Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Discriminant Bag of Words based representation for human action recognition ☆

Alexandros Iosifidis *, Anastastios Tefas, Ioannis Pitas

Department of Informatics, Aristotle University of Thessaloniki, Thessaloniki 54124, Greece

ARTICLE INFO

Article history:
Received 6 March 2014
Available online 30 July 2014

Keywords:
Bag of Words
Discriminant learning
Codebook learning

ABSTRACT

In this paper we propose a novel framework for human action recognition based on Bag of Words (BoWs) action representation, that unifies discriminative codebook generation and discriminant subspace learning. The proposed framework is able to, naturally, incorporate several (linear or non-linear) discrimination criteria for discriminant BoWs-based action representation. An iterative optimization scheme is proposed for sequential discriminant BoWs-based action representation and codebook adaptation based on action discrimination in a reduced dimensionality feature space where action classes are better discriminated. Experiments on five publicly available data sets aiming at different application scenarios demonstrate that the proposed unified approach increases the codebook discriminative ability providing enhanced action classification performance.

1. Introduction

Human action recognition from videos has been intensively studied in the last two decades due to its importance in a wide range of applications, like human–computer interaction (HCI), content-based video retrieval and augmented reality, to name a few. It is, still, an active research field due to its difficulty, which is, mainly, caused because there is not a formal description of actions. Action execution style variations and changes in human body sizes among individuals, as well as different camera observation angles are some of the reasons that lead to high intra-class and, possibly, small inter-class variations of action classes. Recently, several action descriptors aiming at action recognition in unconstrained environments have been proposed, including local sparse and dense space–time features [23,45,10,20,34,35]. Such descriptors capture information appearing in video frame locations that either correspond to video frame interest points which are tracked during action execution, or that are subject to abrupt intensity value variations and, thus, contain information regarding motion speed and/or acceleration, which is of interest for the description of actions. These local video frame descriptors are calculated by using the color (grayscale) video frames and, thus, video frame segmentation is not required.

After describing actions, videos depicting actions, called action videos hereafter, are usually represented by fixed size vectors. Several feature coding approaches have been proposed in order to determine compact (vectorial) representations [14], including Sparse Coding [42], Fisher Vector [28], Local Tangent coding [47] and Salience-based coding [13]. Perhaps the most well studied and successful approach for action representation is based on the Bag of Words (BoWs) model [4]. According to this model, each action video is represented by a vector obtained by applying (hard or soft) quantization on the features describing it and using a set of feature prototypes forming the so-called codebook. This codebook is determined by clustering the features describing training action videos. The BoWs-based action representation has been combined with several classifiers, like Support Vector Machines, Artificial Neural Networks and Discriminant Analysis based classification schemes, providing high action classification performance on publicly available data sets aiming at different application scenarios. However, due to the fact that the calculation of the adopted codebook is based on an unsupervised process, the discriminative ability of the BoWs-based action representation is limited.

In order to increase the quality of the adopted codebook, codebook adaptation processes have been proposed which adopt a generative approach. That is, the initial codebook generated by clustering the features describing training videos is adapted so as to reduce the reconstruction error of the resulted video representation [37]. However, since this generative adaptation process does not take into account the class labels that are available for the training action videos, the discriminative ability of the optimized

☆ This paper has been recommended for acceptance by J. Yang.

* Corresponding author. Tel./fax: +30 2310996304.
E-mail address: aiosif@aiia.csd.auth.gr (A. Iosifidis).

http://dx.doi.org/10.1016/j.patrec.2014.07.011
0167-8655/© 2014 Elsevier B.V. All rights reserved.

codebook is not necessarily increased. In order to increase the discriminative ability of the adopted codebook, researchers have begun to introduce discriminative codebook learning processes [40,29]. However, since the codebook calculation process is, still, disconnected from the adopted classification scheme, the obtained codebook may not be the one that is best suited for the task under consideration, i.e., the classification of actions in our case.

A method aiming at simultaneously learning both a discriminative codebook and a classifier is proposed in [43] for image classification. This method consists of two iteratively repeated steps. The first one involves training images representation by a set of class-specific histograms of visual words at the bit level and multiple binary classifiers, one for each image category, training by using the obtained histograms. Based on the performance of each classifier, the set of training histograms is updated in the second step. While this approach has lead to increased image classification performance, its extension in other classification tasks, e.g., action recognition, is not straightforward. Another approach has been proposed in [25], where a two-class linear SVM-based codebook adaptation scheme is formulated. The adoption of a two class formulation generates the drawback that $C(C-1)/2$ two-class codebooks have to be learned (C being the number of classes) and used in the test phase along with an appropriate fusion strategy. In addition, such an approach is not able to exploit inter-class correlation information appearing in multi-class problems, which may facilitate class discrimination.

In this paper, we build on the BoWs-based action video representation by introducing discriminative criteria on the codebook learning process. The proposed approach integrates codebook learning and action class discrimination in a multi-class optimization process in order to produce a discriminant BoWs-based action video representation. Two processing steps are iteratively repeated to this end. The first one, involves the calculation of BoWs-based representation of the training action videos using a codebook of representative features and learning of an optimal mapping of the obtained BoWs-based action video representations to a discriminant feature space where action classes are better discriminated. In the second step, based on an action class discrimination criterion in the obtained feature space, the adopted codebook is adapted in order to increase action classes discrimination. In order to classify a new, unknown, action video, it is represented by employing the optimized codebook and the obtained BoWs-based action video representation is mapped to the discriminant feature space determined in the training phase. In this discriminant space, classification can be performed by employing several classifiers, like K-Nearest Neighbors (K-NN), Support Vector Machine (SVM) or Artificial Neural Networks (ANN). Here it should be noted that the proposed approach is not aiming at increasing the representation power of the adopted codebook. Instead, it aims at increasing its discrimination power for the classification task under consideration, i.e., the discrimination of the action classes involved in the classification problem at hand.

The rest of the paper is structured as follows. The proposed approach for integrated discriminant codebook and discriminant BoWs-based action representation learning is described in Section 2. Experiments conducted on publicly available data sets aiming at different application scenarios are illustrated in Section 3. Finally, conclusions are drawn in Section 4.

2. Discriminant codebook learning for BoWs-based action representation

In this section we describe in detail the proposed integrated optimization scheme for discriminant BoWs-based action representation. Let $\mathcal{U}$ be a video database containing N_T action videos followed by action class labels $l_i,\ i = 1,\ldots,N_T$ appearing in an action class set $\mathcal{A} = \{\alpha\}_{\alpha=1}^{C}$. Let us assume that each action video i is described by N_i feature vectors $\mathbf{p}_{ij}, \in \mathbb{R}^D,\ i = 1,\ldots,N_T,\ j = 1,\ldots,N_i$, which are normalized in order to have unit l_2 norm. We employ the feature vectors $\mathbf{p}_{ij}$ and the action class labels l_i in order to represent each action video i by two discriminant feature vectors $\mathbf{s}_i \in \mathbb{R}^D$ and $\mathbf{z}_i \in \mathbb{R}^d,\ d < D$, in the feature space determined by the adopted codebook and the discriminant space, respectively.

2.1. Standard BoWs-based action representation

Let us denote by $\mathbf{V} \in \mathbb{R}^{D\times K}$ a codebook formed by codebook vectors $\mathbf{v}_k \in \mathbb{R}^D,\ k = 1,\ldots,K$. This codebook is calculated by clustering the feature vectors $\mathbf{p}_{ij},\ i = 1,\ldots,N_T,\ j = 1,\ldots,N_i$ without exploiting the available labeling information for the training action videos. Several clustering techniques can be employed to this end. K-Means has been widely adopted for its simplicity and fast operation. The codebook vectors $\mathbf{v}_k$ are, usually, determined to be the cluster mean vectors. After determining the codebook $\mathbf{V}$, the standard BoWs-based action representation of action video i is obtained by applying hard or soft vector quantization on the feature vectors $\mathbf{p}_{ij},\ j = 1,\ldots,N_i$. In the first case, the BoWs-based representation of action video i is a histogram of features, calculated by assigning each feature vector $\mathbf{p}_{ij}$ to the cluster of the closest codebook vector $\mathbf{v}_k$. In the second case, a distance function, usually the Euclidean one, is used in order to determine N_i distance vectors, each denoting the similarity of feature vector $\mathbf{p}_{ij}$ to all the codebook vectors $\mathbf{v}_k$, and the BoWs-based representation of action video i is determined to be the mean normalized distance vector [15].

2.2. Discriminant BoWs-based action representation

The proposed discriminant BoWs-based representation exploits a generalization of the Euclidean distance, i.e.,:

$$d_{ijk} = \|\mathbf{v}_k - \mathbf{p}_{ij}\|_2^{-g}. \tag{1}$$

The use of a parameter value $g = 1.0$ leads to a BoWs-based representation based on soft vector quantization, while a parameter value $g \gg 1.0$ leads to a BoWs-based representation based on hard vector quantization. By using the above distance function, each feature vector $\mathbf{p}_{ij}$ is mapped to the so-called membership vector $\mathbf{u}_{ij} \in \mathbb{R}^K$, encoding the similarity of $\mathbf{p}_{ij}$ to all the codebook vectors $\mathbf{v}_k$. Membership vectors $\mathbf{u}_{ij} \in \mathbb{R}^K$ are obtained by normalizing the distance vectors $\mathbf{d}_{ij} = [d_{ij1}\ldots d_{ijK}]^T$ in order to have unit l_1 norm, i.e.:

$$\mathbf{u}_{ij} = \frac{\mathbf{d}_{ij}}{\|\mathbf{d}_{ij}\|_1}. \tag{2}$$

The BoWs-based representation of action video i is obtained by calculating the mean membership vector:

$$\mathbf{q}_i = \frac{1}{N_i}\sum_{j=1}^{N_i}\mathbf{u}_{ij}. \tag{3}$$

Finally, the mean membership vectors $\mathbf{q}_i$ are normalized in order to produce the so-called action vectors:

$$\mathbf{s}_i = \frac{\mathbf{q}_i}{\|\mathbf{q}_i\|_2}. \tag{4}$$

The adopted similarity function as well as the mean similarity to codebook vectors for action video representation allow for better diffusion of the similarity along the codebook vectors. This is more effective, especially for small feature sets, where the resulting