Computer Vision and Image Understanding 114 (2010) 857–864

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

Action categorization by structural probabilistic latent semantic analysis

Jianguo Zhang [a,*], Shaogang Gong [b]

[a] School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, Belfast BT7 1NN, UK
[b] School of Electronic Engineering and Computer Science, Queen Mary University of London, London E1 4NS, UK

ARTICLE INFO

Article history:
Received 4 July 2007
Accepted 28 April 2010
Available online 4 May 2010

Keywords:
Action categorization
pLSA
Structural pLSA
Local shape context

ABSTRACT

Temporal dependency is a very important cue for modeling human actions. However, approaches using latent topics models, e.g., probabilistic latent semantic analysis (pLSA), employ the bag of words assumption therefore word dependencies are usually ignored. In this work, we propose a new approach *structural* pLSA (SpLSA) to model explicitly word orders by introducing latent variables. More specifically, we develop an action categorization approach that learns action representations as the distribution of latent topics in an unsupervised way, where each action frame is characterized by a codebook representation of local shape context. The effectiveness of this approach is evaluated using both the WEIZMANN dataset and the MIT dataset. Results show that the proposed approach outperforms the standard pLSA. Additionally, our approach is compared favorably with six existing models including GMM, logistic regression, HMM, SVM, CRF, and HCRF given the same feature representation. These comparative results show that our approach achieves higher categorization accuracy than the five existing models and is comparable to the state-of-the-art hidden conditional random field based model using the same feature set.

1. Introduction

Human action recognition is an important and challenging task. Numerous methods have been proposed either focusing on building robust action representations or developing recognition models. In general, there are two key elements in modeling and recognizing human actions [24]: local appearance and temporal dependencies. In this section, we briefly review existing work in the area.

1.1. Motion representations

A great deal of work has been done to represent action sequences. They can be broadly classified into two categories: *part-based representations* and *holistic representations*. Part-based motion representation relates to the success of representing images using a sparse set of interesting points in object images [23], which offers invariance to rotational and affine changes, as well as robustness to background clutter. Laptev [19] proposed to interpret local motion changes by the neighborhoods of the spatial–temporal interest points. This work has been further extended to motion recognition by a bag of spatial–temporal words [25,30,20]. Laptev et al. [20] formulated the problem of motion recognition as a matching of corresponding events in image sequences based on an assumption that similar patterns of motion contain similar events with consistent motion across image sequences. The motion descriptor is constructed as a bag of local events. The local velocity adaptation of events allows this approach to recognize motion independently of the scale and Galilean transformations caused by the relative motion of the camera. One limitation of this approach is lack of modeling the relative structure of events in space time because all of the events are treated independently [20]. Dollar et al. [8] used a denser representation by sampling the interesting points as local maxima in the spatial direction only, and achieved better performance than a sparser representation. Using similar features, Niebles et al. [25] achieved comparable performance with an unsupervised approach, and Nowozin et al. [26] further improved the performance using a discriminative approach. It is worth noting that recent study has unveiled some limitations of interest points based approach for object recognition. When the object is small, the detectors cannot produce sufficient detections [10,42]. The detector of spatial–temporal interest point may suffer from insufficient detections as well when the motion scale is relatively small at a distance [9]. Therefore the performance of this approach under such circumstances is still not known.

Holistic approaches usually represent the motion sequences as a whole. These approaches often rely on dense or global representations. For global representations, a simple approach is to accumulate all measurements in a video sequence using a global descriptor, e.g., the motion history image (MHI) [3]. In the MHI based methods, the temporal influence of motion is encoded as the intensity differences in the MHI template where the recent

* Corresponding author.
E-mail addresses: j.zhang@ecit.qub.ac.uk (J. Zhang), sgg@dcs.qmul.ac.uk (S. Gong).

1077-3142/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.cviu.2010.04.006

motions have higher intensity than the old ones. Inspired by this approach, Weinland et al. [40] developed a 3D motion representation called motion history volume using constrained multiple cameras. Boiman and Irani [4] proposed a motion detection approach by computing the correlation between two spatial–temporal volumes in a dense mode with a 3D sliding window approach. Efros et al. [9] used smoothed optical flow as the motion descriptor to classify human actions at a distance. Holistic approaches are known to be sensitive to the large geometrical variations between intra-class samples, moving cameras and non-stationary background. To handle these challenges, those approaches usually deploy some motion based segmentation techniques and then compute the motion descriptors in the segmented regions [9].

Another type of holistic approach is to represent actions as a sequence of human body shapes. This representation is inspired by the observation that a time series of 2D silhouettes in the space–time volume contains both the spatial configuration about the pose of human figure at anytime (location and orientation of the arms, legs, and torso, aspect ratio of different body parts) and the dynamic information (global body motion and motion of the limbs relative to body) [15]. To date, silhouette-based action recognition has received increasing attention [24,36,33,2]. For modeling objects, shape features contain rich and useful information of the object and have been demonstrated one of the key object descriptors for object detection in complex scenes [13,27]. Working in shape is also advantageous to other image representations, such as the appearance features [5]. For instance, Bosch et al. [5] have successfully incorporated a shape descriptor with the appearance features in a pyramid representation, which achieves the state-of-art recognition performance on the Caltech101 [11] and Caltech 256 [16] datasets with accuracies of 98.2% and 69.8% respectively. The object shape variations along time contain dynamic information of motion as well. Wang et al. [39] have demonstrated that it is promising to discover human motion categories from static images, whose features are represented as shape context [39]. Recently, Gorelick et al. [15] worked on 2D silhouette-based action sequences and represented actions as spatial–temporal shapes. Their approach exploited the solution to the Poisson shape representation to extract various shape properties for classification. Our motion representation is essentially a temporal sequence containing body shapes. The shape descriptor is inspired by the pioneering work of the shape context by Belongie et al. [1] and the success of bag of features representation for object recognition [10,42], thus resulting in a codebook representation of local context for each silhouette.

1.2. Motion models

Another important issue for motion recognition is to develop models to learn the temporal dependencies between consecutive frames. To date, numerous methods have been proposed with the vast majority based on graph models. Among them, the hidden Markov model (HMM) is a baseline approach for modeling temporal dependencies mainly interpreted by a transition matrix. Its model parameters are estimated based on the optimization of the class conditional joint probability distribution between the observations and the sequence labels, which is marginalized over a set of hidden variables. Hence, it is a generative method and not optimized based on the conditional Bayesian information. Though HMM has shown good performance in many applications, for the purpose of pattern discrimination, an existing common consensus is that an ideal model in theory should be derived and optimized based on maximizing the discrimination function [14]. Thus, to this point, HMM is not optimal. To overcome this limitation, conditional random field (CRF) was introduced recently [35,33]. However, CRF cannot incorporate the need for labeling a whole sequence as an action, and also cannot capture the intermediate structures using hidden state variables [29]. To overcome these shortcomings, *hidden conditional random field* (HCRF) was proposed in [17,38,29]. Compared to CRF, HCRF is capable of incorporating a sequence label into the optimization of the probabilities of sequence labels conditioned on observations. Recently, Nowozin et al. [26] proposed a discriminative subsequences mining approach to find the optimal discriminative subsequence patterns. In their approach, each video is encoded as a sequence of set of integers. To train a classifier on such a representation, they extended the PrefixSpan [28] subsequence mining algorithm in combination with LPBoost [7].

However, all of the above model-based approaches are *supervised*, i.e., the training set has to be manually labeled with varying degree of supervision. Thus, another interesting direction is to develop *unsupervised* learning methods, e.g., action recognition by *probabilistic latent semantic analysis* (pLSA) [25].

As one of the generic models, pLSA [31] has been successfully used to discover object categories without prior segmentation. For instance, Sivic et al. [31] used pLSA to automatically find the object categories from a large image collection. Fergus et al. [12] developed a translation and scale invariant pLSA model (TSI-pLSA) to localize an object from just its name, which extends pLSA to include spatial information. Inspired by the success of pLSA for object recognition, researchers have recently applied pLSA for motion classification. Niebles et al. [25] developed an unsupervised motion classification approach using pLSA with spatial–temporal words. However, the approach by Niebles et al. is not capable of localizing motion in videos. In order to tackle this problem and inspired by the successful Implicit Shape Model [22] for object detection, Wong et al. [41] recently extended the pLSA approach to localize motion categories by including the geometrical constrains between spatial–temporal patches, which is called pLSA-ISM. This approach is a promising extension of TSI-pLSA to infer the location of motion in video sequences. Experimental results show good localization performance with little presence of background clutter. However, the performance of this method in the presence of strong background motion clutter is still not known. For the purpose of statistical language modeling, Wang et al. [37] presented a directed Markov random field that combines *n*-gram models, probabilistic context free grammars and pLSA to learn the semantic information. However, as other MRF based approaches, the training process of this model needs the labels of language sequences.

The dynamic adjacent dependencies cannot be learnt explicitly by most of the pLSA based approaches, since the pLSA ignores such global dependencies in principle. To incorporate those dependencies into the unsupervised learning process by pLSA, we proposed a *structural pLSA* (SpLSA),[1] where we show that the standard pLSA is a special case of our model. We then develop an action categorization approach learnt by SpLSA with a codebook representation of local shape context. We compared our model with six other existing models using two standard action recognition datasets. In our experiments, all models were given exactly the same feature sets to remove any effect from the choice of different features.

The paper is organized as follows. Section 2 describes the principles of pLSA. We then present our new model SpLSA and the learning method in Section 3. Section 4 develops a novel action categorization approach based on SpLSA with a signature of codebook of local shape context. We evaluate our approach in Section 5. Section 6 concludes the whole paper and points out some future work.

[1] By 'structural', we mean conditional dependencies as in structural learning for probabilistic graph models.