Image and Vision Computing 48–49 (2016) 26–36

Contents lists available at ScienceDirect

Image and Vision Computing

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

A Time Flexible Kernel framework for video-based activity recognition☆

Mario Rodriguez[a,*], Carlos Orrite[a], Carlos Medrano[a,b], Dimitrios Makris[c]

[a] CVLab, I3A, Zaragoza University, c/Mariano Esquillor s/n, 50018 Zaragoza, Spain
[b] EduQTech, E.U. Politecnica, Zaragoza University, c/Ciudad Escolar s/n, 44003 Teruel, Spain
[c] Digital Imaging Research Centre, Kingston University, Penrhyn Road, Kingston upon Thames, Surrey KT1 2EE, UK

ARTICLE INFO

Article history:
Received 21 January 2015
Received in revised form 4 September 2015
Accepted 23 December 2015
Available online 24 February 2016

Keywords:
Activity recognition
Soft-assignment
Kernel methods
Support Vector Machine

ABSTRACT

This work deals with the challenging task of activity recognition in unconstrained videos. Standard methods are based on video encoding of low-level features using Fisher Vectors or Bag of Features. However, these approaches model every sequence into a single vector with fixed dimensionality that lacks any long-term temporal information, which may be important for recognition, especially of complex activities. This work proposes a novel framework with two main technical novelties: First, a video encoding method that maintains the temporal structure of sequences and second a Time Flexible Kernel that allows comparison of sequences of different lengths and random alignment. Results on challenging benchmarks and comparison to previous work demonstrate the applicability and value of our framework.

1. Introduction

Significant research effort has been invested in video-based activity recognition during the last few years, supported by the widespread availability of video cameras, as it may benefit many applications such as video indexing, surveillance or entertainment.

A video is a sequence of frames that can be viewed as a 3-dimensional matrix of pixels, two dimensions provide the space localization and the third one is related to time. When displayed in a screen, as a sequence of images, humans are able to easily distinguish among activities, but the same task is extremely challenging for a computational method. The machine learning community has suggested several approaches with their advantages and disadvantages, but results in unconstrained recordings are still far from what humans may achieve.

The design of sophisticated low-level descriptors [1,2,3] has been central in recent advances for this research challenge. Specifically, space-time feature codification has been present in the state-of-the-art approaches where video sequences are represented by a Bag of Features (BoF) or a Fisher Vector (FV), encoding the extracted features. The recognition process is carried out afterwards by applying a multi-class Support Vector Machine (SVM) [4], which takes advantage of the kernel trick. Despite the promising performance of these approaches there are two drawbacks due to the characteristics of the descriptors: (i) using image or short-term descriptors, the lack of explicit temporal information withhold them from reliable recognition of activities [5], and (ii) mid-term descriptors may describe better the activities [6] but still lack of information of the whole temporal structure making them unreliable for complex activities where the order of sub-actions describes the activity.

On the other hand, some state space models such as Hidden Markov Models (HMM) [7] or more recent Conditional Random Fields [8], codify the long term temporal information of the sequences. Although they have provided satisfactory results in human activities they usually work in constrained scenarios such as ones implied by the datasets KTH, Weizmann or UT-Tower, where there is no camera motion and the point of view is fixed [9,10,11]. Thanks to these restrictions it is possible to train states encompassing common characteristics among the videos and thus to achieve high accuracy. However, new databases, such as HMDB51, UCF50, OlympicSports and Virat Release 2.0 have been produced aiming at challenging tasks, such as indexing events in unconstrained videos in the internet or surveillance in uncontrolled scenarios performing a scene-independent learning and recognition. These datasets were recorded in unconstrained environments with random viewpoints, camera movements and/or dynamic changes in the background.

☆ This paper has been recommended for acceptance by Ioannis Patras, PhD.
Abbreviations: (TFK) , Time Flexible Kernel; (IDT) , Improved Dense Trajectories; (MIP) , Motion Interchange Patterns; (PPK) , Probability Product Kernel.

* Corresponding author.
E-mail addresses: mrodrigo@unizar.es (M. Rodriguez), corrite@unizar.es (C. Orrite), ctmedra@unizar.es (C. Medrano), d.makris@kingston.ac.uk (D. Makris).

http://dx.doi.org/10.1016/j.imavis.2015.12.006
0262-8856/© 2016 Elsevier B.V. All rights reserved.

Some of the best results in these challenging benchmark datasets have been obtained with variations of the mentioned SVM approach [12,4,6], which has been proven to be a convenient method in spite of the lack of long-term dynamic information. Nevertheless, the long-term temporal information is important in the description of complex activities and thus, we propose the recognition framework depicted in Fig. 1 where such information is maintained. Both BoF and FV create a codebook, the former using k-means clustering of the training descriptors and the latter using an EM-GMM algorithm. Following the application of sliding frame-windows, each video is modelled as a sequence of BoFs or FVs. The use of a window with few frames produces sparse data so we minimize its effect by using a soft-assignment approach when using BoFs. These sequences preserve the long-term dynamic information needed for the recognition of complex activities. However, as the sequences length and the pace of actions are variable, standard kernels obtained between vectors of the same length are not applicable and novel approaches, like Spatio-Temporal Pyramid Matching (STPM) [13], keep the long-term information, but they rely on perfect alignment of the sequences with regular pace of actions. Nevertheless, it is worth noting the improvement achieved using several encoding scales, proposed in STPM, that we also apply into our work. So, our contribution in this paper includes the design of a novel kernel formulation between arbitrary length sequences that allows the use of the long-term dynamic information in a SVM with matching flexibility, named Time Flexible Kernel (TFK). In order to validate our contribution we have carried out several experiments in four challenging datasets: HMDB51, UCF50, OlympicSports and Virat Release 2.0.

The rest of the paper is divided as follows. Section 2 reviews some related works. Section 3 explains the proposed framework, focusing on our two main technical contributions: a novel encoding scheme and the Time Flexible Kernel, as well as its application for activity recognition. Section 4 presents our experimental validation and Section 5 concludes the work.

2. Related work

Feature extraction is a key element in recognition systems thus a significant number of methods have been proposed. Descriptors can be divided into global and local or low-level features, however, the former group is sensitive to noise, occlusions and viewpoint variations that can be experienced in challenging datasets such as HMDB51, UCF50, OlympicSports and Virat Release 2.0. Therefore, researchers have mainly focused on local descriptors that can be roughly categorised as: (i) image descriptors like SIFT, HOG, Harris, which lack any kind of time information or (ii) space-time descriptors like their extensions 3D-SIFT [2], HOG3D [3] and spatio-temporal Harris interest points [1]. Some descriptors of the latter group are designed so to capture directly the video information as Motion Interchange Patterns (MIP) [5], HOG-HOF [14] or SCISA [15] do. Most of the approaches use holistic encoding based on BoF or FV to derive a vector for each video sequence. Then, SVM is used for multi-class classification, either using a one-against-one approach or a one-against-all approach for multi-class recognition. Until recently, BoF has been the standard approach in video encoding, quantizing the feature vectors into a predefined codebook. However, recent works [6,4] have adopted the strategy of encoding the feature information into a FV which keeps second-order information in relation to an estimated Gaussian Mixture Model (GMM). Alternatively to the above framework, Can and Manmatha [16] modify the SVM strategy to a ranking problem obtaining encouraging results. Finally, Dense Trajectories [17,18] and the actual state of the art in many benchmarks Improved Dense Trajectories (IDT) [6] are based on short trajectories of local descriptors within a sliding temporal window.

The activity encoding methods such as BoF and FV require a “visual word” dictionary. Usually, such a dictionary is created using clustering of the extracted features in the training examples. In the case of BoF every feature is roughly assigned to a word of the dictionary. However, quantization errors are caused, when only a small amount of samples are used, as in the case of the use of a narrow temporal window. Soft-assignment approaches have been proposed to deal with this problem [19,20,21]. Alternatively, FV models the codebook as a GMM and the encoding produces a vector of $K(2d+1)$ dimensions, K is the number of Gaussians and d is the dimension of descriptors, where second order information of the GMM is used [4,6]. However, the main drawback of both approaches is that the long-term temporal information is lost since the encodings are obtained from an unordered collection of local descriptors.

Many approaches have been designed to account complex temporal structures recognizing complex human activities defined as composite multimedia semantics (e.g., birthday party, wedding ceremony) where orderless sub-actions appear in the video. These approaches consider the order of sub-actions as a distracter (not a discriminant) and hence they perform an alignment of similar sub-scenes disregarding their order. For instance Xu et al. [22] divide the clips into sub-clips which later are matched with other sequence using the earth mover's distance (EMD). Cao et al. [23] have designed a kernel that makes a pooling of the frames into a fixed number of scenes called Scene Alignment Pooling (SAP). In [24] a detection of sub-scene categories and a global scene category are combined in a Multiple Kernel Latent SVM where several features are used. In the work of Li et al. [25] the proposed method identifies the most representative segments of the actions using a dynamic pooling with a latent variable.

As opposed to the previously explained works, the temporal order of the sub-actions performed in an activity is considered essential in this paper, as the objective is to distinguish between complex activities that can be composed of the same sub-actions but in different order, even opposite as for instance “Getting Out of Vehicle” and “Getting Into Vehicle” in Virat dataset, Fig. 2.

Thanks to the kernel trick the SVM can classify in a dimensional space different from the original one where samples may be linear-separable. The standard kernel methods assume a fixed length D-dimensional vector per sample which is projected into a different space where the inner product is performed. However, this is not straightforward in activity recognition videos where the long-term activity dynamic information remains in the encoding because lengths of sequences may be arbitrary. Two solutions have been proposed in the literature: (i) obtaining some sort of inner product by aligning the sequences lengths of the patterns, as Dynamic Time-Alignment Kernel [26] or Fast Global Alignment Kernel [27] do and (ii) training a HMM with a single sequence and posterior obtaining a Probability Product Kernel (PPK) [28] like in [29]. Both solutions have been used in sequence clustering tasks [30,29,31]. Sequence alignment enforces a common start and end in the sequences which is not always the case. On the other hand, the PPK of HMMs implies that each HMM is trained with only one sequence which does not offer sufficient information to train properly the parameters of a complex model. Moreover, the optimization process in the HMM training is performed with the Baum-Welch algorithm which only assures a local optimum.

The long-term temporal information has been used in some other methods. A recent one is the extension of the work of Spatial Pyramid Matching (SPM) [32] called Spatio-Temporal Pyramid Matching (STPM) [13]. The method suggests dividing the videos into equal number of spatio-temporal volumes at several scales, called pyramids, computing in each volume a BoF, and finally obtaining a similarity between two video clips by comparing the corresponding volumes. Fixing the number of divisions and comparing volumes one-to-one constrain the method to regular paced actions losing flexibility. In [33]