Computer Vision and Image Understanding 119 (2014) 27–40

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

Matching mixtures of curves for human action recognition ☆

Michalis Vrigkas [a,*], Vasileios Karavasilis [a], Christophoros Nikou [a], Ioannis A. Kakadiaris [b]

[a] Department of Computer Science, University of Ioannina, Ioannina, Greece
[b] Computational Biomedicine Lab, Department of Computer Science, University of Houston, Houston, TX, USA

ARTICLE INFO

Article history:
Received 28 November 2012
Accepted 22 November 2013
Available online 4 December 2013

Keywords:
Human action recognition
Optical flow
Motion curves
Gaussian mixture modeling (GMM)
Clustering
Dimensionality reduction
Longest common subsequence

ABSTRACT

A learning-based framework for action representation and recognition relying on the description of an action by time series of optical flow motion features is presented. In the learning step, the motion curves representing each action are clustered using Gaussian mixture modeling (GMM). In the recognition step, the optical flow curves of a probe sequence are also clustered using a GMM, then each probe sequence is projected onto the training space and the probe curves are matched to the learned curves using a non-metric similarity function based on the longest common subsequence, which is robust to noise and provides an intuitive notion of similarity between curves. Alignment between the mean curves is performed using canonical time warping. Finally, the probe sequence is categorized to the learned action with the maximum similarity using a nearest neighbor classification scheme. We also present a variant of the method where the length of the time series is reduced by dimensionality reduction in both training and test phases, in order to smooth out the outliers, which are common in these type of sequences. Experimental results on KTH, UCF Sports and UCF YouTube action databases demonstrate the effectiveness of the proposed method.

1. Introduction

Action recognition is a preponderant and difficult task in computer vision. Many applications, including video surveillance systems, human–computer interaction, and robotics for human behavior characterization, require a multiple activity recognition system. The goal of human activity recognition is to examine activities from video sequences or still images. Motivated by this fact our human activity recognition system aims to correctly classify a video into its activity category.

In this paper, we address the problem of human action recognition by representing an action with a set of clustered motion curves. Motion curves are generated by optical flow features which are then clustered using a different Gaussian mixture [1] for each distinct action. The optical flow curves of a probe sequence are also clustered using a Gaussian mixture model (GMM) and they are matched to the learned curves using a similarity function [2] relying on the longest common subsequence (LCSS) between curves and the canonical time warping (CTW) [3]. Linear [1] and nonlinear [4] dimensionality reduction methods may also be employed in order to remove outliers from the motion curves and reduce their lengths. The motion curve of a new probe video is projected onto its own subspace by a projection matrix specified by that video, and then the action label of the closest projection is selected according to the learned feature vectors as the identity of the probe sequence. The LCSS is robust to noise and provides an intuitive notion of similarity between curves. Since different actors perform the same action in different manners and at different speeds, an advantage of the LCSS similarity is that it can handle with motion curves of varied lengths. On the other hand, CTW, which is based on the dynamic time warping [5], allows the spatio-temporal alignment between two human motion sequences. A preliminary version of this work was presented in [6]. One of the main contributions of this paper is that the training sequences do not need to have the same length. When a new probe sequence comes, it is matched against all the training sequences using the LCSS similarity measure. This measure provides a similarity between motion curves without enforcing one-to-one matching. An optimal matching is performed using dynamic programming, which detects similar pairs of curve segments [2].

However, training an action recognition system with only the knowledge of the motion of the current subject it is on its own a challenging task. The main problem is how we can ensure the continuity of the curves along time as an action occurs uniformly or non-uniformly within a video sequence. Unlike other approaches [7,8], which use snippets of motion trajectories, our approach uses the full length of motion curves by tracking the optical flow features. Another question concerns the optimal model that one should adopt for recognizing human actions with high accuracy.

☆ This paper has been recommended for acceptance by Jordi González.

* Corresponding author.
E-mail addresses: mvrigkas@cs.uoi.gr (M. Vrigkas), vkaravas@cs.uoi.gr (V. Karavasilis), cnikou@cs.uoi.gr (C. Nikou), ioannisk@uh.edu (I.A. Kakadiaris).

1077-3142/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.cviu.2013.11.007

This is accomplished by a statistical measure based on the data likelihood. The different lengths of the video sequences and therefore the respective lengths of the motion curves is another problem that is addressed. The large variance between benchmark datasets shows how the algorithm may be generalized. All these problems are discussed here and proper solutions are proposed. To this end, we have conducted experiments on several datasets [9–11] that would help us to understand how human activity recognition works.

Concatenating of optical flow features along time allows us to collect time series that preserve their continuity along time. It is true that correspondence is missing. However, this is the main assumption in many works [12–14]. If data association were used the resulting feature curves would have short duration and would be incomplete, as the features disappear and reappear due to occlusion, illumination, viewpoint changes and noise. In that case, a combination of sparse approach of clustering curves with variant lengths and tracking approaches should be used [15,16]. This is not the central idea in this paper, as the nature of the feature curves drastically changes.

In the rest of the paper, the related work is presented in Section 2, while the extraction of motion curves, the clustering and the curve matching are presented in Section 3. In Section 4, we report results on the KTH [9], the UCF Sports [10] and the UCF YouTube [11] action classification datasets. Finally, conclusions are drawn in Section 5.

2. Related work

The problem of categorizing a human action remains a challenging task that has attracted much research effort in the recent years. The surveys in [17,18] provide a good overview of the numerous papers on action/activity recognition and analyze the semantics of human activity categorization. Several feature extraction methods for describing and recognizing human actions have been proposed [12,9,19,20,13]. A major family of methods relies on optical flow which has proven to be an important cue. Efros et al. [12] recognize human actions from low-resolution sports video sequences using the nearest neighbor classifier, where humans are represented by windows of height of 30 pixels. The approach of Fathi and Mori [13] is based on mid-level motion features, which are also constructed directly from optical flow features. Moreover, Wang and Mori [14] employed motion features as inputs to hidden conditional random fields and support vector machine (SVM) classifiers. Real time classification and prediction of future actions is proposed by Morris and Trivedi [21], where an activity vocabulary is learnt through a three step procedure. Other optical flow-based methods which gained popularity are presented in [22–24].

The classification of a video sequence using local features in a spatio-temporal environment has also been given much consideration. Schuldt et al. [9] represent local events in a video using space–time features, while an SVM classifier is used to recognize an action. Gorelick et al. [25] considered actions as 3D space time silhouettes of moving humans. They take advantage of the Poisson equation solution to efficiently describe an action by utilizing spectral clustering between sequences of features and applying nearest neighbor classification to characterize an action. Niebles et al. [20] addressed the problem of action recognition by creating a codebook of space–time interest points. A hierarchical approach was followed by Jhuang et al. [19], where an input video is analyzed into several feature descriptors depending on their complexity. The final classification is performed by a multi-class SVM classifier. Dollár et al. [26] proposed spatio-temporal features based on cuboid descriptors. An action descriptor of histograms of interest points, relying on [9] was presented in [27]. Random forests for action representation have also been attracting widespread interest for action recognition [28,29]. Furthermore, the key issue of how many frames are required to recognize an action is addressed by Schindler and Van Gool [30].

The problem of identifying multiple persons simultaneously and perform action recognition is presented in [31]. The authors considered that a person has first been localized by performing background subtraction techniques. Based on the Histograms of Oriented Gaussians [32] they detect a human, whereas classification of actions are made by training a SVM classifier. Action recognition using depth cameras are introduced in [33] and a new feature called "local occupancy pattern" is also proposed. A novel multi-view activity recognition method is presented in [34]. Descriptors from different views are connected together forming a new augmented feature that contains the transition between the different views. A new type of feature called the "Hankelet" is presented in [35]. This type of feature, which is formed by short tracklets, along with a BoW approach is able to recognize actions under different viewpoints, without requiring any camera calibration. Zhou and Wang [36] have also proposed a new representation of local spatio-temporal cuboids for action recognition. Low level features are encoded and classified via a kernelized SVM classifier, whereas a classification score denotes the confidence that a cuboid belongs to an atomic action. The new feature act as complementary material to the low-level feature.

Earlier approaches are based on describing actions by using dense trajectories. The work of Wang et al. [37] is focused on tracking dense sample point from video sequences using optical flow. Le et al. [38] discover the action label in an unsupervised manner by learning features directly from video data. A high-level representation of video sequences, called Action Bank, is presented by Sadanand and Corso [39]. Each video is represented as a set of action descriptors which are put in correspondence. The final classification is performed by a SVM classifier. Yan and Luo [27] have also proposed a new action descriptor based on spatial temporal interest points (STIP) [40]. In order to avoid overfitting they have also proposed a novel classification technique by combining the Adaboost and sparse representation algorithms. Wu et al. [41] employed, a visual feature using Gaussian mixture models efficiently represents the spatio-temporal context distributions between the interest point at several space and time scales. An action is represented by a set of features extracted by the interest points over the video sequence. Finally, a vocabulary based approach has been proposed by Kovashka and Grauman [42]. The main idea was to find the neighboring features around the detected interest points quantize them and form a vocabulary. Raptis et al. [8] proposed a mid-level approach extracting that spatio-temporal features construct clusters of trajectories, which can be considered as candidates of an action, and a graphical model is utilized to control these clusters.

Human action recognition using temporal templates has also been proposed by Bobick and Davis [43]. An action was represented by a motion template composed of a binary motion energy image (MEI) and a motion history image (MHI). Recognition was accomplished by matching pairs of MEI and MHI. A variation of the MEI idea was proposed by Ahmad and Lee [44], where the silhouette energy image (SEI) was proposed. The authors have also introduced several variability models to describe an action, and action classification was carried out using a variety of classifiers. Moreover, the proposed model is sensitive to illumination and background changes. In the sense of template matching techniques Rodriguez et al. [10] introduced the Maximum Average Correlation Height (MACH) filter which is a method for capturing intra-class variability by synthesizing a single action MACH filter for a given action class.