Depth Camera Based Hand Gesture Recognition and its Applications in Human-Computer-Interaction

Zhou Ren, Jingjing Meng, Junsong Yuan
School of EEE, Nanyang Technological University
50 Nanyang Avenue, Singapore 639798
Email: {renzhou, jingjing.meng, jsyuan}@ntu.edu.sg

***Abstract*—Of various Human-Computer-Interactions (HCI), hand gesture based HCI might be the most natural and intuitive way to communicate between people and machines, since it closely mimics how human interact with each other. Its intuitiveness and naturalness have spawned many applications in exploring large and complex data, computer games, virtual reality, health care, etc. Although the market for hand gesture based HCI is huge, building a robust hand gesture recognition system remains a challenging problem for traditional vision-based approaches, which are greatly limited by the quality of the input from optical sensors. [16] proposed a novel dissimilarity distance metric for hand gesture recognition using Kinect sensor, called Finger-Earth Mover's Distance (FEMD).**

In this paper, we compare the performance in terms of speed and accuracy between FEMD and traditional corresponding-based shape matching algorithm, Shape Context. And then we introduce several HCI applications built on top of a accurate and robust hand gesture recognition system based on FEMD. This hand gesture recognition system performs robustly despite variations in hand orientation, scale or articulation. Moreover, it works well in uncontrolled environments with background clusters. We demonstrate that this robust hand gesture recognition system can be a key enabler for numerous hand gesture based HCI systems.

I. Introduction

There has been a great emphasis lately on Human-Computer-Interaction (HCI) research to create easy-to-use interfaces by directly employing natural communication and manipulation skills of humans. Among different human body parts, the hand is the most effective general-purpose interaction tool, due to its dexterity. Adopting hand gesture as an interface in HCI will not only allow the deployment of a wide range of applications in sophisticated computing environments such as virtual reality systems and interactive gaming platforms, but also benefit our daily life such as providing aids for the hearing impaired, and maintaining absolute sterility in health care environment using touchless interfaces via gestures [21].

Currently, the most effective tools for capturing hand gesture are electro-mechanical or magnetic sensing devices (data gloves) [9]. These methods employ sensors attached to a glove that transduces finger flexions into electrical signals to determine the hand gesture. They deliver the most complete, application-independent set of real-time measurements of the hand in HCI. However, they have several drawbacks (1) they are very expensive for casual use, (2) they hinder the naturalness of hand gesture, and (3) they require complex calibration and setup procedures to obtain precise measurements.

Vision-based hand gesture recognition serves as a promising alternative to them because of its potential to provide more natural, unencumbered, non-contact interaction. However, despite lots of previous work [5], [17], [19], traditional vision-based hand gesture recognition methods are still far from satisfactory for real-life applications. Because of the limitations of the optical sensors, the quality of the captured images is sensitive to lighting conditions and cluttered backgrounds, thus it is usually not able to detect and track the hands robustly, which largely affects the performance of hand gesture recognition. Recently, [16] presented a novel dissimilarity distance metric, Finger-Earth Mover's Distance, for hand gesture recognition approach using Kinect depth sensor, which performs accurately, efficiently and robustly on a 10-gesture dataset. FEMD metric is specifically designed for hand shapes. It considers each finger as a cluster and penalizes unmatched fingers. In this paper, we compare the FEMD based hand gesture recognition system with traditional corresponding based matching algorithm, Shape Context. And we built several HCI applications on top of this novel hand gesture recognition system and demonstrate its potential in other real-life HCI applications.

II. Related work

First, we give a brief overview of traditional vision-based hand gesture recognition approaches, see [8], [14] for more complete reviews.

From the perspective of extracted feature, vision-based hand gesture recognition methods can be classified into three types:

1. The first type is *High-level feature based approaches*: High-level feature based approaches attempt to infer the pose of the palm and the joint angles from high-level features, such as the fingertip, joint locations or some anchor points on the palm [5]. Colored markers are often used for feature extraction.
   A common problem with the high-level feature based approaches is in feature extraction. Point features are susceptible to occlusions, thus it is difficult to track markers on the image plane due to frequent collisions and/or occlusions [10]. Non-point features, such as protrusions of hand silhouettes [17], were sensitive to segmentation performance. Moreover, none of the proposed approaches of this type, including the ones

with colored markers, operate in the presence of cluttered backgrounds.

2. The second type is *3D feature based approaches*: Use of 3D features is limited to a few studies. In [3], structured light was used to acquire 3D depth data; however, skin color was used for segmenting the hand as well, which requires homogeneous and high contrast background relative to the hand. Another study [6] proposed to track a large number of interest points on the surface of the hand using a stereo camera. Motion information obtained from the 3D trajectories of the points was used to augment the range data. One can also create a full 3D reconstruction of the hand surface using multiple views. However, although 3D data contains valuable information that can help eliminate ambiguities due to self-occlusions which are inherent in image-based approaches, an exact, real-time, and robust 3D reconstruction is very difficult. Besides, the additional computational cost hinders its application in real-life systems.
3. The third type is *Low-level feature based approaches*: In many gesture applications all that is required is a mapping between input video and gesture. Therefore, many have argued that a full reconstruction of the hand is not necessary for gesture recognition. Instead, many approaches have utilized the extraction of low-level features that are fairly robust to noise and can be extracted quickly. In [18], the principle axes defining an elliptical bounding region of the hand was applied for hand gesture recognition. [23] proposed to use the optical flow/affine flow of the hand region in a scene as the low-level feature. Besides, contour and edges are universal features that can be used in any model-based technique [13]. However, low-level feature based measures are not effective under cluttered backgrounds. In [19], skin color model was employed to increase robustness, while also restricted the background setting.

From the perspective of processing scheme, vision-based hand gesture recognition methods can be classified into two categories:

1. The first category is *Machine Learning based approaches*: For a dynamic gesture, by treating it as the output of a stochastic process, the hand gesture recognition can be addressed based on statistical modeling, such as PCA, HMMs [12], [22], and more advanced particle filtering [11] and condensation algorithms [7].
2. The second category is *Rule based approaches*: Rule based approaches consist of a set of pre-encoded rules between feature inputs, which are applicable for both dynamic gestures and static gestures. A set of features of an input gesture are extracted and compared to the encoded rules, the gesture with the rule that matches the input is outputted as the recognized gesture [20].

As we see, traditional hand gesture recognition methods all applied restrictions on the user or environment because of the limitations of the optical sensors, which greatly hinders its widespread use in our daily life. To enable a more robust hand gesture recognition, one effective way is to use other sensors to capture the hand gesture and motion. [16] presented a novel hand gesture recognition method using Kinect depth sensor as the input device, which is stated to be accurate, efficient and robust to cluttered backgrounds, shape variations or distortions.

III. Hand Gesture Recognition

Since the release of Kinect [1], there have been numerous inspiring successes in applying Kinect for articulated human body tracking and face tracking [4], [1]. However, people find that although Kinect works well to track a large object, e.g. the human body, it is difficult to accurately detect and recognize a small object, such as a human hand.

Ren *et al*. [16] proposed a robust hand gesture recognition system based on a novel metric called Finger-Earth Mover's Distance, which not only is accurate and efficient, but also performs robustly to shape variations and distortions in uncontrolled environments. Its definition is in Eq.1.

With this novel distance metric FEMD, I introduce the framework of the hand gesture recognition system in Fig.1. It has the following properties:

- It uses Kinect sensor as the input device, which captures both the color image and its corresponding depth map.
- With the help of the depth cue, it can accurately detect the user′s hand, which is robust to cluttered backgrounds and various lighting conditions.
- After obtaining the hand contour, it relies on a novel hand shape matching algorithm for accurate and fast hand gesture recognition.

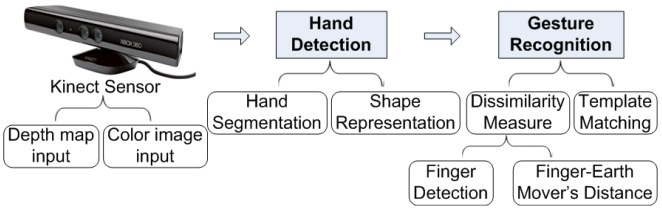

Fig. 1. Framework of the hand gesture recognition system based on FEMD [16].

Now we briefly introduce the framework, which mainly compose of two modules, hand detection and gesture recognition.

A. Hand Detection

Different from traditional methods that use color-markers for hand detection, [16] uses both the depth map and color image obtained by Kinect sensor to segment the hand shapes, which ensures its robustness to cluttered background, as shown in Fig.2. And a hand shape can be represented as a signature with each finger as a cluster.

TABLE I
TTHE MEAN ACCURACY AND THE MEAN RUNNING TIME OF SHAPE CONTEXT AND FEMD BASED HAND GESTURE RECOGNITION METHODS.

	Mean Accuracy	Mean Running Time
Shape Context [2]	83.2%	12.346s
Thresholding Decomposition+FEMD	90.6%	**0.5004**s
Near-convex Decomposition+FEMD	**93.9**%	4.0012s

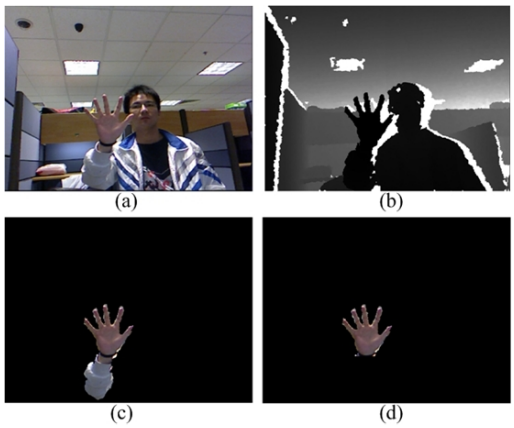

Fig. 2. Hand Detection process. (a). The RGB color image captured by Kinect Sensor; (b). The depth map captured by Kinect Sensor; (c). The area segmented using depth information; (d). The hand shape segmented using RGB information.

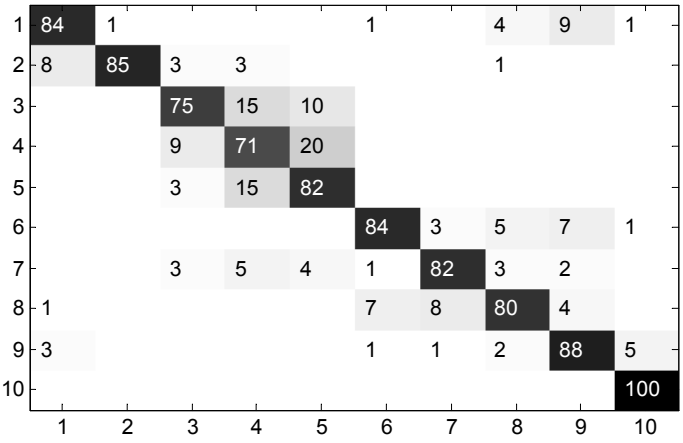

Fig. 3. The confusion matrix of hand gesture recognition using Shape Context [2]. The most confusing classes are among gesture 3, 4, and 5.

B. Gesture Recognition

The core of this hand gesture recognition system is the hand shape matching method for measuring the dissimilarities between different hand shapes, namely Finger-Earth Movers Distance (FEMD), which outperforms the state-of-the-arts in terms of accuracy, efficiency and robustness, especially under severe shape distortions or variations.

After converting the detected hand shapes into signatures, according to [16], the FEMD distance between two signatures, R and T, is formulated as follows:

$$\begin{aligned} \text{FEMD}(R,\ T) &= \beta E_{move} + (1-\beta)E_{empty}, \\ &= \frac{\beta \sum_{i=1}^{m}\sum_{j=1}^{n} d_{ij}f_{ij} + (1-\beta)|\sum_{i=1}^{m} w_{\mathbf{r}_i} - \sum_{j=1}^{n} w_{\mathbf{t}_j}|}{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}}, \end{aligned} \quad (1)$$

where $\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}$ is the normalization factor, f_{ij} is the flow from cluster $\mathbf{r}_i$ to cluster $\mathbf{t}_j$, which constitutes the flow matrix $\mathbf{F}$. d_{ij} is the ground distance from cluster $\mathbf{r}_i$ to $\mathbf{t}_j$. Parameter β modulates the importance between the first and the second terms. For detail explanation please check [16]. To compute the FEMD distance, we need to construct the signature representation of a hand shape, where each cluster represents a finger. So the key for FEMD measure is finger detection.

C. Finger Detection

[16] proposed two ways for finger detection. One is called thresholding decomposition, which defines a finger cluster as a segment in the time-series curve whose height is greater than a threshold h_f . Another one is based on a near-convex shape decomposition method presented in [15], which formulated the shape decomposition as a combinational optimization problem.

IV. PERFORMANCE COMPARISON

With the FEMD based hand gesture recognition system, [16] reported impressive performance by testing on a 10-gesture dataset which contains both color images and their corresponding depth maps. We compare the mean accuracy and mean running time between FEMD based hand gesture recognition system and Shape Context shape matching algorithm in Table I. As we see, because of the compactness of thresholding decomposition based finger detection, it achieves 0.5004 second in Matlab code. And the near-convex decomposition based FEMD achieves 93.9% mean accuracy because of the more accurate finger detection method. We can see that FEMD based methods outperform traditional shape matching algorithm, Shape Context, both in speed and accuracy.

Fig. 4. Some confusing cases for shape context [2] whose shapes are distorted.

Fig.3 illustrates the confusion matrixes of Shape Context [2]. We find that the most confusing classes are among gesture 3, 4, and 5. The reason is that among these classes, fingers are

more easily distorted and make them indistinguishable. Fig.4 shows some confusing cases for shape context whose shapes are distorted.

V. Applications

In the last section, we illustrate the accuracy and efficiency of FEMD hand gesture recognition on a 10-gesture dataset. And now we further demonstrate this FEMD based hand gesture recognition system [16] in a real-life HCI application: Sudoku game.

Our current system is built on an Intel Core TM 2 Quad 2.66 GHz CPU with 3GB of RAM and a Kinect depth camera (driver Sensorkinect version 5.0.0). We developed the system on top of OpenNI open platform for Natural Interaction (version 1.0.0.25 binaries) and OpenNI compliant middleware binaries provided by PrimeSense (NITE version 1.3.0.18). Details of OpenNI can be found at http://www.openni.org/.

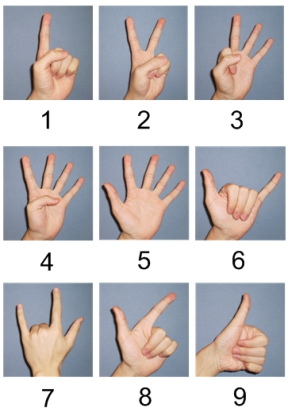

Fig. 5. The 9 gesture commands adopted in Sudoku game.

A. Sudoku game

Sudoku is a popular number-placement puzzle. The objective is to fill a 9×9 grid with digits so that each column, each row, and each of the nine 3×3 sub-grids contain digits from 1 to 9 without repetition. As shown in Fig.5, 9 hand gestures are employed to represent 9 commands, namely number from number 1 to number 9.

The puzzle begins with a partially completed grid and typically has a unique solution. Fig.6 shows an example of Sudoku puzzle. The user selects a square by hovering his hand over it and pushes ("clicks") once. He/She then commands a number to be filled into the square by performing the corresponding hand gesture in Fig.5. The system recognizes the number and fills it into the square and check whether it is correct in the end.

B. Discussion

Although we only introduce one HCI application Sudoku game using the FEMD based hand gesture recognition approach [16], it has the potential to be applied in many other HCI applications that adopt hand as the interface. For instance,

Fig. 6. Illustration of Sudoku game

it can be used for other entertainments, like Rock-paper-scissors game and gobang. Besides, it is very useful for medical systems that require a touchless interface, assisted living for those with disabilities and sign language recognition. The expected market covers every aspect of our society, including health care, education, entertainment, assisted living, human-robot interaction, etc. Among them, the use of our system in education and healthcare are especially of great societal importance.

VI. Conclusion

In this paper, we demonstrate the accuracy, efficiency, and effectiveness of the FEMD based hand gesture recognition system [16] by comparing its performance on a 10-gesture dataset with Shape Context, as well as a real-life HCI application built on top of it. [16] proposed a novel distance metric called Finger-Earth Mover's Distance (FEMD) to measure the hand shape dissimilarities, which not only achieves better performance than Shape Context, in terms of accuracy (a 93.9% mean accuracy) and efficiency (0.5004s). On the other hand, this hand gesture recognition system has a high potential to benefit many real-life HCI applications in entertainment, education, health care, and so on. We build a Sudoku game HCI system on top of this hand gesture recognition method to show its applicability to be a key enabler for many other hand gesture based HCI systems.

Acknowledgment

This work was supported in part by the Nanyang Assistant Professorship (SUG M58040015) to Dr. Junsong Yuan.

References

[1] Microsoft Corp. Redmond WA. Kinect for Xbox 360.
[2] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE Trans. on PAMI*, 24:509–522, 2002.
[3] M. Bray, E. Koller-Meier, and L. V. Gool. Smart particle filtering for 3d hand tracking. In *Proc. of Sixth IEEE International Conf. on Face and Gesture Recognition*, 2004.
[4] Q. Cai, D. Gallup, C. Zhang, and Z. Zhang. 3d deformable face tracking with a commodity depth camera. In *Proc. of IEEE ECCV*, 2010.
[5] C. Chua, H. Guan, and Y. Ho. Model-based 3d hand posture estimation from a single 2d image. *Image and Vision Computing*, 20:191 – 202, 2002.

[6] G. Dewaele, F. Devernay, and R. Horaud. Hand motion from 3d point trajectories and a smooth surface model. In *Proc. of 8th ECCV*, 2004.

[7] A. Doucet, N. de Freitas, and N. Gordon. *Sequential Monte Carlo in Practice*. NewYork: Springer-Verlag, 2001.

[8] A. Erol, G. Bebis, M. Nicolescu, R. D. Boyle, and X. Twombly. Vision-based hand pose estimation: A review. *CVIU*, 108:52 – 73, 2007.

[9] E. Foxlin. Motion tracking requirements and technologies. *Handbook of Virtual Environment Technology*, pages 163–210, 2002.

[10] E. Holden. *Visual recognition of hand motion*. Ph.D thesis, Department of Computer Science, University of Western Australia, 1997.

[11] C. Kwok, D. Fox, and M. Meila. Real-time particle filters. In *Proc. of IEEE*, 2004.

[12] H. Lee and J. Kim. An hmm-based threshold model approach for gesture recognition. *IEEE Trans. on PAMI*, 21:961 – 973, 1999.

[13] D. Lowe. Fitting parameterized three-dimensional models to images. *IEEE Trans. on PAMI*, 13:441 – 450, 1991.

[14] S. Mitra and T. Acharya. Gesture recognition: A survey. *IEEE Trans. on Systems, Man and Cybernetics-Part C: Application and Review*, 37:311 – 324, 2007.

[15] Z. Ren, J. Yuan, C. Li, and W. Liu. Minimum near-convex decomposition for robust shape representation. In *Proc. of ICCV*, 2011.

[16] Z. Ren, J. Yuan, and Z. Zhang. Robust hand gesture recognition based on finger-earth movers distance with a commodity depth camera. In *Proc. of ACM Multimeida*, 2011.

[17] N. Shimada, Y. Shirai, Y. Kuno, and J. Miura. Hand gesture estimation and model refinement using monocular camera-ambiguity limitation by inequality constraints. In *Proc. of Third IEEE International Conf. on Face and Gesture Recognition*, 1998.

[18] T. Starner, J. Weaver, and A. Pentland. Real-time american sign language recognition using desk and wearable computer based video. *IEEE Trans. on PAMI*, 20:1371 – 1375, 1998.

[19] B. Stenger, A. Thayananthan, P. Torr, and R. Cipolla. Filtering using a tree-based estimator. In *Proc. of IEEE ICCV*, 2003.

[20] M.-C. Su. A fuzzy rule-based approach to spatio-temporal hand gesture recognition. *IEEE Trans. on Systems, Man and Cybernetics-Part C: Application and Review*, 30:276 – 281, 2000.

[21] J. P. Wachs, M. Klsch, H. Stern, and Y. Edan. Vision-based hand-gesture applications. *Communications of the ACM*, 54:60–71, 2011.

[22] A. Wilson and A. Bobick. Parametric hidden markov models for gesture recognition. *IEEE Trans. on PAMI*, 21:884 – 900, 1999.

[23] M. H. Yang, N. Ahuja, and M. Tabb. Extraction of 2-d motion trajectories and its application to hand gesture recognition. *IEEE Trans. on PAMI*, 29:1062 – 1074, 2002.