Pattern Recognition Letters 72 (2016) 62–71

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Discriminative human action classification using locality-constrained linear coding☆

Hossein Rahmani [a], Du Q. Huynh [a,*], Arif Mahmood [a,b], Ajmal Mian [a]

[a] School of Computer Science and Software Engineering, The University of Western Australia, Crawley WA 6009, Australia
[b] School of Mathematics and Statistics, The University of Western Australia, Crawley WA 6009, Australia

ARTICLE INFO

Article history:
Available online xxx

Keywords:
Human action classification
Locality-constrained linear coding
Sparse coding
SVM classifier

ABSTRACT

We propose a Locality-constrained Linear Coding (LLC) based algorithm that captures discriminative information of human actions in spatio-temporal subsequences of videos. The input video is divided into equally spaced overlapping spatio-temporal subsequences. Each subsequence is further divided into blocks and then cells. The spatio-temporal information in each cell is represented by a Histogram of Oriented 3D Gradients (HOG3D). LLC is then used to encode each block. We show that LLC gives more stable and repetitive codes compared to the standard Sparse Coding. The final representation of a video sequence is obtained using logistic regression with ℓ_2 regularization and classification is performed by a linear SVM. The proposed algorithm is applicable to conventional and depth videos. Experimental comparison with ten state-of-the-art methods on three depth video and two conventional video databases shows that the proposed method consistently achieves the best performance.

1. Introduction

Human action classification from videos is an important problem because of its many applications in video surveillance, human-computer interaction, sports analysis, and elderly health care [1,33]. However, automatic classification of human actions in videos is a challenging problem. For colour videos, lighting conditions in the environment and clothing worn by the human subject can both affect the performance of the action classification algorithms. While these problems are eliminated in depth videos, issues about occlusion, loose clothing, and variations in style and execution speed of actions remain.

We propose a human action classification algorithm based on *Locality-constrained Linear Coding* (LLC). We demonstrate improved action classification performance for actions captured by both depth and colour videos (Fig. 1). [32] compared LLC versus SC and found that LLC holds more essential information than SC for object classification. In this paper and in our earlier work in this direction [18] we investigated the use of LLC for encoding human actions. We found that to favour sparsity SC tends to select quite different elements from the action feature dictionary even for the same action. This has an adverse effect as it increases the intra-action-class variation. The LLC, on the other hand, exerts locality constraints on the features and tends to select the same elements in the dictionary for the same action. Our research contributions are fourfold:

- We propose using the locality-constrained linear coding for human action classification.
- We propose a sequence descriptor for each action video to be constructed in a hierarchical fashion, including the computation of the cell descriptor, block descriptor, and subsequence descriptor.
- We propose using maximum pooling and a logistic regression classification with ℓ_2 regularization for action classification.
- We demonstrate the effectiveness of the proposed algorithm over the existing techniques for improved action classification accuracy.

To show that LLC is effective for human action classification, we design our sequence descriptor based on LLC and compare its accuracy on human action classification with the descriptor computed using SC. Furthermore, we evaluate these sequence descriptors against ten state-of-the-art techniques. For the benchmark depth video datasets (*MSRGesture3D*, *MSRAction3D*, and *MSRActionPairs3D*), we compare our algorithm against the algorithms of

- [11], where the local descriptor is based on the histogram of oriented 3D spatio-temporal gradients (or HOG3D);
- [30], where the random occupancy patterns (ROP) are used;
- [31], where an actionlet ensemble model is learned;
- [17], where the histogram of oriented 4D normals (HON4D) features are used;

☆ This paper has been recommended for acceptance by Anders Heyden, Ph.D.

* Corresponding author. Tel.: +61 8 64882878; fax: +61 8 64881089.
E-mail address: du.huynh@uwa.edu.au (D.Q. Huynh).

http://dx.doi.org/10.1016/j.patrec.2015.07.015
0167-8655/

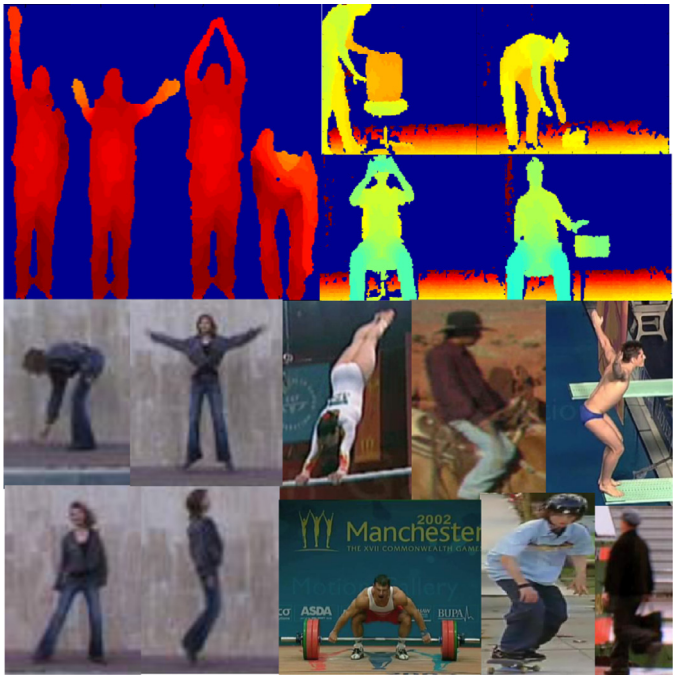

Fig. 1. Our proposed algorithm is capable of classifying a wide range of human actions under varying conditions and captured by different types of sensors. The top two rows show depth images captured by Kinect like sensors. In addition to the subject, other objects may also be present in the depth map. The bottom two rows show the colour images captured by RGB video cameras. These images may have very low resolution, wide variation in backgrounds and viewing angles.

- [35], where space time interest points from depth sequences (DSTIP) are used;
- [20], where local features encoding variations of depths and depth gradients plus skeletal body joints are used with a random decision forest (RDF) classifier.

For the benchmark colour video datasets (*Weizmann* and *UCFSports*), we compare our algorithm with the algorithms of [11] and the following:

- [38], where the mapping between densely-sampled feature patches and the votes in a spatio-temporal action Hough space is trained using random trees;
- [42], where the HOG3D feature was computed within a small 3D cuboid centred at a space-time point and encoded in a sparse coding framework;
- [13], where a figure-centric word representation is used;
- [21], where spatio-temporal structures from clustering of point trajectories of body parts are used;
- [28], where a deformable part model is generated for each action from a collection of examples, with actions being treated as spatio-temporal patterns in the colour videos.

In all of these experiments, our proposed algorithm has consistently shown improved performance compared to the existing algorithms.

2. Related work

Many approaches to human action classification and classification involve analyzing the colour videos directly [5,6,28,40]. Since the release of the Kinect camera in 2011, an increasing number of human action classification papers targeting at analyzing depth videos, colour+depth videos, and/or skeletal data start to emerge [10,12,15,17,19,20,30,31,35,37].

Some action classification algorithms exploited silhouette and edge pixels to form discriminative features. For example, [3] stacked up the sequence of silhouettes from each action video to form a *motion-energy image* and constructed action descriptors using the Hu moments; [39] used a skeletal representation extracted from the human silhouette in the input colour video. While the human silhouette supplies useful cue about the human body pose, to reliably extract the silhouette for human action classification, the background image must be known or the background must not be cluttered.

A large category of approaches to human action classification involve using local descriptors. Some examples are: 2D and 3D SIFT [24] descriptors used in conjunction [26] with motion-energy image; HOG3D features [11] computed in small 3D cuboids and then encoded [42] using *sparse coding* (SC); wavelet-based local descriptor [25] in a bag-of-feature framework; SURF descriptors used to describe cuboids at spatio-temporal interest points [36]. Just like space-time interest points (STIPs) [14] can be extracted from RGB videos, depth-based STIPs have been extracted from depth videos [35]. Histograms of the normal vectors computed from depth images have been used for object classification [27] and these spatial derivatives have been extended to the temporal dimension [17].

Other approaches include operating directly on the depth map or looking for occupancy patterns in the depth videos. For example, [15] sampled 3D points from the depth map and used an action graph to model the dynamics of the action; [29] used space-time occupancy patterns to represent actions. Instead of just labelling each cell in the space-time volume as occupied or not, [30] extracted semi-local features called *random occupancy pattern* (ROP) features. They then used SC to encode these features and an SVM for action classification. Yet another type of feature used for human action classification is the 3D human body joint positions in depth videos. [21] incorporated appearance and motion constraints in a graphical model to govern the body joint trajectory clusters. [20] encoded spatio-temporal variations of both depths and depth gradients together with the 3D body joint information.

In this paper, we tackle the human action classification problem for both colour and depth videos. As our aim is to extract robust features directly from the videos, 3D body joint positions that can be extracted from the depth videos are outside the scope of the paper. Similar to the papers reviewed above, we also treat each input video as a space-time volume and compute local features from the volume. However, instead of small 3D cuboids, we divide the volume into hierarchical structures so a video sequence is composed of subsequences, which contain blocks and then cells (Fig. 2). This structure allows the descriptor that describes each action video to retain some spatial information that is essential for human action classification. We found that depth images have a high level of noise and intra-class variation may be large when the same action is performed in different styles by different human subjects. Unlike the work of [42], we use LLC [32] to overcome this problem. Other coding techniques include the super-vector coding [41] and the Fisher vector coding [23] which have been used for image classification (see [4] for a review paper). In our experiments, we also evaluate the performance of our subsequence descriptors using the Fisher vector encoding.

The rest of the paper is organized as follows. Section 3 describes in detail the steps involved in constructing the sequence descriptor for an input video using both the LLC and SC. Section 4 presents the evaluation of these descriptors versus ten state-of-the-art techniques on human action classification, including the classification performance using the Fisher vector encoding. Lastly, Section 5 gives conclusion and future work.

3. Proposed algorithm

We define an action as a function operating on a three dimensional space. The three independent variables in this space are (x, y, t)