A hand gesture recognition system based on canonical superpixel-graph

Chong Wang, Zhong Liu, Minfeng Zhu, Jieyu Zhao, Shing-Chow Chan

PII: S0923-5965(17)30122-4
DOI: http://dx.doi.org/10.1016/j.image.2017.06.015
Reference: IMAGE 15249

To appear in: *Signal Processing: Image Communication*

Received date : 4 November 2016
Revised date : 27 June 2017
Accepted date : 27 June 2017

Please cite this article as: C. Wang, Z. Liu, M.G. Zhu, J.Y. Zhao, S.C. Chan, A hand gesture recognition system based on canonical superpixel-graph, *Signal Processing: Image Communication* (2017), http://dx.doi.org/10.1016/j.image.2017.06.015

A Hand Gesture Recognition System based on Canonical Superpixel-Graph

Chong Wang, Zhong Liu, Minfeng Zhu, Jieyu Zhao and Shing-Chow Chan

Abstract—**This paper presents a new hand gesture recognition system based on a novel canonical superpixel-graph earth mover's distance (CSG-EMD) metric. It aims to improve the performance of the superpixel earth mover's distance (SP-EMD), a recently proposed distance metric designed for depth-based hand gesture recognition. In real life, people have their own habits while performing certain hand gestures, which yields a variety of hand shapes with different finger poses. Such variety may affect the accuracy of SP-EMD and hence will degrade its performance. In this paper, we propose a new distance metric CSG-EMD to alleviate the problem. Scattered superpixels are organized in the form of canonical superpixel-graph which can factor out non-standard finger poses, resulting a well-structured finger-pose-neutral shape representation for hand gestures. Moreover, a structure stress based fusion scheme is applied to formulate the proposed distance metric, i.e. CSG-EMD, for gesture recognition. Experimental results on five public gesture datasets show that the proposed CSG-EMD-based system can achieve better recognition accuracy than other state-of-the-art algorithms compared. Its superiority is further demonstrated by two real-life applications.**

***Index Terms*—Hand Gesture Recognition, Kinect, Superpixel Earth Mover's Distance, Canonical Forms.**

I. Introduction

Vison-based hand gesture recognition has received great attention in recent years due to its importance in contactless human-computer interaction (HCI) scenarios such as virtual reality, sign language recognition [2, 3]. Although a lot of related work has been done [4-8], robust vision-based hand gesture recognition remains a challenging problem. In particular, reliable hand segmentation is essential to gesture recognition and many hand detection techniques [9] have been developed for tracking and recognizing different kinds of hand gestures. With the recent development of depth cameras, such as Microsoft Kinect, Creative Senz3D or Intel RealSense, the work of hand detection and segmentation can be greatly simplified thanks to the extra depth information provided. For instance, threshold-based methods [10-12] are commonly used for efficient hand segmentation. Moreover, some researchers [12, 13] take advantage of the skeleton recognized by Kinect for more convenient hand localization and tracking.

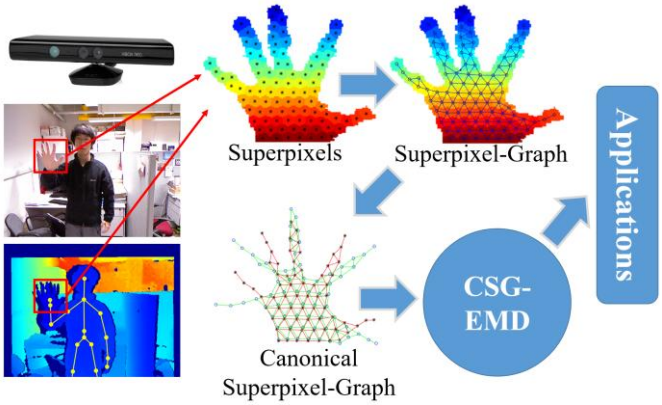

Fig. 1. The workflow of our proposed hand gesture recognition system based on CSG-EMD.

Once the hands have been isolated, a variety of hand features can be extracted and used for hand gesture recognition. The hand features can be roughly classified into three categories, i.e. depth-based, color-image-based and shape-based features. Some typical depth-based features are Histogram of 3D Facets (H3DF) [14] and 3D point distribution histogram [15], which are not sensitive to the lighting conditions. One classical but efficient color-based feature descriptor is Histogram of Oriented Gradients (HOG) [16] which is invariant to geometric and photometric transformations. Meanwhile, there are many shape-based features utilized for gesture recognition such as hand contours [2], hand skeleton [17], shape context [18], inner distance [19] and Finger-Earth Mover's Distance (FEMD) [11]. Most of them are generated based on the hand contour, which is, however, usually noisy and distorted due to the low resolution and accuracy of the current depth cameras. In other words, their performance may suffer from ambiguity due to the orientation,

Manuscript received *. *, *. A preliminary version of this paper appeared in the IEEE International conference on Multimedia and Expo [1]. This work was supported by K.C. Wong Magna Fund in Ningbo University, National Natural Science Foundation of China (61603202, 61571247), the Open Project Program of the State Key Lab of CAD&CG in Zhejiang University (A1606), Zhejiang Provincial Natural Science Foundation of China (LQ16F030001, LZ16F030001, LY17F030002), Zhejiang Open Foundation from Information and Communication Engineering of the Most Important Subjects, China (xkxl1512, xkxl1526), Ningbo Municipal Natural Science Foundation of China (2016A610070), Research Foundation of Education Department of Zhejiang Province, China (Y201533827) and the General Research Fund (GRF) of Hong Kong Research Grant Council (RGC).

C. Wang, M. F. Zhu J. Y. Zhao are with the Faculty of EECS, Ningbo University, Ningbo, P.R.China (e-mail: wangchong@nbu.edu.cn, zhao_jieyu@nbu.edu.cn).

Z. Liu and S. C. Chan are with the Department of EEE, The University of Hong Kong, Hong Kong. (E-mail: liuzhong@eee.hku.hk, scchan@eee.hku.hk).