Pattern Recognition 48 (2015) 1302–1314

Contents lists available at ScienceDirect

Pattern Recognition

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

Conditional distance based matching for one-shot gesture recognition

Ravikiran Krishnan *, Sudeep Sarkar

Department of Computer Science, University of South Florida 4202 E. Fowler Ave, Tampa, FL 33613, United States

ARTICLE INFO

Article history:
Received 13 December 2013
Received in revised form
10 July 2014
Accepted 20 October 2014
Available online 28 October 2014

Keywords:
Warp vector
Gesture recognition
Conditional distance
One-shot
Distance measure
Level building

ABSTRACT

A problem of matching gestures, where there are one or few samples per class, is considered in this paper. The proposed approach shows that much better results are achieved if the distance between the pattern of frame-wise distances of two gesture sequences with a third (anchor) sequence from the modelbase is considered. Such a measure is called as conditional distance and these distance pattern are referred to as "warp vectors". If these warp vectors are similar, then so are the sequences; if not, they are dissimilar. At the algorithmic core, there are two dynamic time warping processes, one to compute the warp vectors with the anchor sequences and the other to compare these warp vectors. In order to reduce the complexity a speedup strategy is proposed by pre-selecting "good" anchor sequences. Conditional distance is used for individual and sentence level gesture matching. Both single and multiple subject datasets are used. Experiments show improved performance above 82% spanning 179 classes.

1. Introduction

Analyzing and recognizing human gestures is important for human computer interaction. The large number of human gesture categories such as sign language, traffic signals, everyday actions and also subtle cultural variations in gesture classes makes gesture recognition an interesting and challenging problem. In most naturally occurring scenarios, gestures are connected together in continuous varying stream, without any obvious break between individual gestures [21]. Identifying each one of these individual gestures gives a good representation for ultimately translating visual communication into speech or other form of interaction. Such labeling tasks have many challenges.

Labeling theses continuous gesture stream or query involves matching temporally segmented individual gestures to a model-base. If the modelbase has many samples per class, statistics of that particular class can be learnt. Having one or few samples deters any class statistic learning approaches to classification, as the full range of variation is not covered. One of the key components of a matching algorithm, apart from feature extraction, would be gesture to gesture distances. These distances should define, in a concrete way, what it means for data points of such a class space to be "near to" or "far away from"[1] each other. One commonly used approach would be to take pair-wise distances (using a distance function) between all available models with the query and discern which are closer (classified as same) or far away from (classified as not same). A commonly used distance function would be dynamic time warping with 1-Nearest Neighbor approach for classification [12].

In our work, a matching algorithm based on a level building approach is proposed. This algorithm is based on a framework of one-shot learning (single sample per class). The proposed algorithm is capable of handling – (a) Isolated and continuous gesture queries; (b) eliminates the need for temporal segmentation; and (c) single sample per class scenarios. A new distance function is defined and this serves as the center of our algorithm. Each gesture sequence is seen as a curve and each curve as a data point on a space that is formed by all the gesture classes. Pair-wise distance pattern between two gesture sequences conditioned on a third (anchor) sequence is considered. These distance pattern vectors are called as "warp vectors". And such a process is called as "conditional distance". At the algorithmic core, there are two dynamic time warping processes, one to compute the warp vectors with the anchor sequence and the other to compare these warp vectors. Such measures have been proposed earlier, for example, Mahalanobis distance, where intra-class variations, or the variation between the instances, accounts for the scaling that the distance measure needs. Our work explores the variations between the classes itself as the modelbase is framed as a single sample per class.

Given a situation where the model base is large (number of classes is also large); the disadvantage of such a distance would be the computational cost. There is a need for pre-selecting the anchor gesture (or class). A speedup strategy is proposed based on pre-selection of anchor gestures from the modelbase. The

* Corresponding author. Tel.: +1 813 367 7401.
E-mail addresses: rkrishn2@cse.usf.edu (R. Krishnan), sarkar@cse.usf.edu (S. Sarkar).

[1] http://en.wikipedia.org/wiki/Distance

http://dx.doi.org/10.1016/j.patcog.2014.10.026
0031-3203/

proposed distance is computed with every gesture to every other gesture in the modelbase. For each such distance, an anchor gesture is determined. Majority anchor gesture is selected and distances between query and model is computed only on this chosen anchor gesture.

Conditional distance gives the distance between two isolated gestures. In order to label multiple connected gestures, a simultaneous segmentation and recognition matching algorithm called level building algorithm is used. Dynamic programming implementation of the level building algorithm is employed. The core of this algorithm depends on a distance function that compares two gesture sequences. This distance function is replaced by conditional distance. Hence, this version of level building is called as conditional level building (clb).

Earlier version of this work was proposed in [19]. In this paper, a more detailed use, analysis and results for conditional distance is provided. The main differences are listed as follows:

1. This paper shows that conditional distance increases performance over the baseline distances in two gesture modelbase contexts – single category-single subject and multiple category-multiple subjects and shows that conditional distance satisfies metric properties in practice.
2. An anchor pre-selection strategy to speed up computation of the proposed distance is proposed. Anchor behavior and selection with and without the proposed strategy is analyzed.
3. A conditional distance version of level building algorithm for recognizing connected gestures is also proposed.
4. Results are shown on gesture challenge datasets [1] (Fig. 1) and compared our results with state-of-the-art on those datasets.

Conventions: A time sequence or a gesture sequence is an array of images taken at certain times. The sampling rate is same as the regular video sampling rate. Gesture sequence can have a length n and are indexed from 1 to n. The l_2 distance between feature vectors x and y is $\|x-y\|_2$ and it satisfies the triangle inequality $\|x-z\|_2 \leq \|x-y\|_2+\|y-z\|_2$. A summary of frequently used conventions is provided in Table 1.

2. Related work

Any gesture recognition task, might that be a series of gestures connected to form a single query or a single gesture query, involves comparing an incoming query against a training set of gestures. A collection of all the instances of all the classes available for training is referred to as a modelbase. These modelbases can have many instances per gesture class or they might have just one instance per class. If there are many instances then the recognition can be based on learning statistics of class features from the instances of the modelbase with methods such as Hidden Markov Model and its variants [9,29], Finite State Machines [15], Dynamic Bayesian Networks (DBN) [35], topology-preserving self-organized networks [11], Conditional Random Fields (CRF) [33] and other methods [41]. Table 2 gives the summary of matching techniques, features used and samples per class used by various state-of-the-art techniques. HMMs are widely accepted as the choice of recognition method. This method has been used to solve both single gesture as well as connected gestures. Dynamic Bayesian Networks (DBN) are however a more general form of probabilistic models which combine HMMs and Kalman filters and is a generalized version of Bayesian Networks. Such approaches have their share of problems, such as requiring large amounts of data to cover all variations of gesture classes or less of such leading to over-fitting. The number of samples required in all of these methods varies based on factors such as the input space, category of gestures and the learning method chosen. These methods have been summarized in Table 2.

In this paper, for related work, the focus is more on one-shot or single sample per class methods. There has been increasing interest in computer vision to avoid problems such as collecting and labeling large amounts of data, in a one-shot learning approach for gesture recognition [1,23,7,25]. While the term "one-shot" learning has been loosely used in the literature as one or few training instances, our work refers to one-shot as only one instance per class. A recognition problem in such a context is considered. In [25], different combination of features are used to get the best result on one-shot learning gesture dataset [1]. HOG, HOF and motion signatures are some of the common features used in one-shot based approaches. In [39], a bag of features approach is used with the low level features being the 3D motion sift where both RGB and Depth are combined. In [26], HOG/HOF features are used with HMM for matching. One of the state-of-the-art approach also uses sub-sampled depth images and use HOG features on it. Even though these methods provide different matching technique, it is important to mention that DTW remains [12], by the far the most used technique for matching for one-shot gesture recognition. Methods in [24,17,10] also propose new one-shot distance measures for images but none of them use these techniques for gesture classification.

Dynamic time warping is commonly used as a distance measure when comparing two gesture sequences. DTW as a measure, was first proposed in [30]. Efficient methods have been proposed to compute this measure in [16,20,31,2,40]. This distance measure

Fig. 1. Examples of common gestures include sign language signals, traffic signals, hand signals and body language from ChaLearn Gesture Challenge Dataset [1]. The RGB images representing gesture sequences are not used in our experiments. Only depth gesture sequences corresponding to such RGB sequences are used.