multi-class SVM and a distribution on all the labels for each movement is the output.

3.5. Set-based Soft Discriminative Model

As pointed out that the intra-class variances of our dataset is large since one gesture may have multiple rules to be assembled. In the discriminative model, the rules and their corresponding probabilities can be learned from training data. Movement sets, both meaningful and meaningless, are considered as rules in our work. Suppose x is the movements set (observation) and y is the gesture label (hidden state). Normal Discriminative classifiers model the posterior $p(y|x)$ directly. Suppose each movement in the set has a probability P_{w_mov} to be wrongly classified. Therefore, the gesture consisting of movements in such set has a probability of $P_{w\_ges} = 1 - (1 - P_{w\_mov})^{n_s}$ to be wrongly recognized, where n_s is the number of movements in one set. Through P_{w_mov} may be small, P_{w_ges}can be quite large so that it can hardly be tolerated in real-time applications.

In our work, P_{w_mov} is relatively small by the proper feature and classifier. Less promotion can be further made. For this reason, a soft discriminative model is proposed to directly decrease the P_{w_ges}. We define m as the movement with a distribution on all labels, which are denoted as M. p_m^M represents the probability of movement m classified to class M. Ms represents the trained movement set, also known as assemble rule, which serves as observation. G_j is the gesture label and serves as hidden state. The probability $p(G_j)$ is computed as follows:

$$\begin{cases} Ms_i =< M_{i1}, M_{i2}, \ldots, M_{in_i} > i = 1, 2, 3 \ldots \\ p\left(Ms_i\right) = \prod_{k=1}^{n_i} p_{m_k}^{M_{ik}} \\ p\left(A_j\right) = \max_i \sqrt[n_i]{\left[p\left(A_j|Ms_i\right) \times p\left(Ms_i\right)\right]} \end{cases} \tag{2}$$

where n_i-th root is used for normalizing since the number of the movements in one set is variant.

Actually, Eq. 2 replaces traditional x with a set Ms and fully use the distribution of classification output p_m^M. Each rule Ms_i is evaluated on subsets of a given movement sequence along time. The Ms can be regarded as a sparse joint distribution of the movement in sequence. It provides a soft observation of the discriminative model. The detailed implementation is described by Algorithm 1 and Fig. 7 gives a simple example of this model.

4. Experiment

Dataset and Correct Rate. We collect a new dataset of interactive presentation gestures that contains five classes: up, down, go, back and click intuitively corresponding to the up, down, left, right and enter in the keyboard. Part of the assemble rules of gestures is shown in Fig. 8, in which movement sets are represented in the form of MHI. The number of movements in one action and the appearance of the same class of MHI are variant, since different subjects act in a quiet different way.

In our dataset, each gesture was performed by three subjects for five times and at two different light conditions: projection light and normal light. Each subject performs three times at a normal speed (4 second/gesture) and the other two times at a fast speed (1 second/ gesture) under each light conditions. The distance between Kinect and subjects is 1.5-2.5m. The depth maps were captured at about 30 frames per second and the size is 640×480. In addition, the PARs are normalized to 120×260. Altogether, the dataset has $3(subjets) \times 5(times) \times 5(actions) \times 2(illuminations) =$

Train k rules $Ms = \{Ms_1, Ms_2 ... Ms_k\}$;
repeat
  Add a new movement $m_{n+1}$to a queue of movements;
  for *rule index* $i = 1$ *to* k **do**
    Compute the probability $p(Ms_i)$ for all the sets with equal length n_i in queue of movements;
    Multiply $p(Ms_i)$ by posterior $p(G_j|Ms_i)$;
    Normalize $p(G_j)$;
  end
  Choose the max $p(G_j)$;
  if $p(G_j) >$ *threshold* **then**
    Detect an action and label it as G_j;
    Delete the movements from queue;
  else
    No action is detected;
  end
until *no movement is detected any more*;

Algorithm 1: Gesture recognition from movement sets

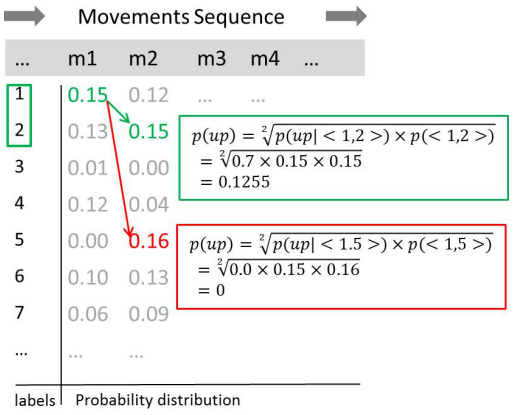

Figure 7. Suppose the priori $p(up| < 1,2 >) = 0.7$ and $p(up| < 1,5 >) = 0$. See from the table, this movement set has the largest joint probability on $< 1,5 >$ after classification, and has no chance to be labeled as "up". However, in soft discriminative model, this set still has the probability of 0.1255 to be labeled as "up" when $< 1,2 >$ is chosen.

150 gestures, 30 samples for each gesture. To test the classification of movement and the recognition of gesture, 60 gestures $(3(subjets) \times 2(times) \times 5(actions) \times 2(illuminations))$ serve as training data and the rest serve as test data. The dataset are labeled manually before training and test. The confusion matrix of 12 classes of movements is shown in Table 1. The confusion matrix of 5 classes of gestures is shown in Table 2. The correct classification rate of movement achieves 95.23% in 5-folds cross validation and the correct recognition rate of gesture achieves 90.00% in the test.

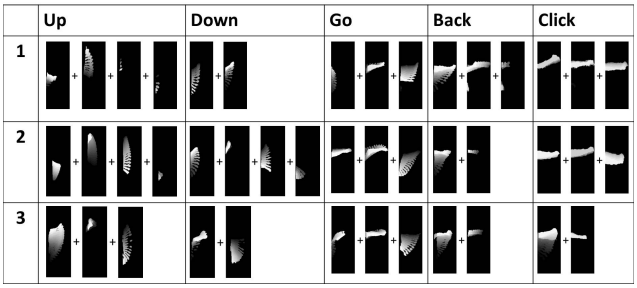

Figure 8. Samples of five gestures performed by 3 subjects.

Table 1. Confusion matrix of 12 classes of movements in the test. Our method achieves a correct recognition rate of 95.23% in 5-folds Cross Validation.

	1	2	3	4	5	6	7	8	9	10	11	12
1	.841		.02				.02			.02	.04	.06
2		.85			.07							.08
3			1									
4				.91		.04			.05			
5					1							
6						.93					.07	
7							.89	.05				.06
8								1				
9									1			
10										1		
11	.06		.06				.05				.83	
12				.01								.99

Table 2. Confusion matrix of 5 classes of gestures. None means no gesture is those frames. Achieve a correct rate of 90.00% excluding "None".

	Up	Down	Go	Back	Click	None
Up	16	2				
Down		17				1
Go		1	16			1
Back				15		3
Click					18	
None			2		1	

Complexity reduction. In our work, the gesture detection complexity adapts to the arm's movement, that is, the detection frequency is controlled by detection mode derived from PARs and hand position in the previous depth map. In the dataset, 18 videos are captured for testing. When collecting the dataset, one subject successively performs five different gestures in one video, and each video contains 600 frames. Among these depth frames, only partial of them are selected as active frame for gesture detection. As the results, the computing complexity is vastly reduced, and its reduction ratio is proportional to the number of inactive frame. Fig. 9 (Top) shows the active frame ratio over all frames of 18 videos. The active frame ratio is around 50%, which means that half of depth frames will not calculated for detection. When speaker spend more time on presentation instead of interaction, the active frame ratio will further reduced.

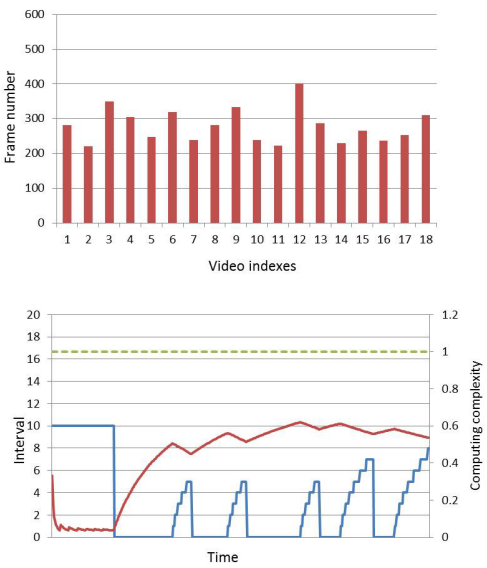

Figure 9. Top: active frames numbers of all the 18 videos. Bottom: interval (blue line) and computational complexity curve of one of our samples. Red line represents our computational complexity while green dash line represents normal computational complexity (Best view in color).

Table 3. Computational complexity in different detection mode. n is the number of pixels in PARs. Chamfer distance maps can be executed in linear (O(n)) time.

Detecting Mode	Processing	Frequency
Inactive mode	Detecting ($O(n)$)	Every 10 frames
Semi-active mode	Detecting ($O(n)$)	Every k frames ($1 < k < 10$)
Active mode	Detecting ($O(n)$) + Recording ($O(n)$) + Feature extraction($O(n)$)	Every frame

Table 3 gives the computational complexity in different detection mode. Fig. 9 (Bottom) shows the complexity from one of the 18 videos, which containing four gestures. At the beginning, hand is detected every 10 frames in an "inactive" mode. Once hand is detected in the PARs and is close enough to the screen, the detecting mode is switched to "active". If the hand leaves the PARs or become far away from screen, the number decreases gradually. See from Table 4, the active frame ratio reduces to 46.47% and the correct recognition rate does not decrease much. This method can be potentially used for wireless transfer of frames.

Comparison on Features. In the field of gesture recognition, the trajectories of gestures are always represented as a set of points (e.g., sampled positions of the head, hand, and eyes) in a 2-D space before being decomposed. For example, HoGS is a descriptor proposed by Sadeghipour et al. [11]. They combine this feature with SVM to solve the challenging problem of gesture-based object recognition. Though trajectory is obtained by tracking hand in the first place, MHI is our final feature. The reason lies in two folds. First, compared with sensitive point detection, the method to generate MHI is more robust since it simulates original silhouettes. Second, the MHI implicitly represents the direction of movement while a trajectory of hand points has less information of that information. To compare the two features, an experiment using trajectory as feature is conducted on our dataset.

Table 4. The total frame number and correct recognition rate. (AFP: Active Frame Percentage. CCR: Correct Recognition Rate.)

	AFP	CCR
All frames	100%	90.00%
Selected frames	46.47%	85.56%

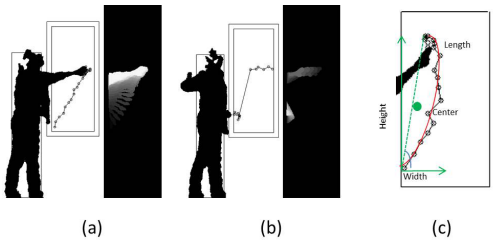

Figure 10. (a) Comparison on two features. (b) Trajectory has problem with outliers (clothes), the straight line is wrongly connected. (c) The five attributes for trajectory in our experiment.

As trajectories have been segmented by MLS, some attributes can be extracted from the curve of segmentations like the method in [11]. Five attributes are used: height, width, length, orientation and center of the curve (see Fig. 10 (c)). Combined with SVM, this feature has a low movement correct rate and gesture correct rate (see Table 5). As stated above, the main reason is the sensitiveness of points on trajectories. For example, pictures (a) and (b) in Fig. 4 are the comparisons between trajectory feature and MHI feature. In (a), the two features have almost the same discriminating power and are both correctly classified in experiment. In (b), since the clothes of the subject used to enter the bounding box and produce some outliers at the left bottom region, trajectory fails to describe this movement because of a wrongly connected straight line while the MHI feature is correctly classified. That mainly owns to the abundant original information the MHI feature contains.

Set-based Soft Discriminative Model. On the training data set including 60 samples, a posterior $p(G|Ms)$ is trained. In traditional discriminative model, Ms includes the movement sets that occur at least once in the training

Table 5. Comparison result. Movement correct classification rate are computed by 5-folds CV. (DM: Discriminative Model; MCCR: Movement Correct Classification Rate; GCRR: Gesture Correct Recognition Rate).

Feature	Model	MCCR	GCRR
MHI + Semi-circle	Traditional DM	95.23%	76.67%
MHI + Rectangle	Soft DM	79.31%	76.67%
Trajectory	Soft DM	48.32%	62.23%
MHI + Semi-circle	**Soft DM**	**95.23%**	**90.00%**

data. New movement sets have no chance to be changed to the nearest one in the Ms while soft discriminative model does. That is because soft discriminative model and choose the best one according to the distribution of movements m_i (see Eq. 2). The traditional model is also tested on the test set including the rest 90 samples and the correct recognition rate is 76.67%. This result shows that the correct recognition rate is limited by the scale of train set since traditional discriminative models directly model a conditional distribution over gesture labels given the observations. Besides, our observation is set-based, some observation sets on test set never occurs on training set. In our approach, this observation set is mapped to the one exists on the soft discriminative model trained by small scale train set. After using the soft discriminative model, the correct recognition rate is 90.00% (see Table 5).

5. Conclusions and Future Work

We propose a framework for gesture recognition and test it in our dataset. Experimental results show that our method is efficiency due to the detection mode transfer and robust due to the MHI feature and soft discriminative model. In addition, depth maps can further compressed in our framework, which will enhance the efficient. To recognize gesture from compressed images is our future work.

References

[1] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. In *ICCV*, volume 2, pages 1395–1402. IEEE, 2005. 2

[2] A. Bobick. Movement, activity and action: the role of knowledge in the perception of motion. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 352(1358):1257–1265, 1997. 3

[3] A. Bobick and J. Davis. The recognition of human movement using temporal templates. *PAMI*, 23(3):257–267, 2001. 4

[4] X. Feng and P. Perona. Human action recognition by sequence of movelet codewords. In *3D Data Processing Visualization and Transmission*, pages 717–721. IEEE, 2002. 3

[5] K. Fujimura and X. Liu. Sign recognition using depth image streams. In *FG*, pages 381–386. IEEE, 2006. 2

[6] Y. Jia, C. Huang, and T. Darrell. Beyond spatial pyramids: Receptive field learning for pooled image features. In *CVPR*, pages 3370–3377. IEEE, 2012. 5

[7] A. Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. *Advances in neural information processing systems*, 14:841, 2002. 3

[8] A. Klaser, M. Marszalek, and C. Schmid. A Spatio-Temporal Descriptor Based on 3D-Gradients. In M. Everingham, C. Needham, and R. Fraile, editors, *BMVC*, pages 275:1–10, Leeds, Royaume-Uni, 2008. British Machine Vision Association. 3

[9] I. Laptev. On space-time interest points. *IJCV*, 64(2):107–123, 2005. 3

[10] R. Poppe. A survey on vision-based human action recognition. *IVC*, 28(6):976–990, 2010. 1

[11] A. Sadeghipour, L. Morency, and S. Kopp. Gesture-based object recognition using histograms of guiding strokes. *BMVC*, 2012. 2, 3, 7

[12] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A local svm approach. In *ICPR*, volume 3, pages 32–36. IEEE, 2004. 2

[13] P. Scovanner, S. Ali, and M. Shah. A 3-dimensional sift descriptor and its application to action recognition. In *Proceedings of the 15th international conference on Multimedia*, pages 357–360. ACM, 2007. 3

[14] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from single depth images. In *CVPR*, pages 1297–1304. IEEE, 2011. 2

[15] Y. Song, D. Demirdjian, and R. Davis. Tracking body and hands for gesture recognition: Natops aircraft handling signals database. In *FG*, pages 500–506. IEEE, 2011. 2, 3

[16] P. Turaga, R. Chellappa, V. Subrahmanian, and O. Udrea. Machine recognition of human activities: A survey. *Circuits and Systems for Video Technology, IEEE Transactions on*, 18(11):1473–1488, 2008. 1

[17] J. Wang, Z. Liu, Y. Wu, and J. Yuan. Mining actionlet ensemble for action recognition with depth cameras. In *CVPR*, pages 1290–1297. IEEE, 2012. 2

[18] Q. Wang, X. Chen, and W. Gao. Skin color weighted disparity competition for hand segmentation from stereo camera. In *BMVC*, pages 66–1, 2010. 2

[19] D. Weinland, E. Boyer, and R. Ronfard. Action recognition from arbitrary views using 3d exemplars. In *ICCV*, pages 1–7. IEEE, 2007. 3

[20] G. Willems, T. Tuytelaars, and L. Van Gool. An efficient dense and scale-invariant spatio-temporal interest point detector. *ECCV*, pages 650–663, 2008. 3