Pattern Recognition Letters 39 (2014) 65–73

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Online gesture recognition from pose kernel learning and decision forests

Leandro Miranda [a], Thales Vieira [a], Dimas Martínez [a], Thomas Lewiner [b,*], Antonio W. Vieira [c,d], Mario F. M. Campos [c]

[a] Institute of Mathematics, UFAL, Maceió, Brazil
[b] Department of Mathematics, PUC–Rio, Rio de Janeiro, Brazil
[c] Department of Computer Science, UFMG, Belo Horizonte, Brazil
[d] Department of Mathematics, UNIMONTES, Montes Claros, Brazil

ARTICLE INFO

Article history:
Available online 16 October 2013

Communicated by Carla Dal Sasso Freitas

Keywords:
Online gesture recognition
Key pose identification
Skeleton representation
Depth sensors
3D motion
Natural user interface

ABSTRACT

The recent popularization of real time depth sensors has diversified the potential applications of online gesture recognition to end-user natural user interface (NUI). This requires significant robustness of the gesture recognition to cope with the noisy data from the popular depth sensor, while the quality of the final NUI heavily depends on the recognition execution speed. This work introduces a method for real-time gesture recognition from a noisy skeleton stream, such as those extracted from Kinect depth sensors. Each pose is described using an angular representation of the skeleton joints. Those descriptors serve to identify key poses through a Support Vector Machine multi-class classifier, with a tailored pose kernel. The gesture is labeled on-the-fly from the key pose sequence with a decision forest, which naturally performs the gesture time control/warping and avoids the requirement for an initial or neutral pose. The proposed method runs in real time and its robustness is evaluated in several experiments.

1. Introduction

Human gesture recognition is an active topic of research and covering a wide range of applications including originally monitoring, control and analysis. More recently, several applications use real-time (online) gesture recognition to control entertainment devices such as game consoles, virtual reality setups, motion capture to graphics model animation, or automatic control of domestic utilities.

The variety of potential applications have intensified the efforts to improve the automatic recognition of human gestures. The evolution and popularization of depth sensors, which currently generate depth maps in real time as well as their skeletons, is paving the way for the development of high quality natural user interface (NUI) beyond their use as game consoles. Improving the quality of a NUI essentially means to increase the execution speed of the gesture identification and its robustness, in particular with noisy data such as skeletons extracted from Kinect sensors, and this is the objective of the present work.

The gesture recognition problem can be stated as the process of automatically labeling gestures performed by a person based on sensory data, usually captured as sequences of positions in space. This is a particularly challenging task, specially considering that different users perform gestures with different speeds and distinct sequences of poses. In this work, we propose a gesture recognition method from captured skeletons in real time that tackles the aforementioned issues. More specifically, all our experiments are performed using the popular Kinect platform, a real-time depth sensing system that parses a depth-map stream at 30 frames per second, from which positions of the skeleton nodes for each frame can be estimated in real time (Shotton et al., 2011).

Contributions: A human gesture can be formally described as the continuous evolution of body poses over time. Interestingly, we verbally describe such gestures by sequentially identifying a few extreme poses, referred to as *key poses* (Lv and Nevatia, 2007), as illustrated in Fig. 1. In this case, we recognize a gesture by extracting those key poses, classifying them, and then identifying sequences of key poses as a gesture. Following this observation, we focus here on improving and tailoring the three main ingredients of key pose gesture recognition: pose descriptor, pose identification, and labeling of pose sequences.

Our pose descriptor relies on spherical angular representations of joints, similarly to the recent work of Raptis et al. (2011). However, our method is more robust for usual gestures, and it allows

* Corresponding author. Tel.: +55 21 3527 1747; fax: +55 21 3527 1282.
E-mail addresses: leandrobotelhoalves@gmail.com (L. Miranda), thomas@lewiner.org (T. Lewiner).
URLs: http://www.im.ufal.br/professor/thales/ (T. Vieira), http://www.im.ufal.br/professor/dimas/ (D. Martínez), http://thomas.lewiner.org/ (T. Lewiner), http://homepages.dcc.ufmg.br/~awilson/ (A.W. Vieira), http://www.verlab.dcc.ufmg.br/ (M.F. M. Campos).

0167-8655/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.patrec.2013.10.005

Fig. 1. Gesture representation from key poses: our method represents a body gesture as a sequence of a few extreme body poses, referred to as *key poses*. In the example above, a gesture composed by opening arms and then clapping hands is represented by the key poses in black.

for real-time pose classification. In particular, it improves the representation of secondary joints (arms, hands, legs and feet) to better suit NUI applications.

The pose identification process combines several Support Vector Machine (SVM) classifiers (Vapnik, 2000), one per reference key pose. We propose a *pose kernel* that entails the angular nature of our representation, and use that pose kernel distance in feature space as the confidence measure. This pose kernel significantly improves the robustness of the method over a kernel based on Euclidean distances (Miranda et al., 2012b), as seen in the results section. Moreover, the small pose descriptor size allows for online training and recognition.

Finally, we propose a scheme for gesture recognition based on decision forests. Each forest node is a key pose, eventually including time constraints, and the leaves are the gesture labels. This decision forest is learned during the training phase. Each tree is rooted at a key pose, and a leaf-to-root path represents a possible sequence of key poses of that leaf gesture. At each identification of a new key pose, the tree rooted at that pose is used to check if it completes a gesture. This allows for real time gesture recognition without the need for a neutral/initial pose. Moreover, the decision state machine produces a natural and robust time warping. The whole process is robust even with noisy depth-based skeletons (Shotton et al., 2011) as shown in the results section.

This paper presents an extension of a previous work (Miranda et al., 2012b). While in that paper the pose kernel relies on Euclidean distances, in the present work we propose a different metric more suitable to our problem. Experiments with this new kernel show significative improvements on the recognition rate. Moreover, we adopted a combination of an out-of-sample approach with a gradient descent search to accurately calibrate kernel parameters. Finally, time constrained gestures are further evaluated through new experiments.

2. Related work

Human gesture recognition has been extensively studied, and a large body of literature has been produced for application in areas such as surveillance, home monitoring and entertainment. We summarize here the most related work for gesture recognition according to the spatial representation used: local, global or parametric.

In the local category, the methods typically use point-wise descriptors evaluated at some points of interest, and then use a bag of features (BoF) strategy to represent actions. This approach has attracted much attention in the past few years (Sun et al., 2009; Cao et al., 2010; Kovashka and Grauman, 2010; Niebles et al., 2010), and an example of largely used local feature is the Space–Time Interest Point (STIP) presented by Laptev and Lindeberg (2003). A drawback of local features approaches is that they lose spatial context information between interest points.

The methods in the global category use features such as silhouettes (Lv and Nevatia, 2007; Weinland and Boyer, 2005; Li et al., 2008) or template based representations (Chen et al., 2007; Bobick and Davis, 2001), where spatial context information is preserved. However, global features usually miss some precise details of the pose such as body joints identification.

Finally, parametric methods try to reconstruct a model of the human body with identification of joints to obtain an skeleton. Skeletons can be obtained by strategies such as the Motion Capture (MoCap) models (e.g., public databases available at `http://mocap.cs.cmu.edu`, `http://www.moves.com`, and `http://www.mpi-inf.mpg.de/resources/HDM05/`), where a set of markers is attached to the human body at specific points that are tracked during motion. Using this representation, spatial features are constructed using some geometric measures and relations, e.g., angles (Kovar, 2004; Forbes and Fiu, 2005) and body joints positions (Müller and Röder, 2006; Müller et al., 2009). In particular, Vieira et al. (2012) show that the matrix of distances between body joints completely describe a pose up to rigid movements, and such matrices serve to define low-dimensional invariant features for classifying actions.

MoCap skeletons strongly depend on rather sophisticate capture systems. A more accessible way to generate skeletons is proposed by Shotton et al. (2011), who obtain skeletons without markers by computing joint coordinates in real time from depth maps. In particular, such skeletons can be obtained from the popular *Kinect* sensor. Compared to MoCap data, skeletons from Kinect are easier to obtain, which have driven the popularity of this sensor. However, they show a high level of noise and spatial discontinuity, turning gesture recognition from depth data a sizable challenge, and that is the focus of the present work.

Gesture recognition using skeletons from Kinect has received a lot of attention recently. In particular, Li et al. (2010) published the MSR Action3D dataset, a database with depth maps sequences and their respective skeletons, composed of 20 different short and non-repetitive action classes, each performed by several subjects (Liu, 2011). They distinguish three different subsets and present their classification rate obtained for each test. This dataset became a benchmark for several recent works (Vieira et al., 2012b; Yang and Tian, 2012; Yang et al., 2012). We also use this dataset to validate our approach for action classification and present comparative results with other state-of-the-art works in the literature.

Reyes et al. (2011) obtain 3D coordinates of skeletal models using a reference coordinate system to make the description view point invariant and tolerant to corporal differences among subjects. Results are reported for only five different categories. Raptis et al. (2011) introduce a method for classifying dance gestures using Kinect skeletons where a large number of gesture classes are used. Their pose descriptor uses spherical coordinates in a frame obtained by Principal Component Analysis on a subset of the body points that define the torso. In spite of the high recognition rate reported, their classification method is limited to dance gestures and is conceived based on the strong assumption that the input motion adheres to a known music beat pattern. We extend their pose descriptor to a completely invariant angular representation. Vieira et al. (2012b) propose a global feature, called Space–Time Occupancy Patterns for action recognition from depth map sequences where space and time axes are used to define a 4D grid. A saturated histogram of point count in the grid cells is used as features for action classification. Yang and Tian (2012) propose a new type of feature based on position differences of joints which combines action information including static posture, motion, and offset. They employ the Naïve-Bayes-Nearest-Neighbor classifier for multi-class action classification and also explore the number of frames that are needed to classify the actions. Yang et al. (2012) project depth maps onto three orthogonal planes and accumulate global activities across entire video sequences to generate the Depth Motion Maps (DMM). Histograms of Oriented Gradients (HOG) are then computed from DMM as the representation of an action video. Note that the pose descriptor used in the works of Li et al. (2010), Vieira et al. (2012b), Yang and Tian (2012), Yang et al. (2012) are sensitive to the orientation of the capture device.