Three Dimensional Motion Trail Model for Gesture Recognition

Bin Liang
Charles Sturt University
Wagga Wagga NSW, Australia
bliang@csu.edu.au

Lihong Zheng
Charles Sturt University
Wagga Wagga NSW, Australia
lzheng@csu.edu.au

Abstract

In this paper an effective method is presented to recognize human gestures from sequences of depth images. Specifically, we propose a three dimensional motion trail model (3D-MTM) to explicitly represent the dynamics and statics of gestures in 3D space. In 2D space, the motion trail model (2D-MTM) consists of both motion information and static posture information over the gesture sequence along the xoy-plane. Considering gestures are performed in 3D space, depth images are projected onto two other planes to encode additional gesture information. The 2D-MTM is then extensively combined with complementary motion information from additional two planes to generate the 3D-MTM. Furthermore, the Histogram of Oriented Gradient (HOG) feature vector is extracted from the proposed 3D-MTM as the representation of a gesture sequence. The experiment results show that the proposed method achieves better results on two publicly available datasets namely MSR Action3D dataset and ChaLearn gesture dataset.

1. Introduction

Recognition of human gestures has always remained an active research topic of great interest over the last three decades. There are many promising gesture recognition applications in computer vision, such as automatic environment surveillance, assisted living, video indexing, sport video analysis and human computer interaction. Sensors used for gesture recognition include wearable sensors and vision sensors. Compared with the extensive calibration and restricted natural movement of the wearable sensors, vision sensors address these issues. In recent years, vision sensors such as video cameras are widely used for gesture recognition and research on vision-based interaction has been actively studied. The task of gesture recognition has made significant advances using video cameras. Despite the research efforts in the past decade and many encouraging successes, accurate gesture recognition is still a challenging task. The advances have been limited to the use of RGB images captured by video cameras, ignoring the important information of depth. Depth information has long been regarded as an essential part of successful gesture recognition [15]. In addition, how to model the 3D human gestures that are dynamic, static and ambiguous in an efficient way is another major issue.

The Kinect camera provides depth information through collecting a sequence of depth images for human gestures. Depth images collect depth information as 8-bit gray level for each pixel that presents the distance between the captured object and the camera. Figure 1 shows RGB images and corresponding depth images for one gesture sequence from ChaLearn gesture dataset [14]. As seen, the motion ambiguity of the video camera, such as the huge color and texture variability induced by clothing, hair, skin and background, could be bypassed. This paper studies the recognition of human gestures from sequences of depth images.

(a) RGB images

(b) Depth images

Figure 1. RGB images and depth images for one sample gesture sequence from ChaLearn gesture dataset [14]

Depth images contain a great amount of data which could result in high computational costs. An effective human gesture recognition method using a depth camera is presented in this paper. The general framework of our method is illustrated in Figure 2. Based on depth data, gesture regions can be easily segmented from background with a suitable threshold of depth value. After that, smoothing is used in order to make sure there is less noise in depth images so that gesture analysis could not be negatively influenced by the noise. We then use the proposed model (3D-MTM) to represent gesture motion information and static posture

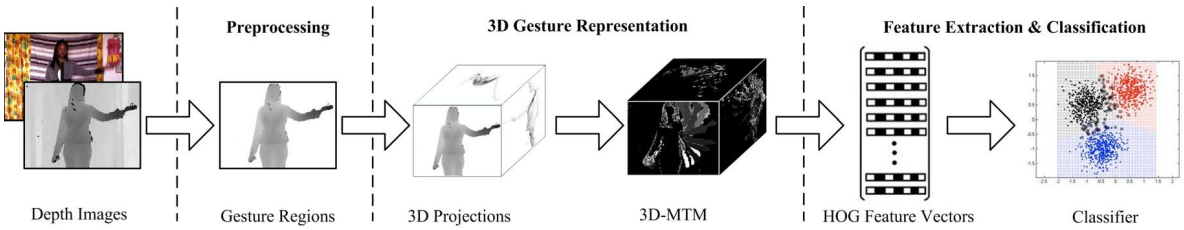


Figure 2. The general framework of the proposed method

information in 3D space. The next step is feature extraction and classification. The Histogram of Oriented Gradients (HOG) [11] descriptors are exploited to describe the local distribution of gradients for feature vector extraction. A classifier compares the features of training samples and test samples to identify the gesture types.

More importantly, we propose a new 3D model for human gestures, called the *Three Dimensional Motion Trail Model (3D-MTM)*. Gestures can be divided into two distinctive categories: dynamic and static [5]. A dynamic gesture is intended to move over a period of time while a static gesture is observed at the gesture in which a single posture is held for a certain duration. To understand the meaning of the gesture, it is necessary to interpret all the static and dynamic gestures over a period of time. Thus, gesture recognition is to interpret a continuous gesture sequence. The proposed motion trail model in 2D space (2D-MTM) explicitly models the temporal motion information and static posture information of human gestures. We use the depth motion history image (D-MHI) and the average motion image (AMI) to encode motion information of gestures. Similarly, the static posture history image (SHI) and the average static posture image (ASI) of gestures are generated containing static posture information. The 2D-MTM consists of both motion information and static posture information over the gesture sequence along the xoy-plane. Furthermore, human gestures captured by video cameras can only encode the information induced by the lateral movement of the scene parallel to the image plane. As human bodies and motions are performed in 3D space, the information loss in the depth channel could cause significant degradation of the representation and discriminating capability for human gestures. For better modeling human gestures in 3D space, our idea is to obtain the complementary motion information from other views in 3D space to compensate for the lost gesture information in 2D space. After projecting depth images onto other two planes in 3D space, the 2D-MTM is then extensively combined with complementary motion information from additional two planes to generate the 3D-MTM. The 3D-MTM of a gesture contains disparity information from the corresponding depth images.

Our key contributions are the following: 1) a 2D motion trail model (2D-MTM) is proposed to represent the motion information and static posture information of a gesture sequence; 2) a novel 3D model (3D-MTM) is extensively proposed by projecting depth images onto other two planes, and it is shown to be robust to model gestures in 3D space; 3) the proposed method based on 3D-MTM achieves competitive performance on two benchmark datasets: MSR Action3D dataset [18] and one-shot learning ChaLearn gesture dataset [14]. The experiments have shown that not only does our method achieve better performance on the benchmark dataset that has multiple training data, but also it generalizes well on the one-shot learning gesture dataset.

2. Related work

Gesture recognition approaches can be categorized into three main categories: template based approaches, volumetric approaches and machine learning based approaches. Template based approaches usually convert a gesture sequence into a static shape pattern (e.g., MHI [6]), and then the extracted features are used to compare to the prestored prototypes during recognition. Template matching approaches are easy to implement and require less computational load. Volumetric approaches consider the whole gesture sequence as a 3D volume of pixel intensities instead of extracting features on a frame basis. These approaches generally extend typical image features to the 3D case. Interest point based methods [16] and geometric methods [20] are two commonly used methods. These kinds of approaches have disadvantages of high computational cost and complexity. Machine learning techniques are employed for gesture recognition in recent years. k-NN based methods [10] are simple to implement, but all the training data will be used when recognizing gestures, which is memory and time expensive. SVM based methods [4] are the most popular application for gesture recognition as a group of associated supervised learning techniques. Bag-of-features methods [27] represent the gesture sequence as unsorted sets of features, and then the features are quantized into discrete vocabularies by learning. The HMM based methods [8] have been frequently used for modeling human motions as they efficiently abstract time series data, and can be used for subsequent motion recognition. These methods usually require complex interactive computation.

There have been many surveys on human gesture recog-

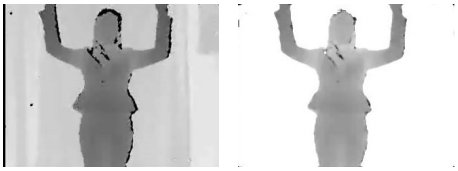
nition and analysis [1, 2, 22, 13], most of which have cited the motion history image (MHI) method [6] as one of the most important methods. In the MHI, the silhouette sequence is condensed into one gray scale image, while dominant motion information is preserved. Therefore, it can represent a gesture sequence in a compact manner. Besides, the MHI is not so sensitive to silhouette noises, like holes, shadows, and missing parts [3]. These advantages make MHI a suitable candidate for motion and gait analysis [19]. Tian *et al.* [26] use Harris detector and local HOG descriptor on MHI to perform gesture recognition and detection. The MHI expresses the motion history by the intensity of every pixel in a temporal manner. However, the traditional MHI method has the limitation of scalability because only lateral motion of the gesture is analyzed. Human gestures are performed in 3D space, which means MHI performed in 2D space may miss some motion information of the gesture performed in the real world. We propose a novel model to represent distinctive features of human gestures, and apply a classifier to compare the features extracted from the proposed model. The advantages of the proposed model are: 1) the proposed model is based on depth images, so gesture information is condensed into one model using depth data, while MHI uses silhouette images which could generate ambiguous gestures representation; 2) the 3D-MTM contains the information of motion history, static posture history, average motion and average static posture through the entire gesture sequence, while MHI only records the most recent movements; 3) our model represents movements from xoy-plane, xoz-plane and yoz-plane, *i.e.* three orthogonal projections of depth images, while MHI only keeps motion information on a single plane.

3. Proposed method

To represent the gesture motion information and static posture information in 3D space, a novel 3D model (3D-MTM) is proposed. We project depth images onto three planes and obtain D-MHI, SHI, AMI and ASI on xoy-plane, and D-MHI on the other two planes. HOG descriptors are extracted from 3D-MTM and concatenated as the final gesture representation. This section gives a detailed description of the proposed gesture recognition method based on the 3D-MTM, including preprocessing, 3D gesture representation and feature extraction and classification.

3.1. Preprocessing

Preprocessing consists of two steps: background removal and image smoothing. In order to represent human gestures well, to segment the gesture regions out of each video frame is an essential step for gesture recognition. In the case of having no prior background image and dynamic background, it is a challenging task for traditional methods. To this end, we adopt depth images to segment gesture regions. In depth images, the values of pixels belonging to background have a great difference from those belonging to the object. Utilizing the property, the gesture regions in a sequence can be easily segmented from the background by using Otsu's method [23] to classify the pixels. Besides, the depth images have other drawbacks, one of which is the noise at the edge of objects. With missing bits and a pretty serious flickering issue, noise in depth images resembles a type of salt and pepper noise. Motion information is sensitive to silhouette noise, so smoothing of depth images is necessary. We then adopt a median filter [24] for spatial filtering to replace the pixel value with the median value of the sub-image. Thus, it removes randomly generated noise and smooths the original image. After applying noise reduction to the depth image, the resulting motion description is less prone to faulty defects from the depth sensor. The preprocessing operation is highly effective for the following process in our method. Figure 3 shows the original depth image and the image after preprocessing operation from one gesture sequence. As seen, the gesture regions are clearly segmented and noise has been removed.



(a) Original depth image (b) Preprocessed image

Figure 3. Preprocessing of original depth image

3.2. 3D Gesture representation

3.2.1 Motion history image

Motion history image (MHI) and motion energy image (MEI) templates were proposed by Bobick and Davis [6] to describe where there is motion and how the object is moving. All the frames in one gesture sequence are projected onto a single image over the range of time. The MHI template has the advantage that the temporal motion information may be encoded into one image, and the MHI spans the time scale of human gestures [7]. The MHI $\mathrm{H}(x, y, t; \tau)$ can be obtained from an updating function $\Psi(x, y, t)$:

$$\mathrm{H}(x, y, t; \tau) = \begin{cases} \tau & \text{if } \Psi(x, y, t) = 1, \\ \max(0, \mathrm{H}(x, y, t-1; \tau) - \delta) & \text{otherwise.} \end{cases} \quad (1)$$

where x, y represent pixel position and t is time. $\Psi(x, y, t)$ signals object presence in the current gesture image with coordinate (x, y) at the t^{th} frame of the gesture sequence. τ decides the temporal duration of the motion, and δ is the decay parameter. The update function $\Psi(x, y, t)$ is called for every frame analyzed in the gesture sequence. Finally,

a scalar-valued image can be generated through the computation, presenting more recent moving pixels brighter and vice-versa [6, 21]. A final MHI template $\mathrm{H}(x, y, t; \tau)$ records the temporal history of motion in the corresponding gesture sequence.

3.2.2 2D motion trail model

In the MHI, the gesture sequence is condensed into a single gray scale image, preserving dominant motion history information. It keeps a record of temporal changes at each pixel location, which decays over time [28]. MHI presents the motion history using binary cumulative motion images. Binary frame-to-frame difference methods are widely used for motion representation. Although binary images or silhouette based images are able to represent a wide variety of body configurations, they could produce ambiguities in the represented motion. However, in the presence of occlusions of body, or improper implementation of the update function, the MHI fails to cover most of the motion regions. For instance, most of the arm and hand movements performed in front of a person's body become "invisible". In addition, the information of static posture history, repetitive movements and repetitive static postures is ignored in the MHI template. To overcome the limitations, the proposed 2D model in this paper employs four templates generated from depth images, *i.e.* depth motion history image (D-MHI), average motion image (AMI), static posture history image (SHI) and average static posture image (ASI), to encode supplementary essential information of gestures to increase the robustness for representation.

Assume $I_t = (I_1, I_2, \ldots, I_T)$ is a depth image sequence and let $D_t = (D_1, D_2, D_3, \ldots, D_T)$ be a difference image sequence indicating the absolute difference between consecutive frames:

$$D_t = \begin{cases} I_1 & \text{if } t = 1 \\ |I_t - I_{t-1}| & \text{otherwise} \end{cases} \tag{2}$$

where t is the time and T is the total number of frames in one gesture sequence. To indicate the regions of motion and static posture, the motion information and static posture information of I_t can be obtained through motion update function $\Psi_{\mathrm{M}}(x, y, t)$ and static posture update function $\Psi_{\mathrm{S}}(x, y, t)$:

$$\Psi_{\mathrm{M}}(x, y, t) = \begin{cases} 1 & \text{if } \ D_t > \varsigma_M, \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

$$\Psi_{\mathrm{S}}(x, y, t) = \begin{cases} 1 & \text{if } \ I_t - D_t > \varsigma_S, \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

where ς_M and ς_S are the thresholds for motion and static information between consecutive frames. The motion information image and static information image from two frames of one sample gesture are illustrated in Figure 4. It can be seen that the dynamic regions and static regions are highlighted using motion update function and static posture update function.

(a) I_{t-1} (b) I_t

(c) $\Psi_{\mathrm{M}}(t)$ (d) $\Psi_{\mathrm{S}}(t)$

Figure 4. Motion information and static information

It is obvious that in Equation(1) the larger τ is, the more information of the gesture could be encoded. In our work, τ is set as the total number of frames T for the whole gesture to preserve the whole motion trail information. We define depth motion history image (D-MHI) as:

$$\mathrm{H}_{\mathrm{M}}(x, y, t) = \begin{cases} T & \text{if } \Psi_{\mathrm{M}}(x, y, t) = 1 \\ \mathrm{H}_{\mathrm{M}}(x, y, t-1) - 1 & \text{otherwise} \end{cases} \tag{5}$$

Additionally, we extend D-MHI and utilize the static posture update function $\Psi_{\mathrm{S}}(x, y, t)$ to get static posture history image (SHI) indicating to compensate for static regions over the whole gesture sequence, which can be obtained in the similar way as D-MHI:

$$\mathrm{H}_{\mathrm{S}}(x, y, t) = \begin{cases} T & \text{if } \Psi_{\mathrm{S}}(x, y, t) = 1 \\ \mathrm{H}_{\mathrm{S}}(x, y, t-1) - 1 & \text{otherwise} \end{cases} \tag{6}$$

Note that there are no maximum operators in Equation(5) and Equation(6) because parameter τ is set as the whole duration T causing non-negative values of $\mathrm{H}_{\mathrm{M}}(x, y, t)$ and $\mathrm{H}_{\mathrm{S}}(x, y, t)$. Figure 5 shows the D-MHI and SHI generated from one sample gesture.

(a) D-MHI (b) SHI

Figure 5. D-MHI and SHI of one sample gesture

In the proposed motion trail model, another two elements, average motion image (AMI) and average static posture image (ASI), are employed to provide complementary

information of the gesture for repetitive movements and repetitive static postures over the whole gesture sequence.

Average motion image (AMI) is to compensate for the repetitive movements information. The summation of all motion information using $\Psi_{\mathrm{M}}(x, y, t)$ and normalization of the pixel value defines the AMI:

$$\mathrm{A_M} = \frac{1}{T}\sum_{t=1}^{T}\Psi_{\mathrm{M}}(x, y, t) \tag{7}$$

Average static posture image (ASI) is used to recover the average static posture information, which can be defined using static posture information of each frame:

$$\mathrm{A_S} = \frac{1}{T}\sum_{t=1}^{T}\Psi_{\mathrm{S}}(x, y, t) \tag{8}$$

Figure 6 shows the AMI and ASI of one sample gesture.

(a) AMI (b) ASI

Figure 6. AMI and ASI of one sample gesture

3.2.3 3D motion trail model

Our previous research on 2D-MTM based gesture representation has shown that the proposed model carries more essential gesture information in 2D space than traditional MHI method. However, the proposed 2D-MTM has some limitations. It can only encode the information induced by the lateral movement of the scene motion parallel to the image plane. As human bodies and motions are performed in 3D space, the information loss in the depth channel could cause significant degradation of the representation and discriminating capability for human gestures. With depth images, we can now extend the proposed 2D-MTM to 3D space, generating a 3D-MTM which is capable of encoding the motion information along two additional planes (xoz-plane and yoz-plane) besides xoy-plane. Thus the 3D-MTM uses disparity information of the gesture from xoy-plane, xoz-plane and yoz-plane, which can robustly discriminate each gesture using information from additional viewpoints with only one model.

The information from xoy-plane is dominant for the gesture and the projections onto xoz-plane and yoz-plane can be very coarse due to the resolution of the depth images, so only D-MHI templates are generated from the projections on xoz-plane and yoz-plane respectively. The 3D-MTM of one gesture sequence is demonstrated in Figure 7. Therefore, one gesture depth image sequence can be modeled as six templates using the proposed 3D-MTM.

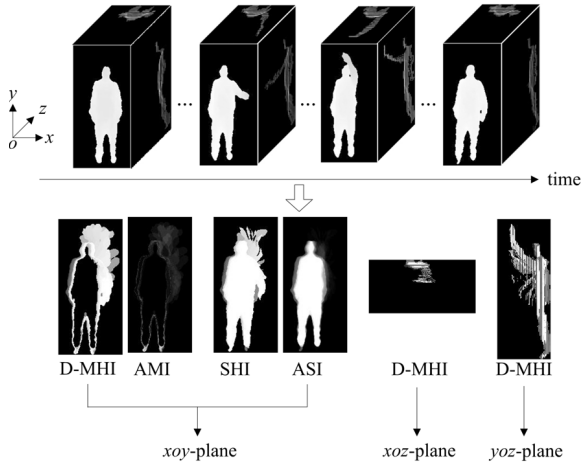


Figure 7. The 3D-MTM generated from one sample gesture

3.3. Feature extraction and classification

HOG [11] is able to characterize the local appearance and shape on 3D-MTM by the distribution of local intensity gradients. Thus, each template can be represented as a feature vector of $N \times N \times B$ dimension, and then we concatenate six vectors into one feature vector. In our experiment, the size $N \times N$ of lattices is 3×3, and the bin number B is 9. In this way, the 3D-MTM generates a descriptor vector with the dimension of $6 \times 3 \times 3 \times 9 = 486$.

As for MSR Action3D dataset [18], support vector machine (SVM) is adopted for the final stage to classify the gestures. A well-known SVM library LIBSVM [9] is used to train our model and test the performance of the model. Then we have used RBF kernel which non-linearly maps samples into a higher dimensional space so it can handle the case when the relation between class labels and attributes is non-linear.

One-shot learning ChaLearn gesture dataset [14] only has one training sample for each gesture class, so we employ a nonparametric method *Maximum Correlation Coefficient* as the matching metric by avoiding the issue of overfitting.

4. Experimental results

We choose MSR Acion3D dataset [18] and one-shot learning ChaLearn gesture dataset [14] to evaluate the proposed gesture recognition method. The experimental results show that the proposed method outperforms than other methods.

4.1. MSR Action3D dataset

MSR Action3D dataset [18] is an action dataset of depth image sequences captured by a depth sensor similar to the Kinect device. This dataset contains 20 action types: *high arm wave*, *horizontal arm wave*, *hammer*, *hand catch*, *forward punch*, *high throw*, *draw x*, *draw tick*, *draw circle*, *hand clap*, *two hand wave*, *side-boxing*, *bend*, *forward kick*, *side kick*, *jogging*, *tennis swing*, *tennis serve*, *golf swing*, *pick up & throw*. Each action is performed by 10 subjects for 2 or 3 times. There are 567 gesture sequences in total. The resolution is 320×240. Some frames of the gesture sequences are shown in Figure 8. Those gestures cover various movements of arms, legs, torso and their combinations. The background of the gesture is clean in this dataset, but this dataset is still challenging because many of the gestures are highly similar to each other.

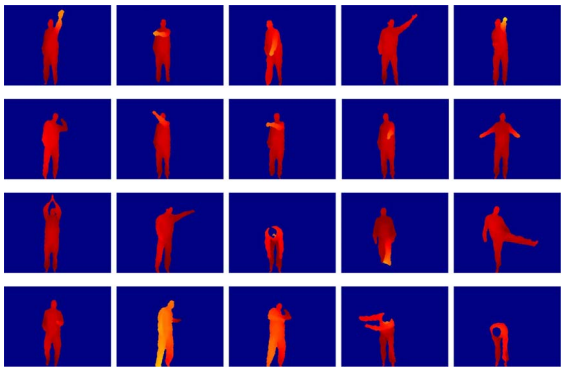


Figure 8. Sample frames from the MSR Action3D dataset

In order to evaluate the performance of the proposed method, our experiments are conducted using different number of training samples. We follow the same experimental settings as [18] to divide the 20 actions into three subsets, each having 8 actions as listed in Table 1. For each subset, there are three different tests, *i.e.* Test One (T1), Test Two (T2), and Cross Subject Test (CST). In Test One, 1/3 of the samples are used as training and the rest as testing; in Test Two, 2/3 samples are used as training and the rest as testing; in the Cross Subject Test, half of the subjects are used as training and the rest subjects are used as testing.

We compare our proposed method with the state-of-the-art method [18] on the MSR Action3D dataset in Table 2. As shown in this table, the proposed method considerably outperforms the Bag-of-3D-Points. The average recognition accuracies of our method in Test One, Test Two, and Cross Subject Test are 96.1%, 99.2% and 78.9%, which increase the average accuracies in [18] by 4.5%, 5.0%, and 4.2%, respectively. The results reflect the robustness of the proposed 3D-MTM, and demonstrate the 3D-MTM can represent distinctive features of human gestures.

Action Set 1 (AS1)	Action Set 2 (AS2)	Action Set 3 (AS3)
Horizontal arm wave	High arm wave	High throw
Hammer	Hand catch	Forward kick
Forward punch	Draw x	Side kick
High throw	Draw tick	Jogging
Hand clap	Draw circle	Tennis swing
Bend	Two hand wave	Tennis serve
Tennis serve	Forward kick	Golf swing
Pickup & throw	Side boxing	Pickup & throw

Table 1. The three action subsets

	Bag-of-3D-Points [18]	Proposed Method
T1-AS1	89.5%	**96.0%**
T1-AS2	89.0%	**94.9%**
T1-AS3	96.3%	**97.3%**
T2-AS1	93.4%	**100.0%**
T2-AS2	92.9%	**97.5%**
T2-AS3	96.3%	**100.0%**
CST-AS1	72.9%	**73.7%**
CST-AS2	71.9%	**81.5%**
CST-AS3	79.2%	**81.6%**

Table 2. Recognition accuracy comparison for MSR Action3D dataset of different subsets that are Test One (T1), Test Two (T2), and Cross Subject Test (CST)

4.2. One-shot learning ChaLearn gesture dataset

The data in ChaLearn gesture dataset [14] are recorded using a Kinect camera including both hand and arm gestures. It was used for a one-shot learning challenge of gesture recognition. The key aspect of the dataset is that each gesture class has only one training sample. Our experiments are performed on the first 10 data batches of the dataset, each of which is made of 47 gesture sequences and split into a training set and a test set. In the dataset, each test video contains 1 to 5 gestures. Note that in order to verify the performance of our approach, we use the temporal segmentation annotation provided by [14]. Thus, a test video is firstly segmented so that we can get several single gesture sequences from a test video. Some sample frames from this dataset are shown in Figure 9

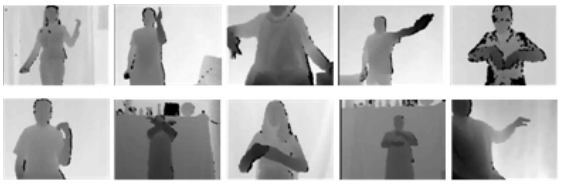


Figure 9. Sample frames from the ChaLearn gesture dataset

Method	Average error rate
Baseline [14]	62.8%
Dynamic Time Warping [14]	43.1%
Principle Motion [12]	37.4%
MHI [6]	37.6%
2D-MTM (ours)	**24.4%**
3D-MTM (ours)	**21.7%**

Table 3. Performance comparison on one-shot learning ChaLearn gesture dataset

The recognition performance is evaluated using the Levenshtein distance [17]. Table 3 compares the average recognition error rate of the proposed method with results from other methods: baseline method [14], dynamic time warping (DTW) [14], principle motion method [12], MHI [6] method and the proposed 2D-MTM. It can be observed that the proposed 3D-MTM shows a better performance which is competitive to the other methods. Our approach achieves 21.7% average error rate, which illustrates that the 3D-MTM can be effectively adopted for gesture recognition.

Figure 10 shows the recognition error rates of each data batch using baseline, DTW, principle motion, MHI, 2D-MTM and 3D-MTM, respectively. It can be found that the proposed 3D-MTM performs better than other methods on average. Note that for data batch 7 and 9, the 3D-MTM does not perform better than 2D-MTM. The reason is that in these two data batches the motion information from xoz-plane and yoz-plane induces negative effect instead of providing complementary information to the dominant gesture information on xoy-plane. The human movements in these two data batches are mainly performed on the xoy-plane and less discriminative movements on xoz-plane and yoz-plane. Besides, there are other particular batches, data batch 3 and 10, where the main movements are hand movements and finger movements. Theses subtle movements dominate the whole gesture causing it to be confused with other similar gestures. We reason the higher error rates in these batches that there is no hand detector in our approach to locate the hand position. Thus it can be concluded that our method performs well when there is large amount of motion presenting in a gesture.

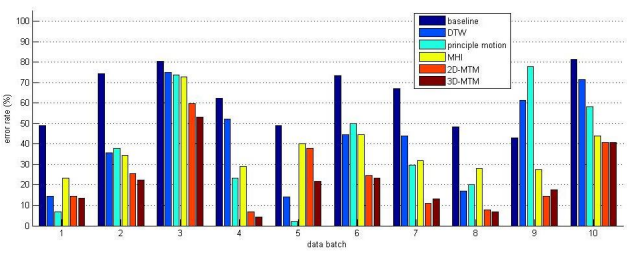


Figure 10. Results of the 10 data batches

5. Conclusions

In this paper, we have proposed an effective gesture recognition method by using a novel model *3D-MTM*. The 3D-MTM is able to represent the motion information and static posture information of human gestures along xoy-plane, and additional motion information from xoz-plane and yoz-plane. The experimental results on MSR 3DAction dataset demonstrate that our method outperforms the state-of-the-art method. In addition, our method also performs well on the one-shot learning ChaLearn gesture dataset. The future work will focus on gesture recognition on variant subjects to improve recognition performance in the Cross Subject Test.

References

[1] J. K. Aggarwal and S. Park. Human motion: Modeling and recognition of actions and interactions. In *3D Data Processing, Visualization and Transmission, 2004. 3DPVT 2004. Proceedings. 2nd International Symposium on*, pages 640–647. IEEE, 2004.

[2] M. Ahad, J. Tan, H. Kim, and S. Ishikawa. Human activity recognition: various paradigms. In *Control, Automation and Systems, 2008. ICCAS 2008. International Conference on*, pages 1896–1901. IEEE, 2008.

[3] M. A. R. Ahad, J. K. Tan, H. Kim, and S. Ishikawa. Motion history image: its variants and applications. *Machine Vision and Applications*, 23(2):255–281, 2012.

[4] E. Ardizzone, A. Chella, and R. Pirrone. Pose classification using support vector machines. In *Neural Networks, 2000. IJCNN 2000, Proceedings of the IEEE-INNS-ENNS International Joint Conference on*, volume 6, pages 317–322. IEEE, 2000.

[5] H. Birk and T. B. Moeslund. *Recognizing gestures from the hand alphabet using principal component analysis*. Aalborg Universitet, 1996.

[6] A. F. Bobick and J. W. Davis. The recognition of human movement using temporal templates. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(3):257–267, 2001.

[7] G. R. Bradski and J. W. Davis. Motion segmentation and pose recognition with motion history gradients. *Machine Vision and Applications*, 13(3):174–184, 2002.

[8] M. Brand, N. Oliver, and A. Pentland. Coupled hidden markov models for complex action recognition. In *Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on*, pages 994–999. IEEE, 1997.

[9] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[10] T. Cover and P. Hart. Nearest neighbor pattern classification. *Information Theory, IEEE Transactions on*, 13(1):21–27, 1967.

[11] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005.

[12] H. J. Escalante and I. Guyon. Principal motion: Pca-based reconstruction of motion histograms. Technical report, Technical report, ChaLearn Technical Memorandum, June 2012. http://www. causality. inf. ethz. ch/Gesture/principal_motion. pdf, 2012.

[13] D. M. Gavrila. The visual analysis of human movement: A survey. *Computer vision and image understanding*, 73(1):82–98, 1999.

[14] I. Guyon, V. Athitsos, P. Jangyodsuk, H. J. Escalante, and B. Hamner. Results and analysis of the chalearn gesture challenge 2012. 2013.

[15] A. Janoch, S. Karayev, Y. Jia, J. T. Barron, M. Fritz, K. Saenko, and T. Darrell. A category-level 3d object dataset: Putting the kinect to work. In *Consumer Depth Cameras for Computer Vision*, pages 141–165. Springer, 2013.

[16] I. Laptev. On space-time interest points. *International Journal of Computer Vision*, 64(2-3):107–123, 2005.

[17] V. I. Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. In *Soviet physics doklady*, volume 10, page 707, 1966.

[18] W. Li, Z. Zhang, and Z. Liu. Action recognition based on a bag of 3d points. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pages 9–14. IEEE, 2010.

[19] J. Liu and N. Zheng. Gait history image: a novel temporal template for gait recognition. In *Multimedia and Expo, 2007 IEEE International Conference on*, pages 663–666. IEEE, 2007.

[20] Y. M. Lui, J. R. Beveridge, and M. Kirby. Action classification on product manifolds. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 833–839. IEEE, 2010.

[21] O. Masoud and N. Papanikolopoulos. A method for human action recognition. *Image and Vision Computing*, 21(8):729–743, 2003.

[22] T. B. Moeslund, A. Hilton, and V. Krüger. A survey of advances in vision-based human motion capture and analysis. *Computer vision and image understanding*, 104(2):90–126, 2006.

[23] N. Otsu. A threshold selection method from gray-level histograms. *Automatica*, 11(285-296):23–27, 1975.

[24] S. Park, S. Yu, J. Kim, S. Kim, and S. Lee. 3d hand tracking using kalman filter in depth space. *EURASIP Journal on Advances in Signal Processing*, 2012(1):1–18, 2012.

[25] R. Poppe. Vision-based human motion analysis: An overview. *Computer vision and image understanding*, 108(1):4–18, 2007.

[26] Y. Tian, L. Cao, Z. Liu, and Z. Zhang. Hierarchical filtered motion for action recognition in crowded videos. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 42(3):313–323, 2012.

[27] J. Willamowski, D. Arregui, G. Csurka, C. R. Dance, and L. Fan. Categorizing nine visual classes using local appearance descriptors. *illumination*, 17:21, 2004.

[28] T. Xiang and S. Gong. Beyond tracking: Modelling activity and understanding behaviour. *International Journal of Computer Vision*, 67(1):21–51, 2006.