Neurocomputing 100 (2013) 134–143

Contents lists available at SciVerse ScienceDirect

Neurocomputing

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

Supervised learning probabilistic Latent Semantic Analysis for human motion analysis

Jin Wang [a,*], Ping Liu [b], Mary F.H. She [a], Abbas Kouzani [a], Saeid Nahavandi [c]

[a] Institute for Technology Research and Innovation, Deakin University, Geelong, VIC 3217, Australia
[b] Department of Computer Science and Engineering, University of South Carolina, Columbia, USA
[c] Centre for Intelligent Systems Research, Deakin University, Geelong, VIC 3217, Australia

ARTICLE INFO

Available online 9 May 2012

Keywords:
Motion analysis
Pyramid HoG
pLSA
Supervised learning

ABSTRACT

Latent topic models such as Latent Dirichlet Allocation (LDA) and probabilistic Latent Semantic Analysis (pLSA) have demonstrated success in computer vision tasks. Most existing approaches train LDA and pLSA in an unsupervised manner, where the training data does not include any class label information. However, the class labels in training data are very important for the task of classification. In this paper, we propose to train a pLSA model in a supervised manner for the task of human motion analysis using the bag-of-words representation. Each frame in a video is treated as a word, and all the frames in the training videos are clustered to construct a codebook. The class label information is used to learn the pLSA model in a supervised manner, which not only makes the training more efficient, but also improves the overall recognition accuracy significantly. In addition, we employ the pyramid Histogram of orientation Gradient (HoG) to encode a human figure in each frame. The pyramid HoG descriptor does not require extraction of silhouettes, and is invariant to translations and rotations to some extent. The method is validated using two standard datasets. The experimental results show that our method can accurately recognize human motion in video sequences. Moreover, the overall recognition accuracy is rather stable with respect to the codebook size.

1. Introduction

With the advancement of modern imaging technology and reduction in electronic hardware cost, more surveillance systems are being installed in public places. Automatic analysis of human motion in large number of video sequences is a challenging task. When human figures are recorded in a crowded scene, the resulting video sequences often contain occlusions. Moreover, there may exist substantial variations within the same class of motions performed by different subjects. Even the same activities may be performed by the same subjects at different speeds, giving rise to temporal variations of the same activities. Due to these challenges, human motion analysis algorithms that can model complex scenarios, and simultaneously be robust to viewpoints, noise and occlusion are highly desirable. A general framework for automatic human behavior understanding system may consist of video acquisition, human detection, motion representation, motion recognition, and motion semantic description [1–3]. There exist two important questions associated with the human motion analysis. The first question is how to effectively encode human figures in video sequences. The second question is how to model the temporal dynamic motion sequences so that the variations and similarity of test and reference sequences can be exploited in the training and recognition algorithms.

1.1. Overview of the proposed approach

Our approach is inspired by the success of the bag-of-words method in computer vision fields including image segmentation, object categorization and activity recognition [4–8]. In this paper, a general framework is proposed for the analysis of human motion in videos based on the bag-of-words representation and the probabilistic Latent Semantic Analysis (pLSA) model (see Fig. 1). This framework consists of detecting human subjects in videos, extracting pyramid Histogram of oriented Gradients (HoG) descriptors, constructing a visual codebook by k-means clustering, and supervised learning the pLSA model for recognition.

Once interest regions containing human figures are extracted by tracking or detection algorithms, such as particle filters [9] and background subtraction [10], we characterize the human figures in each frame at different spatial scales using the pyramid HoG descriptor [11]. The pyramid HoG descriptor can encode a human figure in a compact way without extracting human silhouettes. Moreover, the proposed method is invariant to rotation and translation variations to some extent [11].

Each frame described by the pyramid HoG descriptor is treated as a word in the bag-of-words representation. All the unordered groups

* Corresponding author. Tel.: +61 430 47 98 74.
E-mail address: jay.wangjin@gmail.com (J. Wang).

0925-2312/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2011.10.033

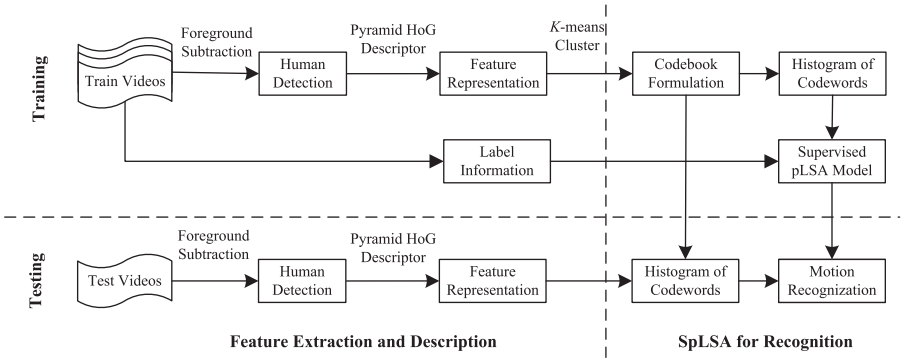

Fig. 1. The flowchart of the proposed approach for human motion recognition based on the supervised pLSA (SpLSA).

of these words from the training video sequences become a bag of words. Although the temporal order information between image frames is lost, the bag-of-words representation remains effective and discriminative due to the characteristics of the supervised learning of the pLSA model. In order to construct the codebook, we cluster the entire pyramid HoG features extracted from training frames using the *k*-means algorithm based on the Euclidean Distance Metric. The center of each cluster is defined as a codeword, and the centers clustered from the training frames produce the codebook.

The pLSA [12] is firstly proposed to model text collection in an unsupervised way. It assumes that the words are generated from a mixture of latent aspects which can be decomposed from a document. Here, we regard each aspect in the pLSA as one particular motion class. In another word, the number of aspects is equal to the number of motion classes in video. As such, the class label of a new video can be determined by the distribution of the aspects in the pLSA. We notice the importance of the class label information in training data for the classification task. Considering this important information, we propose to learn the pLSA model in a supervised manner, which not only simplifies the learning process of the pLSA, but also improves its recognition accuracy.

1.2. Contribution and organization

The main contribution of this paper is threefold. Firstly, we propose to encode human figures in videos by the pyramid HoG descriptor for motion analysis, which does not require extraction of human silhouettes or contours. Secondly, we extend the standard pLSA model to make use of the class label information in the training data, and propose to train the pLSA in a supervised fashion. Specifically, the parameters are directly counted from training videos and no further iteration is required during training. Furthermore, experiments on two public activity datasets are conducted. We achieved comparative or even higher recognition accuracies compared to the other state-of-the-art methods in the literatures. The remainder of the paper is organized as follows. Section 2 briefly reviews the related work on human motion analysis, and the topic models in computer vision. Section 3 gives the details of the proposed approach including motion representation, codebook formulation and the supervised pLSA model. Section 4 analyzes the experimental results associated with our method on two public datasets. Finally, conclusion remarks are given in Section 5.

2. Related work

2.1. Motion representations

Various features and classification methods have been proposed to recognize human motions in video sequences. Appearance-based features focus on the shapes of silhouettes or the contours of the human body, which have the advantage of low computational costs. Wang and Suter [13] divided the raw silhouette sequences into many sub-blocks as visual features. Although relatively simple, the features are effective and efficient. Later, they [14] applied the manifold learning to the Distance Transform of silhouettes to match dynamic shapes of humans. The manifold learning algorithm automatically uses of the structural information in the temporal dynamic shapes.

Fig. 2. Space–time volume of human actions [15].

Some works constructed a space–time volume by stacking the sequences of silhouette or action sequences frame by frame over a given time interval, which can be treated as a three-dimensional shape where the third dimension is the temporal axis. Gorelick et al. [15] stacked silhouette sequences over a temporal interval to construct a 3D space–time volume shown in Fig. 2. They formulated a Poisson equation to encode the 3D shape. These spatial–temporal features were extracted based on the solution of the Poisson equation. Instead of modeling the whole 3D shapes, Batra et al. [16] employed the space–time volumes to construct the mid-level features called Space-Time shapelets. They concatenated these shapelets into a histogram to represent an action. Based on the result of an experiment, the shapelets representation was found to be more robust to partial occlusions and lighting variations.

The silhouette-based methods require detection and localization of subjects in videos, which are usually variant to noise, viewpoint and occlusions to some extent. By contrast, methods based on the local features do not require accurate background subtraction or tracking. Dollár et al. [17] convolved a spatial Gaussian with a pair of Gabor filters along the temporal domain. The spatial–temporal points corresponding to a strong response were defined as interest points. Local patches around the detected interest points were encoded by pixel values, brightness gradient and optical flow. In [18], the 2D Harris corner detector was extended to spatial–temporal domain for detection of interest points in videos. Histogram of orientation Gradient (HoG) and Histogram of optical Flow (HoF) were used to capture appearance and motion information.