Signal Processing 104 (2014) 248–257

Contents lists available at ScienceDirect

Signal Processing

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

Energy-based model of least squares twin Support Vector Machines for human action recognition

Jalal A. Nasiri, Nasrollah Moghadam Charkari*, Kourosh Mozafari

Electrical and Computer Engineering, Tarbiat Modares University, Tehran, Iran

ARTICLE INFO

Article history:
Received 28 October 2013
Received in revised form
19 February 2014
Accepted 9 April 2014
Available online 15 April 2014

Keywords:
Twin Support Vector Machine
Least square method
Multi-class classification
Spatio-temporal feature
Human action recognition

ABSTRACT

Human action recognition is an active field of research in pattern recognition and computer vision. For this purpose, several approaches based on bag-of-word features and support vector machine (SVM) classifiers have been proposed. Multi-category classifications of human actions are usually performed by solving many one-versus-rest binary SVM classification tasks. However, it leads to the class imbalance problem. Furthermore, because of environmental problems and intrinsic noise of spatio-temporal features, videos of similar actions may suffer from huge intra-class variations. In this paper, we address these problems by introducing the Energy-based Least Square Twin Support Vector Machine (ELS-TSVM) algorithm. ELS-TSVM is an extended LS-TSVM classifier that performs classification by using two nonparallel hyperplanes instead of a single hyperplane, as used in the conventional SVM. ELS-TSVM not only could consider the different energy for each class but also it handles unbalanced datasets' problem. We investigate the performance of the proposed methods on Weizmann, KTH, Hollywood, and ten UCI datasets which have been extensively studied by research groups. Experimental results show the effectiveness and validity of noise handling in human action and UCI datasets. ELS-TSVM has also obtained superior accuracy compared with the related methods while its time complexity is remarkably lower than SVM.

1. Introduction

Human action recognition is one of the important research areas in computer vision and pattern recognition. It has a wide range of applications such as surveillance systems, human computer interaction, video retrieval, and gesture recognition. In the past decade, with growing in video quality and personal video recording, the need to automatic video analysis and the recognition of events has bean increased. The difficulty of human action recognition problems may have been originated from several challenges such as illumination changes, partial occlusions, and intra-class differences [1].

Recently, Bag of Words (BoWs) representation and support vector machine (SVM) for human action recognition have attracted much interest [2–4]. Accordingly, the feature descriptors are extracted from all the training sequences to build a codebook by clustering similar features. The cluster centroids, called as video words, are the members of this codebook. Each feature descriptor is assigned to a certain video word (cluster centroid). An action video is represented as a histogram of the number of occurrences of particular video words. Then, classification methods are exploited to build models for each action class.

The support vector machine was originally proposed by Cortes and Vapnik [5] for the purpose of binary classification. SVM has been successfully applied in a wide spectrum

* Corresponding author. Tel.: +98 21 82883301.
E-mail addresses: j.nasiri@modares.ac.ir (J.A. Nasiri), charkari@modares.ac.ir (N. Moghadam Charkari).

http://dx.doi.org/10.1016/j.sigpro.2014.04.010
0165-1684/

of research areas like face recognition, object categorization, and biomedicine [6–9]. The computational complexity of SVM is $O(l^3)$, where l denotes the total size of training data. However, this drawback restricts the application of SVM to large-scale problem domains. Since the optimal hyperplane obtained by SVM depends on only a small part of samples (support vectors), it is very sensitive to the outliers and noisy samples. Moreover, multi-category classification of human actions is usually done by solving many one-versus-rest binary SVM classification tasks. Each binary SVM is trained with all of the patterns, so it easily leads to the class imbalance problem.

To deal with these issues, we propose a fast classifier to understand activity recognition based on Twin Support Vector Machines (TSVM). TSVM were proposed by Jayadeva et al. in [10] for binary classification. This method generates two nonparallel hyperplanes by solving two smaller-sized Quadratic Programming Problems (QPPs) such that each hyperplane is closer to one class and as far as possible from the other. The idea of solving two smaller-sized QPPs rather than a single larger-sized QPP in SVM makes the learning of TSVM four times faster than the conventional SVM [10]. Least Squares Twin Support Vector Machine (LS-TSVM) [11] is an extension of TSVM as a way to replace the convex QPPs in TSVM with a convex linear system by using a squared loss function instead of the hinge one. This formulation leads to the extremely simple and fast algorithm. The constraints of the LS-TSVM are converted to an energy model which could reduce the adverse effects of noisy data and outliers. In addition, in one-versus-rest protocol of ELS-TSVM for multi-class classification, imbalance datasets will not affect the model learning.

The paper is organized as follows: we first review the related work in Section 2. In Section 3, we describe the proposed human action recognition framework and introduce the ELS-TSVM. In Section 4, the experimental results on common datasets are given. Finally, Section 5 contains concluding remarks.

2. Related work and background

A comprehensive review of the human action recognition approaches can be found in some interesting survey papers such as [1,12–14]. In general, feature representations of video sequences can be divided into two categories: top-down (global) [15,16], and bottom-up (local) [17–19] strategy representations. The global strategy first localizes region of the person in the video by background subtraction, and then represents the interest region as a whole. In this way, global strategies are sensitive to noise, variations in viewpoint and partial occlusion. However, the local strategy, which is based on the spatio-temporal interest points, has a better adoption to the environment.

Recently, local representation methods have become popular and some new interest point detectors and descriptors have been proposed. One of the earliest works in space–time local features were discussed in [20]. Accordingly, the previous Harris local feature detector was extended to employ the notions of interest points into the spatio-temporal domain. Then, the detector finds the spatio-temporal corners in a 3-D space. The proposed detector is able to capture various types of motion patterns. In [21] a spatio-temporal feature detector was presented. The detector extracts space–time points with local periodic motions. At each interest point, a cuboid is extracted which contains the spatio-temporally windowed pixel values. Classic SIFT descriptor was extended in [22] and a 3-D version of SIFT descriptor was developed similar to the cuboid features. The study in [18] introduced histogram descriptors of space–time volumes in the neighborhood of detected points to describe motion and appearance of local features (Histograms of Oriented Gradient (HOG) and Optical Flow (HOF)). They divided each space–time volume into grids of cuboids and computed the histograms for these cuboids. Wang et al. [23] introduced motion boundary histograms (MBHs) to eliminate noise caused by the background motion. Although local histograms have been widely employed in the fields of object recognition and action recognition. They have some disadvantage of hard binning.

In the classification phase, many methods have been applied in the field of human action recognition, including the (k-NN) [19], support vector machine (SVM) [17,18, 22,23], boosting-based classifiers [24], Hidden Markov Models [15,16]. The recent research has witnessed the evolution of support vector machines as a powerful paradigm for pattern classification. But in the application of human action recognition, SVM has some drawbacks such as high computational time, sensitivity to outliers and the unbalanced dataset problem. In this paper, we propose a TSVM based classifier for human action recognition to deal with the above issues.

Twin Support Vector Machine: TSVM is a binary classifier that performs classification by the use of two non-parallel hyperplanes unlike SVM uses a single hyperplane [10]. Consider the dataset D that is divided into d^+, as the training set with label $+1$, and d^-, as the training set with label -1, in the m-dimensional real space R^m. Let matrix $A \in R^{d^+ \times m}$ represent the training data which belongs to the class $+1$ and matrix $B \in R^{d^- \times m}$ represent the training data which belongs to the class -1. The linear TSVM searches for two non-parallel hyperplanes in R^m as follows:

$$x^T w_{(1)} + b_{(1)} = 0 \quad \text{and} \quad x^T w_{(2)} + b_{(2)} = 0 \tag{1}$$

Such that each hyperplane is closest to the training data of one class and farthest from the training data of another class. A new data sample is assigned to class $+1$ or -1, depending on which of the two hyperplanes lies closest to the point. The linear TSVM solves two quadratic programming problems (QPPs) (2) and (3) with objective function corresponding to one class and constraints corresponding to the other class

$$\begin{aligned} \min_{w_{(1)},b_{(1)}} \quad & \tfrac{1}{2}\|Aw_{(1)} + e_1 b_{(1)}\|^2 + c_1 e_2^T y_2 \\ \text{s.t.} \quad & -(Bw_{(1)} + e_2 b_{(1)}) + y_2 \geq e_2, \quad y_2 \geq 0 \end{aligned} \tag{2}$$

and

$$\begin{aligned} \min_{w_{(2)},b_{(2)}} \quad & \tfrac{1}{2}\|Bw_{(2)} + e_2 b_{(2)}\|^2 + c_2 e_1^T y_1 \\ \text{s.t.} \quad & (Aw_{(2)} + e_2 b_{(2)}) + y_1 \geq e_1, \quad y_1 \geq 0, \end{aligned} \tag{3}$$