Fast and Robust Dynamic Hand Gesture Recognition via Key Frames Extraction and Feature Fusion

Hao Tang[1], Hong Liu[2*], Wei Xiao[3], Nicu Sebe[1]

[1] *Department of Information Engineering and Computer Science, University of Trento, Trento, Italy*
[2] *Key Laboratory of Machine Perception, Shenzhen Graduate School, Peking University, Beijing, China*
[3] *Lingxi Artificial Intelligence Co., Ltd, Shen Zhen, China*

Abstract

Gesture recognition is a hot topic in computer vision and pattern recognition, which plays a vitally important role in natural human-computer interface. Although great progress has been made recently, fast and robust hand gesture recognition remains an open problem, since the existing methods have not well balanced the performance and the efficiency simultaneously. To bridge it, this work combines image entropy and density clustering to exploit the key frames from hand gesture video for further feature extraction, which can improve the efficiency of recognition. Moreover, a feature fusion strategy is also proposed to further improve feature representation, which elevates the performance of recognition. To validate our approach in a "wild" environment, we also introduce two new datasets called HandGesture and Action3D datasets. Experiments consistently demonstrate that our strategy achieves competitive results on Northwestern University, Cambridge, HandGesture and Action3D hand gesture datasets. Our code and datasets will release at https://github.com/HaoTang/HandGestureRecognition.

Keywords: Hand gesture recognition; Key frames extraction; Feature fusion; Fast; Robust.

1. Introduction

Gesture recognition is to recognize category labels from an image or a video which contains gestures made by the user. Gestures are expressive, meaningful body motions involving physical movements of the fingers, hands, arms, head, face, or body with the intent of: conveying meaningful information or interacting with the environment.

Hand gesture is one of the most expressive, natural and common type of body language for conveying attitudes and emotions in human interactions. For example, in a television control system, hand gesture has the following attributes: "Pause","Play", "Next Channel", "Previous Channel", "Volume Up", "Volume Down" and "Menu Item". While in a recommendation system, hand gesture can express "Like" or "Dislike" emotions of users. Thus, it is one of the most fundamental problems in computer vision and pattern recognition, and has a wide range of applications such as virtual reality systems [1], interactive gaming platforms [2], recognizing sign language [3, 4, 5], enabling very young children to interact with computers [6], controlling robot [7, 8], practicing music conducting [9], television control [10, 11], automotive interfaces [12, 13], learning and teaching assistance [14, 15], and hand gesture generation [16].

There has been significant progress in hand gesture recognition, however, some key problems e.g., fast and robust are still challenging. Prior work usually puts emphasis on using whole data series, which always contain redundant information, resulting in degraded performance. For examples, Wang et al. [1] present a superpixel-based hand gesture recognition system based on a novel superpixel earth mover's distance metric. Ren et al. [2] focus on building a robust part-based hand gesture recognition system. Hikawa and Kaida [3] propose a posture recognition system with a hybrid network. Moreover, there are many approaches are also proposed for action or video recognition task, such as [17, 18, 19, 20, 21, 22, 23, 24, 25, 26]. Liu and Shao [17] introduce an adaptive learning methodology to extract spatio-temporal features, simultaneously fusing the RGB and depth information, from RGB-D video data for visual recognition tasks. Liu et al. [26] propose to combine the Salient Depth Map (SDM) and the Binary Shape Map (BSM) for human action recognition task. Simonyan et al. [27] propose a two-stream ConvNet architecture which incorporates spatial and temporal networks to extract spatial and temporal features. Feichtenhofer et al. [28] study a number of ways of fusing ConvNet towers both spatially and temporally in order to best take advantage of this spatio-temporal information. In sum, all these efforts endeavor to decrease the computation burden in each solo frame, while overlooking all processing schemes in the whole frames would incur more computation burden than

☆E-mail: {hao.tang, niculae.sebe}@unitn.it; hongliu@pku.edu.cn; xiaoweithu@163.com;
*Corresponding author.