Pattern Recognition Letters 50 (2014) 101–111

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Combining multiple depth-based descriptors for hand gesture recognition

Fabio Dominio, Mauro Donadeo, Pietro Zanuttigh [*]

Department of Information Engineering, University of Padova, Italy

ARTICLE INFO

Article history:
Available online 25 October 2013

Communicated by Dmitry Goldgof

Keywords:
Gesture recognition
Support Vector Machines
Depth
Kinect

ABSTRACT

Depth data acquired by current low-cost real-time depth cameras provide a more informative description of the hand pose that can be exploited for gesture recognition purposes. Following this rationale, this paper introduces a novel hand gesture recognition scheme based on depth information. The hand is firstly extracted from the acquired data and divided into palm and finger regions. Then four different sets of feature descriptors are extracted, accounting for different clues like the distances of the fingertips from the hand center and from the palm plane, the curvature of the hand contour and the geometry of the palm region. Finally a multi-class SVM classifier is employed to recognize the performed gestures. Experimental results demonstrate the ability of the proposed scheme to achieve a very high accuracy on both standard datasets and on more complex ones acquired for experimental evaluation. The current implementation is also able to run in real-time.

1. Introduction

Hand gesture recognition is an intriguing problem for which many different approaches exist. Even if gloves and various wearable devices have been used in the past, vision-based approaches able to capture the hand gestures without requiring any physical device to be worn allow a more natural interaction with computers and many other devices. This problem is currently raising a high interest due to the rapid growth of application fields where it can be efficiently applied, as reported in recent surveys (e.g., Wachs et al., 2011; Garg et al., 2009). These include human–computer interaction, where gestures can be used to replace the mouse in computer interfaces and also to allow a more natural interaction with mobile and wearable devices like smartphones, tablets or newer devices like the Google glasses. Also the navigation of 3D virtual environments is more natural if controlled by gestures performed in the 3D space. In robotics gestures can be used to control and interact with the robots in a more natural way. Another key field is computer gaming, where devices like Microsoft's Kinect have already brought gesture interfaces to the mass market. Automatic sign-language interpretation will also allow to help hearing and speech impaired people to interact with the computer. Hand gesture recognition can be applied in the healthcare field to allow a more natural control of diagnostic data and surgical devices. Gesture recognition is also being considered for vehicle interfaces.

Several hand gesture recognition approaches, based on the analysis of images and videos, can be found in literature (Wachs et al., 2011; Zabulis et al., 2009). Images and videos provide a bidimensional representation of the hand pose, which is not always sufficient to capture the complex movements and inter-occlusions characterizing hand gestures. Three dimensional representations offer a more accurate description of the hand pose, but are more difficult to acquire. The recent introduction of low-cost consumer depth cameras, such as Time-Of-Flight cameras and Microsoft's Kinect^{TM}, has made depth acquisition available to the mass market, thus widely increasing the interest in gesture recognition approaches taking advantage from three-dimensional information.

In order to recognize the gestures from depth data the most common approach is to extract a set of relevant features from the depth maps and then exploit machine learning techniques to the extracted features. Kurakin et al. (2012) uses a single depth map and extract silhouette and cell occupancy features for building a shape descriptor that is then fed into a classifier based on action graphs. Suryanarayan et al. (2010) extract 3D volumetric shape descriptors from the hand depth to be classified with a Support Vector Machine. Volumetric features and an SVM classifier are also used by Wang et al. (2012). In Keskin et al. (2012) the classification is instead performed using Randomized Decision Forests (RDFs). RDFs are also used by Pugeault and Bowden (2011) that also combines together color and depth information to improve the accuracy of the classification. Another approach consists in analysing the segmented hand shape and extract features based on the

[*] Corresponding author. Address: Dept. of Information Engineering, Via Gradenigo 6/B, 35131 Padova, Italy. Tel.: +39 049 827 7782; fax: +39 049 827 7699.

E-mail addresses: dominiof@dei.unipd.it (F. Dominio), donadeom@dei.unipd.it (M. Donadeo), zanuttigh@dei.unipd.it (P. Zanuttigh).

0167-8655/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.patrec.2013.10.010

convex hull and on the fingertips positions as in Wen et al. (2012) and Li (2012). A similar approach is used also by the Open-source library *XKin* (Pedersoli et al., 2012). Finally, Ren et al. (2011b) and Ren et al. (2011a) compare the histograms of the distance of hand edge points from the hand center.

If the target is the recognition of dynamic gestures, motion information and in particular the trajectory of the hand's centroid in the 3D space can be exploited (Biswas and Basu, 2011). In Doliotis et al. (2011) a joint depth and color hand detector is used to extract the trajectory that is then fed to a Dynamic Time Warping (DTW) algorithm. Finally, Wan et al. (2012) exploits both the convex hull on a single frame and the trajectory of the gesture. A related harder problem is the estimation of the hand pose from the depth data (Oikonomidis et al., 2011; Ballan et al., 2012; Keskin et al., 2011).

In most of the previously cited works depth data is mainly used to reliably extract the hand silhouette in order to exploit approaches derived from hand gesture recognition schemes based on color data. This paper instead uses a set of three-dimensional features to properly recognize complex gestures by exploiting the 3D information on the hand shape and finger posture contained in depth data. Furthermore instead of relying on a single descriptor extraction scheme, different types of features capturing different clues are combined together to improve the recognition accuracy. In particular the proposed hand gesture recognition scheme exploits four types of features: the first two sets are based on the distance from the palm center and the elevation of the fingertips, the third contains curvature features computed on the hand contour and the last set of features is based on the geometry of the palm region accounting also for fingers folded over the palm. The constructed feature vectors are then combined together and fed into an SVM classifier in order to recognize the performed gestures. The proposed approach introduces several novel elements: it jointly exploits color and depth data to reliably extract the hand region and is able to extract wrist, palm and finger regions; it fully exploits three-dimensional data for the feature extraction, and finally it combines features based on completely different clues to improve the recognition rate.

The paper is articulated as follows: Section 2 introduces the general architecture of the proposed gesture recognition system, Section 3 explains how the hand region is extracted from the acquired depth data and segmented into arm, palm and fingers regions. Section 4 describes the computation of the proposed feature descriptors, and Section 5 presents the classification algorithm. Section 6 reports the experimental results and finally Section 7 draws the conclusions.

2. Proposed gesture recognition system

The proposed gesture recognition system (Fig. 1) encompasses three main steps. In the first step the hand samples are segmented from the background exploiting both depth and color information. The previous segmentation is then refined by further subdividing the hand samples into three non overlapping regions, collecting palm, fingers and wrist/arm samples respectively. The last region is discarded, since it does not contain information useful for gesture recognition. The second step consists in extracting the four feature sets that will be used in order to recognize the performed gestures, i.e.:

- *Distance features:* this set describes the Euclidean 3D distances of the fingertips from the estimated palm center.
- *Elevation features:* this set accounts for the Euclidean distances of the fingertips from a plane fitted on the palm samples. Such distances may also be considered as the *elevations* of the fingers with respect to the palm.
- *Curvature features:* this set describes the curvature of the contour of the palm and fingers regions.
- *Palm area features:* this set describes the shape of the palm region and helps to state whether each finger is raised or bent on the palm.

Finally, during the last step, all the features are collected into a *feature vector* to be fed into a multi-class Support Vector Machine classifier in order to recognize the performed gesture.

3. Hand segmentation

The first step in the proposed method is the segmentation of the hand. Although depth information alone may be enough for this purpose, we exploit both depth and color information in order to recognize the hand more robustly. The data acquired by the Kinect^{TM} color camera is first projected on the depth map and both a color and a depth value are associated to each sample. Note that the Kinect^{TM} depth and color cameras have been previously jointly calibrated by the method proposed in Herrera et al. (2012). After projection, the acquired depth map $D(u,v)$ is thresholded on the basis of color information. More specifically, the colors associated to the samples are converted into the CIELAB color space and compared with a reference skin color that has been previously acquired.[1] The difference between each sample color and the reference skin color is evaluated and the samples whose color difference is below a pre-defined threshold are discarded. This first thresholding will only retain depth samples associated with colors compatible with the user's skin color that are very likely to belong to the hand, the face or other uncovered body parts. After the skin color thresholding the hand region has a higher chance to be the object nearest to the Kinect^{TM}. Note that this is the only step of the algorithm where color data is used. In applications where the hand is proven to be always the closest object to the sensor, the usage of color information may be skipped in order to simplify the acquisition of the data and to improve computation performances.

Let us denote with $\mathbf{X}_{u,v}$ a generic 3D point acquired by the depth camera, i.e., the back-projection of the depth sample in position (u,v). A search for the sample with the minimum depth value $D_{\min}$ on the thresholded depth map is performed. The corresponding point $\mathbf{X}_{\min}$ is chosen as the starting point for the hand detection procedure. In order to avoid to select as $\mathbf{X}_{\min}$ an isolated artifact due to measurement noise, our method verifies the presence of an adequate number of samples with a similar depth value in a 5×5 region around $\mathbf{X}_{\min}$. If the cardinality threshold is not satisfied we select the next closest point and repeat the check.

Let us now denote by $\mathcal{H}$ the hand samples set. Points belonging to $\mathcal{H}$ cannot have a depth that differs from $\mathbf{X}_{\min}$ of more than a value T_{depth} that depends on the hand size. $\mathcal{H}$ may be then expressed as:

$$\mathcal{H} = \{\mathbf{X}_{u,v} | D(u,v) < D_{\min} + T_{depth}\} \tag{1}$$

T_{depth} can be measured from a reference user's hand, but we experimentally noted that an empirical threshold of $T_{depth} = 10$ cm is acceptable in most cases (we used this value for the experimental results). In order to remove also most of the retained arm samples, we perform a further check on $\mathcal{H}$, namely we remove each $\mathbf{X}_{u,v} \in \mathcal{H}$ that has a distance in the 3D space from $\mathbf{X}_{\min}$ larger than a threshold T_{size} that also depends on the hand size (for the experiments we set $T_{size} = 20$ cm). Note how T_{depth} and T_{size} only depend on the physical hand size and not on the hand position or the sensor resolution.

[1] A reference hand or alternatively a standard face detector Viola and Jones (2001) can be used to extract a sample skin region.