Image and Vision Computing 32 (2014) 522–532

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

Real-time fingertip localization conditioned on hand gesture classification☆

Xavier Suau [a,*], Marcel Alcoverro [a], Adolfo López-Méndez [b], Javier Ruiz-Hidalgo [a], Josep R. Casas [a]

[a] Universitat Politècnica de Catalunya, 1-3, Jordi Girona, 08034 Barcelona, Spain

[b] FEZOO S.C.P., 13, Avinguda Corts Catalanes, 2F, 08173 Sant Cugat del Vallès, Barcelona, Spain

ARTICLE INFO

Article history:
Received 25 July 2013
Received in revised form 6 March 2014
Accepted 30 April 2014
Available online 9 May 2014

Keywords:
Hand gesture recognition
Fingertip classification
Range camera
Interactivity
Dataset

ABSTRACT

A method to obtain accurate hand gesture classification and fingertip localization from depth images is proposed. The Oriented Radial Distribution feature is utilized, exploiting its ability to globally describe hand poses, but also to locally detect fingertip positions. Hence, hand gesture and fingertip locations are characterized with a single feature calculation. We propose to divide the difficult problem of locating fingertips into two more tractable problems, by taking advantage of hand gesture as an auxiliary variable. Along with the method we present the ColorTip dataset, a dataset for hand gesture recognition and fingertip classification using depth data. ColorTip contains sequences where actors wear a glove with colored fingertips, allowing automatic annotation. The proposed method is evaluated against recent works in several datasets, achieving promising results in both gesture classification and fingertip localization.

1. Introduction

Until recent years, interaction between humans and computer systems has been typically performed with peripherals (e.g. mouse, keyboard). Current trends focus on improving the user experience when interacting with computers and systems. Apple's trackpad [1] or multi-touch devices are commercially successful examples of new interaction paradigms; they have created a *habit* in users thanks to their easy-to-learn and intuitive gestures combining simple movements and finger configurations. However, these methods are limited to systems where the user is required to physically *touch* the device.

Touch-less interaction is an emerging alternative providing a more immersive and intuitive experience. In this paper, we propose a touch-less interaction paradigm that aims at extending multi-touch gestural interaction to spatial processing. Consequently, hand gestures and finger configurations are combined with simple movements in order to deliver a large number of interaction options with a rather low number of one-hand gestures. In order to put this strategy into operation, accurate real-time fingertip detection is required.

Self-occlusions, limited resolution, the degrees of freedom of a hand and the intra-class variability of gestures render hand gesture recognition and fingertip detection as challenging problems. Fortunately, a number of traditional vision challenges involving scale and illumination changes have been overcome thanks to the irruption of consumer depth cameras (e.g. Kinect). Such depth cameras have paved the way towards touch-less interactivity; as a remarkable example, it has taken a few years for full-body pose interaction [2] to achieve commercial success in gaming. However, few works tackle fingertip detection from depth cameras [3–6] (see Section 2).

The objective of this work is to locate fingertips in real-time. More precisely, to know *where* fingers are placed (detection), and *which* finger is each (classification). Instead of facing the problem from raw data, as in [2], we propose an intermediate step to restrict the search space. Here, the intuition that fingertip locations are conditioned by hand gestures is exploited, and formulated such that the statistical correlation between gestures and fingertip locations restricts the search space of fingertip configurations, rendering faster and more accurate inference. In a first step, the most probable hand gesture labels are inferred from data. Next, fingertip locations are inferred from depth data and most probable hand gesture label, which is used as a discriminative auxiliary variable. For this second step, a graph matching approach exploiting fingertip structure is proposed. Instead of hand gesture labels, other variables, such as hand orientation, could have been chosen to restrict the search space. However, gestures have a semantic purpose on their own, and are easier to annotate.

In this work, we also propose a novel usage of the Oriented Radial Distribution (ORD) feature, presented in [7]. The ORD feature characterizes a point cloud such that end-effectors have high responses, whereas flat parts have low responses. Therefore, ORD is suitable to both globally characterize the structure of a hand gesture and to locally locate its end-effectors. Such property nicely fits in the above mentioned two-step method: globally for the hand gesture classification task and locally

☆ This paper has been recommended for acceptance by Matthew Turk.

* Corresponding author.
E-mail address: xavier.suau@upc.edu (X. Suau).

http://dx.doi.org/10.1016/j.imavis.2014.04.015
0262-8856/© 2014 Elsevier B.V. All rights reserved.

for the fingertip detection step. That is, a single ORD calculation suffices for both tasks.

Available datasets for gesture recognition with depth data have been proposed to typically respond to specific needs of the aimed interaction paradigms. Ganapathi et al. [8] provide a body pose estimation dataset using a Time-of-Flight (TOF) camera. Pugeault and Bowden [9] propose a hand gesture dataset using Kinect, which is intended for American Sign Language (ASL) purposes. Although useful for the evaluation of some tasks, such as hand gesture recognition, none of the available datasets is suitable to test the interaction paradigm proposed in this paper. For that matter, we present ColorTip [10], a depth-based dataset consisting of 7 subjects performing 9 different hand gestures (Figs. 1 and 2). Ground-truth annotations for hand positions, hand gestures, fingertip locations and finger labels are also provided. Finger positions are obtained using a colored glove during capture, enabling a non-costly color-wise segmentation. Furthermore, each subject performs two sequences (*Set A* and *Set B*), with increased intra-gesture variability in the latter.

Summarizing, in this work we propose the following main contributions:

- A practical touch-less interaction concept, combining finger configurations, hand gesture and simple movements.
- A real-time method to obtain fingertip locations and labels, as well as hand gestures, using Kinect. We propose to exploit the statistical correlation between hand gestures and fingertip locations.
- A novel use of the Oriented Radial Distribution feature, exploiting its global structure for hand gesture characterization and its local values for fingertip detection.
- ColorTip, a public dataset intended for hand gesture classification and fingertip localization.

The effectiveness of the proposed method is experimentally evaluated in different datasets. Furthermore, we conduct experiments assessing the performance of each aspect of our approach. At the feature level, we show the validity of the ORD feature by means of a 3D feature benchmark. Next, hand gesture classification accuracy is evaluated in the ASL database provided by [9]. Finally, fingertip localization results are compared to a state-of-the-art Random Forest (RF) approach using the ColorTip dataset.

The remainder of the paper is organized as follows. A summary of related work is provided in Section 2. In Section 3 we present the ColorTip dataset. The Oriented Radial Distribution feature is described in Section 4. Section 5 contains the theoretical description of the proposed method, followed by experimental results in Section 6. Finally, conclusions are drawn and discussed in Section 7.

2. Related work

An early hand gesture recognition from depth data [11] tackles hand gesture recognition using a laser-based camera to produce low-resolution depth images. They interpolate hand pose using sets of finger poses and inter-relations. Liu and Fujimura [12] recognize dynamic hand gestures using Time-of-Flight depth images. Hands are detected measuring shape similarity by means of the Chamfer distance. They analyze the trajectory of the hand and classify gestures using shape, location, trajectory, orientation and speed features. As range sensors become progressively cheaper, the number of approaches towards touch-less interactivity grows, most remarkably full body pose estimation [2,13,8,14] and hand gesture recognition [15–17].

Many authors have explored how to control a virtual environment with hands (e.g. PC desktop, 3D model). Such applications generally involve dynamic hand gesturing. In this regard, Soutschek et al. [15] propose a user interface for the navigation through 3D datasets using a Time-of-Flight (TOF) camera. They perform a polar crop of the hand over a distance threshold to the centroid, and a subsequent NN classification into five hand gestures. With a similar objective, Van den Berg and Van Gool [18] improve their work in [17] by combining RGB and depth to construct classification vectors. Their alphabet consists of four gestures that enable selecting, rotating, panning and zooming of a 3D model on a screen. Hackenberg et al. [3] estimate hand pose by identifying palm and finger candidates, after a pixel-wise classification into tips and pipes. The final hand structure is obtained with optical flow techniques. Ren et al. [19] segment the hand under some restrictive assumptions and adapt the earth mover's distance to a finger signature, finding the NN according to this metric. Malassiotis and Strintzis [16] extract PCA features from depth images of synthetic 3D hand models for training.

Obtaining hand gestures with the Nearest Neighbor (NN) classification has proven to be a promising approach when dealing with depth data [15,19,20]. However, most recent works use features that are not specifically designed for depth data.

Other works have focused on finger-spelling using the American Sign Language (ASL). While still being an alphabet, the ASL contains 26 hand poses and their accurate classification becomes a challenging task. We remark that 24 of the 26 hand poses are static gestures and 2 of them are dynamic (involve trajectory). Most of the related works are focused on the static subset. Keskin et al. [21] take advantage of Randomized Decision Trees to classify hand shapes. Zhang et al. [22] recently propose a descriptor for depth data which encodes 3D facets into a histogram. They prove the suitability of this descriptor for hand gesture recognition on ASL datasets. Zhu and Wong [23] propose to fuse common color and depth descriptors and use linear SVMs to predict the hand gesture. Kollorz et al. [20] obtain a fast NN classification using simple feature projection on two axes, which they apply to the first 12 letters of the ASL (static gestures). Uebersax et al. [24] perform an iterative hand segmentation by optimizing the center, orientation and size of the hand. They aggregate three classifiers that take shape and orientation into account. Pugeault and Bowden [9] propose a multi-resolution Gabor filtering of the hand patch to train a Random Forest classifier. In their work, they provide a complete dataset of the 24 American Sign Language (ASL) static gestures captured with the Kinect sensor, with both color and depth information available. Their dataset contains patches roughly centered at the hand centroid.

Fewer works tackle fingertip localization. In [3], fingertips are detected but not labeled, as well as in [4] where also the palm and fingers orientations are estimated. Both approaches exploit geometric features to detect fingertips on the hand point cloud. The body part classification

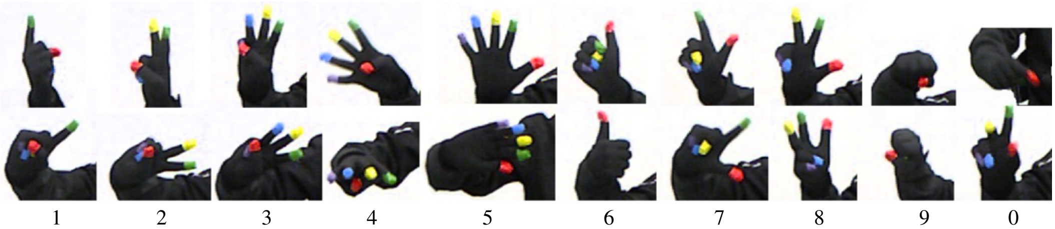

Fig. 1. Sample of the annotated gestures in the ColorTip dataset. Two examples per gesture are shown (columns). These examples are extracted from a *Set B* sequence, with a high intra-gesture variation. Note the rotations and translations. Label 0 corresponds to *no gesture* (i.e. other gestures, transitions).