Pattern Recognition 55 (2016) 137–147

Contents lists available at ScienceDirect

Pattern Recognition

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

# An image-to-class dynamic time warping approach for both 3D static and trajectory hand gesture recognition ☆

Hong Cheng [a,*], Zhongjun Dai [a], Zicheng Liu [b], Yang Zhao [a]

[a] Center for Robotics, University of Electronic Science and Technology of China, Chengdu 611731, China
[b] Microsoft Research Redmond, One Microsoft Way, Redmond, WA 98052, USA

ARTICLE INFO

*Article history:*
Received 21 July 2015
Received in revised form
7 January 2016
Accepted 13 January 2016
Available online 3 February 2016

*Keywords:*
Image-to-class distance
Fingerlets
Strokelets
Dynamic time warping
3D hand gesture recognition
Human computer interaction

ABSTRACT

In this paper, we present an Image-to-Class Dynamic Time Warping (I2C-DTW) approach for the recognition of both 3D static hand gestures and 3D hand trajectory gestures. Our contribution is twofold. First, we propose a technique to compute the image-to-class dynamic time warping distance instead of the Image-to-Image distance. By doing so, we obtain better generalization capability using the Image-to-Class distance than the Image-to-Image distance. Second, we propose a compositional model called fingerlets for static gesture representation, and a compositional model called strokelets for trajectory gesture representation. The compositional models make it possible to compute the DTW distance between a data sample and a gesture category. We have evaluated the static gesture recognition performance on several public 3D hand gesture datasets. For better evaluating the performance on trajectory gesture recognition, we collected a 3D hand trajectory gesture dataset, called UESTC-HTG, using a Kinect device. The experiment results show that the proposed I2C-DTW approach significantly improves the recognition accuracy on both static gestures and trajectory gestures.



## 1. Introduction

The use of hand gestures has been around since the beginning of time and is much older than speech. Hand gestures are natural and ubiquitous ways to interact among different people [16]. Inspired by human interaction mainly by vision and sound, the use of hand gestures is one of the most powerful and efficient ways in Human Computer Interaction (HCI) [13,12,26].

There are three basic types of sensors which are capable of sensing hand gestures: mount based sensors (Microsofts Digits hand-gesture sensor bracelet, IR gesture-sensing systems, 5DT data glove), touch based sensors (iPhone, iPad), and vision based sensors [13]. Vision based sensors have the following advantages. First, they are less cumbersome and more comfortable than the other two sensors thanks to no physical contact with users. Second, vision based sensors allow a user to interact with a computer at a distance. Finally, vision based sensors can be robust for gesture recognition in noisy environments where sound based sensors would be in trouble. In particular, the emergence of the Kinect devices has drawn much attention on 3D vision based gesture recognition [30,12,35].

Vision based hand gesture recognition approaches in general consist of three modules including detection, tracking and recognition. The detection module defines and extracts visual features which encode hand gestures in the camera's field of view. The tracking module is the process of locating the hand over time. Note that tracking is not a mandatory step in the gesture recognition pipeline. For example, some approaches may just consider the current frame in the static gestures recognition. The recognition module clusters the spatiotemporal features generated from the first two modules into groups and labels these groups with gesture classes. Diverse hand gesture approaches have been proposed in the past decade with different types of visual features and classifiers.

Dynamic Time Warping (DTW) is an efficient classification approach and much research has been done to improve its speed and accuracy. However, this approach has two serious limitations. (1) A traditional dynamic time warping approach works in an image-to-image manner to compute the distances between samples. Thus Image-to-Image Dynamic Time Warping (I2I-DTW) works well only when the testing sample is holistically similar to one of the training samples. This limits its generalization capability beyond the training samples. (2) Dynamic time warping approaches usually model objects as holistic time-series curves. As we know, people not only use global ways to perceive patterns but

☆The preliminary version of this paper has been previously published in ICME 2013 [8].

* Corresponding author. Tel.: +8602861830797.
*E-mail addresses:* hcheng@uestc.edu.cn (H. Cheng), diana.d6523@gmail.com (Z. Dai), zliu@microsoft.com (Z. Liu), zhaoyang1025@gmail.com (Y. Zhao).

http://dx.doi.org/10.1016/j.patcog.2016.01.011
0031-3203/

also analyze the patterns in detail. Thus the global features limit its flexibility.

In this paper, we propose an Image-to-Class Dynamic Time Warping (I2C-DTW) approach for both 3D static hand gesture and 3D hand trajectory gesture recognition. There are two main contributions. First, instead of finding an optimal image-to-image match, the proposed approach first searches for the minimal warping path between a test sample and a training sample's compositional features. The gesture recognition is done by using the ensemble of multiple image-to-class DTW distances each corresponding to a compositional feature. Second, we propose compositional models for static and trajectory gesture representation. For static hand gesture, we divide the time-series curve of a gesture into various finger combinations, called fingerlets. For hand trajectory gesture, we divide a hand trajectory curve into a set of atomic strokes, called strokelets. The compositional models allow us to perform partial matching thus improving the generalization capability. To evaluate the performance on 3D trajectory gesture recognition, we have collected a hand trajectory gesture dataset, called UESTC-HTG, using a Kinect device. It has 16 trajectory gesture classes performed by 100 people with 1600 samples in total. The proposed approach is evaluated on two public 3D static hand gesture datasets and the UESTC-HTG trajectory gesture dataset. The experiment results show that the proposed I2C-DTW approach significantly improves the recognition performance.

The rest of this paper is organized as follows. Section 2 reviews the related work. We introduce the proposed image-to-class dynamic time warping approach in Section 3. Sections 4 and 5 present the application of the proposed I2C-DTW on static gesture recognition and trajectory gesture recognition, respectively. Section 6 presents the experimental results and analysis. Section 7 concludes this paper and the future work.

## 2. Related work

In this section, we first review the state-of-the-art dynamic time warping approaches. Following that, we review gesture representation, modeling, and classification approaches for both static gestures and hand trajectory gestures. We will focus on 3D gesture understanding.

Thanks to its small training data requirement and high accuracy, a dynamic time warping approach has been widely used in both sequential and non-sequential models such as human activity recognition [36,37], hand gesture recognition [2,9], hand-writing recognition [31], and face recognition [33]. For a survey on dynamic time warping, the readers are referred to [29]. After DTW was widely adopted in various pattern recognition tasks, a lot of algorithms emerged to reduce the computational burden and improve the classification performance. Many researchers use global constraints to reduce the computational complexity, such as Sakoe–Chiba band [28], Itakura Parallelogram [14] and R-K band [24]. The R-K band learns constraints to make time-series classification more accurate. The new framework learns arbitrary constraints on the warping path of DTW computation and speeds up DTW by a wide margin. Another aspect to improve a classic approach is DTW transformation. Derivative Dynamic Time Warping (DDTW) [17] is one of them, which is designed to deal with singularities and missing natural alignments of DTW. Xie et al. introduced a Feature based Dynamic Time Warping (FDTW) [34] approach to align two sequences based on each points' local and global feature. Concurrently, Dynamic Image-to-Class Warping (DICW) [33] is proposed to deal with partially occluded face recognition. As we know, DTW cannot be directly used in multi-model sequences (e.g., videos and motion capture data) since there is no fusion strategy in the traditional DTW. To address this issue, Zhou et al. proposed a Canonical Time Warping (CTW), which combines DTW with Canonical Correlation Analysis (CCA) [37]. More generally, Zhou et al. proposed Generalized CTW to overcome three main limitations for efficient spatio-temporal alignment of multiple time series [36].

Recognition of static hand gestures involves many aspects such as hand detection, tracking, hand contour representation, and gesture classification [22,25,38]. Static hand gesture recognition can be classified into three basic categories: 2D, 2.5D and 3D approaches. Overviews of 2D static hand gesture recognition can be found in [22]. Thanks to the emergence of commercial 2.5D depth sensors, depth information can be easily used in gesture recognition. It is both used in the hand detection and segmentation phase [10,11], and also in the feature extraction [26,8,4,23,21]. Ren et al. proposed a novel distance metric for hand dissimilarity measure, called Finger-Earth Movers Distance (FEMD), to handle the noisy hand data captured by a Kinect sensor [26]. Bonansea proposed a complete solution for the one-hand 3D gesture recognition problem [4].

Hand trajectory gestures may involve body parts like waving a hand or greeting. Alon et al. proposed a method for simultaneous localization and recognition of dynamic hand gestures. The core is a Dynamic Space Time Warping (DSTW) algorithm [1], which aligns a pair of query and model gestures in both space and time. It performs better than classic DTW in hand-signed digit recognition. However, users have to stand in front of a special background without hand-skin-colored objects. Reyes et al. presented Feature Weighting Dynamic Time Warping (FWDTW) [27] which calculates feature weights based on inter–intra-class gesture variability. They described human joints as feature vectors, assign weights to features, and use them to recognize the beginning and end of gestures. Guo proposed a 3D hand hierarchy model for 3D hand articulation tracking [13]. The hand gesture is extracted first as static postures and the particle swarm optimization was used to solve for the articulation parameters.

## 3. The proposed image-to-class dynamic time warping approach

### 3.1. Traditional image-to-image dynamic time warping approach

The DTW distance is an extremely efficient technique that allows a non-linear mapping of one feature to another by minimizing the distance between those two features which is superior to Euclidean distance.

More specifically, the DTW approach deriving from dynamic programming aims to formulate a cost matrix and find the optimal path with certain constraints between two time-series curves. For better addressing the proposed I2C-DTW, we review the traditional dynamic time warping approach in this section and follow the notation described in [29,27].

Mathematically, given two features $\boldsymbol{f}_a = \{a_1, a_2, \ldots, a_m\} \in R^m$ and $\boldsymbol{f}_b = \{b_1, b_2, \ldots, b_n\} \in R^n$, we calculate a $m \times n$ cost matrix $C = [c(a_i, b_j)]$ as

$$c(a_i, b_j) = \| b_j - a_i \|_p, \tag{1}$$

where the Euclidean distance is used when $p=2$.

The total cost $c_p(\boldsymbol{f}_a, \boldsymbol{f}_b)$ of a warping path $p$ between $\boldsymbol{f}_a$ and $\boldsymbol{f}_b$ with respect to the local cost measure $c(\cdot)$ is defined as

$$c_p(\boldsymbol{f}_a, \boldsymbol{f}_b) = \sum_{l=1}^{L} c(a_{i,l}, b_{j,l}), \tag{2}$$

where a $(n,m)$-warping path $p = (p_1, \ldots, p_L)$ defines an alignment between $\boldsymbol{f}_a$ and $\boldsymbol{f}_b$ by assigning the elements $a_{i,l}$ of $\boldsymbol{f}_a$ to the