Image and Vision Computing 27 (2009) 1814–1825

Contents lists available at ScienceDirect

Image and Vision Computing

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

Sparse B-spline polynomial descriptors for human activity recognition

Antonios Oikonomopoulos [a,*], Maja Pantic [a,b], Ioannis Patras [c]

[a] Department of Computing, Imperial College London, UK
[b] Faculty of Electrical Engineering, Mathematics and Computer Science, University of Twente, The Netherlands
[c] Department of Electronic Engineering, Queen Mary University of London, UK

ARTICLE INFO

Article history:
Received 19 May 2008
Received in revised form 10 April 2009
Accepted 16 May 2009

Keywords:
Optical flow
Median filter
Spatiotemporal salient points
B-splines
Gentleboost
Relevance Vector Machine

ABSTRACT

The extraction and quantization of local image and video descriptors for the subsequent creation of visual codebooks is a technique that has proved very effective for image and video retrieval applications. In this paper we build on this concept and propose a new set of visual descriptors that provide a local space-time description of the visual activity. The proposed descriptors are extracted at spatiotemporal salient points detected on the estimated optical flow field for a given image sequence and are based on geometrical properties of three-dimensional piecewise polynomials, namely B-splines. The latter are fitted on the spatiotemporal locations of salient points that fall within a given spatiotemporal neighborhood. Our descriptors are invariant in translation and scaling in space-time. The latter is ensured by coupling the neighborhood dimensions to the scale at which the corresponding spatiotemporal salient points are detected. In addition, in order to provide robustness against camera motion (e.g. global translation due to camera panning) we subtract the motion component that is estimated by applying local median filters on the optical flow field. The descriptors that are extracted across the whole dataset are clustered in order to create a codebook of 'visual verbs', where each verb corresponds to a cluster center. We use the resulting codebook in a 'bag of verbs' approach in order to represent the motion of the subjects within small temporal windows. Finally, we use a boosting algorithm in order to select the most discriminative temporal windows of each class and Relevance Vector Machines (RVM) for classification. The presented results using three different databases of human actions verify the effectiveness of our method.

1. Introduction

Due to its practical importance for a wide range of vision-related applications like video retrieval, surveillance, vision-based interfaces, and human–computer interaction, vision-based analysis of human motion is nowadays one of the most active fields of computer vision. One of the main goals in this field is to efficiently represent an activity captured by a camera and to accurately classify it, that is, assign it into one or more known action categories.

Given a video sequence, humans are usually able to deduce quickly and easily information about its content. In particular, they can discriminate relatively easily between a wide range of activities, even if they have observed each of the activities only a few times. By contrast, the development of computational methods for robust activity recognition still remains a very challenging task. Moving camera conditions, dynamic background, occlusions, abrupt illumination changes and multiple subjects in the scene, introduce significant difficulties in the development of a robust motion analysis framework. This is evident from the abundance of different motion analysis approaches that have been developed [1–5].

The aim of this work is to obtain a good representation of the human activity depicted in an image sequence and classify it into one or more activity categories. For robustness against occlusion and clutter, we opt for a sparse representation that is based on visual descriptors extracted around a set of spatiotemporal interesting points. An important issue that we address is handling general motions caused by a moving camera. We do so by detecting the space-time interesting points on the estimated optical flow of the image sequence. In order to filter out vectors that correspond solely to camera motion, we locally subtract the median of the optical flow vectors prior to the interesting point detection. In this way, the detected points correspond to areas of independent motion in the scene. Inspired by the success of 'bag of word' models, which rely on a codebook constructed by clustering static spatial descriptors, we build a 'bag of verbs' model by clustering local space-time descriptors. We use a boosting algorithm in order to select the most discriminant sets of codewords for each class in the training set. Finally, we use Relevance Vector Machines (RVM) for classification. The kernel of the RVM is defined according to the proximity between the test examples and the selected features of each class

[*] Corresponding author. Address: Department of Computing, Imperial College London, 180 Queensgate SW7 2AZ, London, UK. Tel.: +44 7950585131.
E-mail addresses: aoikonom@imperial.ac.uk (A. Oikonomopoulos), m.pantic@imperial.ac.uk (M. Pantic), ioannis.patras@elec.qmul.ac.uk (I. Patras).

0262-8856/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.imavis.2009.05.010

in the training set. In order to demonstrate the efficiency of our approach, we present experimental results on three different datasets of human activities. The latter range from simple aerobics exercises to common everyday actions, like walking and running.

1.1. Related work

Activity recognition systems can be divided into two main categories. Within the first category fall methods that use tracking in order to represent the actions. Several methods have been proposed, including tracking of articulated models (e.g. [6–8]), tracking of landmark points (e.g. [9–12]), or methods that attempt to track specific shapes (e.g. [13,14]), like silhouettes and hand shapes. The result is subsequently used for recognition, either by taking into account the resulting trajectories of the landmark points or by taking into account the temporal transitions of the tracked models and shapes.

The difficulty in acquiring reliable trajectories for recognition has lead several researchers to assume that they are known a-priori (e.g. [11,12]). This difficulty originates from the fact that articulated objects (e.g. the human body) can undergo various changes in appearance and geometry due to rotations, deformations, rapid non-linear motions, and partial occlusions. Furthermore, the high dimensionality of the problem, and appearance changes that lead to the so-called drifting problem [15], make tracking of body parts cumbersome and in most cases unreliable.

Within the second category fall methods that rely on local spatiotemporal feature-descriptor representations. Their success in object detection and localization, their sparsity, and robustness against illumination, clutter, and viewpoint changes [16] have inspired a number of methods in the area of motion analysis and activity recognition. Detection of keypoints, in particular, has been very popular, due to their detection simplicity. A typical example are the space-time interest points [17,18], which correspond roughly to points in space-time where the direction of motion changes abruptly. A similar approach is used by Dollar et al. [19], where an activity is summarized using sets of space-time cuboids. Entropy based spatiotemporal salient point representations are used in [20], as a temporal extension of the salient point detector proposed in [21]. The method takes into account the information content of pixels within a spatiotemporal neighborhood and detects areas where there is a significant amount of motion. One of the most common type of descriptors stems from the Scale Invariant Feature Transform (SIFT). Introduced in [22], it has been used widely in a variety of applications, including object classification (e.g. [23,22]) and scene classification (e.g. [24,25]). The underlying concept in SIFT is the use of a cascade of Gaussian filters of variable width. Keypoints are subsequently detected as the extrema of the Difference of Gaussian filters (DoG) across different scales. Shape contexts [26] constitute an alternative local representation, in which a log polar histogram of the object's edges is used in order to capture local shape. Its robustness against scale changes and its ability to capture local spatial structure have made it very appealing for applications related to human detection (e.g. [27]). A similar and very effective approach for capturing local structure in space and time are the histograms of oriented gradients (HoG), extensively used for activity recognition (e.g. [28–31]). Biologically inspired representations, such as the C features, have been proposed in [32,33]. The method works in an hierarchical way and the obtained features are invariant to scale changes in space and time. Finally, Wong and Cipolla [34] use global information in terms of dynamic textures in order to minimize noise and detect their interesting points.

Illumination variability, smooth motions and, moving camera conditions, have lead several researchers to implement their methods in domains other than the intensity values at the image pixels. Optical flow, in particular, has been a popular choice. Ke et al. [35] use optical flow fields in their volumetric feature detector in order to represent and recognize actions. The authors claim that their method is robust to camera motion, however they do not explicitly handle it, making the method sensitive to less smooth motions of the camera. Shape flows [36,37] has been another method for dealing with camera motion. In this method, motion flow lines acquired by tracking [37] or using MPEG motion vectors [36] are used in order to represent the activities. Matching is done directly using the optical flow lines. However, the matching problem is NP-hard, and while relaxation methods can reduce the computational complexity, it still remains high. Fathi and Mori [38] use mid-level features consisting of optical flow and spatial gradient vectors and use two rounds of boosting in order to train their classifier. Ahmad and Lee [39] use a combination of shape flow and image moments in order to build their descriptors. However, their method relies on silhouettes that are extracted by background subtraction. Shechtman and Irani [40] propose an algorithm for correlating spatiotemporal event templates with videos without explicitly computing the optical flow. Their work, in conjunction with the temporal templates of Bobick and Davis [41] is used in [42] in order to construct a descriptor of shape and flow for detecting activities in the presence of clutter (e.g. crowds).

Exemplar-based methods, like the ones mentioned above, often require a large amount of training examples. Furthermore, in order to classify an unknown instance of an activity, the number of comparisons that have to be performed is equal to the number of the exemplars in the training set. This makes classification a time consuming process. To remedy this, a number of recent works use visual codebooks in order to detect and recognize objects and/or humans. The visual codebook creation is performed by clustering the extracted feature descriptors in the training set [43]. Each of the resulting centers is considered to be a codeword and the set of codewords forms a 'codebook'. In a 'bag of words' approach, each instance (for example an image) is represented as a histogram of codewords. Recognition is then performed by means of histogram comparison.

Visual codebooks have been extensively used for detecting objects, humans and activities. Aiming at object recognition, in [23], SIFT-like descriptors are extracted hierarchically and a visual codebook is created for each level of the hierarchy. Then, the histogram of the descriptors at each level of the hierarchy is classified using Support Vector Machines (SVM). SIFT descriptors in a bag-of-words framework are also used in [24] for the combined problem of event, scene, and object classification, with application to sports images. In [44], a voting scheme similar to the one by Leibe et al. [45] is implemented for localization and recognition of activities. An interesting work is presented in [46], where oriented rectangles are fitted on human silhouettes and matched against a visual codebook. However, the use of silhouettes assumes knowledge of the background and is sensitive to noise and camera motion. Furthermore, the system in [46] ignores dynamic information, and a human activity is considered as a sequence of static poses.

The major weakness of 'bag of words' approaches is that, by histogramming the descriptors, any information about their relative position is lost. In an attempt to remedy this, several researchers have proposed approaches that attempt to encode the spatial relationships between the features. One such approach is the relative encoding of the feature positions by considering a reference point, i.e. the center of the object on which the feature is extracted. Notable works which employ this concept for modeling static objects are those by Marszalek and Schmid [47] and by Leibe et al. [45]. A similar method is used in [48], where the features consist of fragments belonging to object edges, while the position of each fragment is stored relatively to the object's center. Alternatives to this concept of structure include the 'doublets' of Sivic et al. [49]