Pattern Recognition Letters 49 (2014) 224–230

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Bag-of-words with aggregated temporal pair-wise word co-occurrence for human action recognition ☆

Pau Agustí, V. Javier Traver *, Filiberto Pla

[a] Institute of New Imaging Technologies, Universitat Jaume I, 12071 Castelló de la Plana, Spain
[b] Department of Computer Languages and Systems, Universitat Jaume I, 12071 Castelló de la Plana, Spain

ARTICLE INFO

Article history:
Received 2 September 2013
Available online 29 July 2014

Keywords:
Bag-of-words
Human action recognition
Temporal constraints

ABSTRACT

The bag-of-words (BoW) representation has successfully been used for human action recognition from videos. However, one limitation of the standard BoW is that it ignores spatial and temporal relationships between the visual words. Although several approaches have been proposed to deal with this issue, we propose an extension which is arguably simpler yet quite effective. The proposed representation, *t*-BoW, captures only temporal relationships between pairs of words in an aggregated way by counting co-occurrences at several temporal differences. Unlike other approaches, neither spatial nor hierarchical information is accounted for explicitly, and no significant change is required in the quantization or classification procedures. Performance improvements over the traditional BoW and other BoW extensions are experimentally observed in the KTH, the ADL, the Keck, and the HMDB51 action/gestures datasets.

1. Introduction

Human action recognition from video has been one of the most active research fields in computer vision for the last decade. The main reason is its immediate application in surveillance, human–machine interaction, elderly care, etc. Despite the progress made, the problem is very challenging and, consequently, robust, efficient, general and scalable solutions are elusive.

There are a lot of different approaches to deal with the goal of automating the recognition of human actions, and surveys like [1] provide useful taxonomies. Because of the sequential nature of actions, Hidden Markov Models (HMMs) were initially explored two decades ago for action or gesture recognition [59,40], with many variants, improvements and alternative probabilistic graphical models being proposed since then [15,30,32,33,53–55].

Nevertheless, the bag-of-words (BoW) model is one of the most used approaches lately [23,6,8,4]. Despite its remarkable success, the traditional BoW has some drawbacks. One of them is the heavy computational demand of the clustering procedure (usually, *K*-means) used to build the codebook. To address this issue, trade-offs between accuracy and computational cost have been explored [2]. Another possibility includes dimensionality reduction [46].

The first step for BoW usually consists of detecting spatio-temporal interest (salient) points within the video spatio-temporal volume. This may result in points belonging to noise or background being detected, which somehow contaminates the representation. One possibility to overcome this problem is being more selective and filter unwanted points [11,47].

Our work addresses another well-known limitation of the conventional BoW approach: only orderless occurrences of words are counted, but no spatial or temporal information is considered. However, it has been shown in the context of computer vision problems [3,18] that including this kind of contextual constraints to the BoW model usually leads to higher performance.

In the context of action recognition, the addition of spatio-temporal constraints to the BoW (or other techniques) has also been proposed. Wang et al. [52] provide contextual information for an interest point by considering several spatio-temporal scales and spatio-temporal distances among neighboring interest points. In [29], probability maps capturing spatio-temporal relations between features are studied. Niebles et al. [34] decompose an action sequence into several temporal segments, so that an action model includes information of segments of varying lengths, at several temporal scales, and their temporal order. In [39], relationships of visual words within different neighborhoods (cubic-shaped "kernels") are first computed and then subject to another clustering procedure. Ta et al. [45] concatenate in pairs the (visual + geometric) descriptors of neighboring spatio-temporal interest points, and build two codewords by considering the appearance and geometric components of these pairs separately.

☆ This paper has been recommended for acceptance by G. Borgefors.

* Corresponding author at: Department of Computer Languages and Systems, Universitat Jaume I, 12071 Castelló de la Plana, Spain.

E-mail addresses: pagusti@uji.es (P. Agustí), vtraver@uji.es (V.J. Traver), pla@uji.es (F. Pla).

http://dx.doi.org/10.1016/j.patrec.2014.07.014
0167-8655/© 2014 Elsevier B.V. All rights reserved.

Similarly, Bilinski and Bremond [7] concatenate pairs of visual descriptors but the geometric information is kept implicit by distinguishing several levels of locality.

All these works report some performance improvement in some dataset or scenario. Although some of them are quite general or particularly useful for long-term activities, they include relatively complex procedures. In this work an alternative technique is explored which, unlike some of these approaches:

- includes only *temporal* co-occurrences between words; no spatial distributions nor hierarchical or multiple neighborhoods are explicitly accounted for;
- only counts of word pairs are considered; no distance measure is used to account for either appearance or spatio-temporal similarity between descriptors;
- is based on a single codebook (vocabulary); no additional quantization is required in an additional or subsequent space; and
- the procedure within the standard BoW pipeline is altered minimally; in particular, no additional or alternative learning schemes are required, and only histogram calculation is modified.

The conventional BoW represents a video as the number of occurrences of every visual word in a vocabulary. Thus, for a vocabulary of K words, a video is represented as a histogram of K bins. The proposed representation enriches the BoW by counting word co-occurrences at several temporal differences. Therefore, if L temporal differences are considered, a video would be represented with a histogram of $K \cdot L$ bins, with $K \gg L$ since K is usually in the order of hundreds or thousands, whereas L is in the order of tens. Despite its conceptual and computational simplicity, the proposed algorithm exhibits good performance: as revealed experimentally, it outperforms the classical BoW and even other state-of-art extensions of BoW.

2. Methodology

Under the conventional BoW for action recognition, local spatio-temporal interest points $\{\mathbf{p}_1, \ldots, \mathbf{p}_P\}$ are first extracted from an action video snippet, and then locally described, resulting in their corresponding descriptors $\{\mathbf{d}_1, \ldots, \mathbf{d}_P\}$. These descriptors are then clustered into K visual words $\{w_1, \ldots, w_K\}$ and each word is represented by the centroid of the corresponding cluster. These centroids are used for building *per video* histograms of word counts: each descriptor is assigned to the word of the nearest centroid. We propose an extension of this BoW where this representation is enriched with pairwise word relationship. More specifically, co-occurrences of pairs of words are considered for different temporal differences within a temporal window. This allows the inclusion of temporal information that is ignored in the traditional, orderless BoW. We will refer to this approach as *t*-BoW, because of its temporal information.

2.1. Spatio-temporal interest points detection and description

Spatio-temporal descriptors [6,4,52] have shown to be discriminative enough to describe salient points in videos and recognize human actions. The algorithm used in this work to extract the (salient) interest points (IPs) is the Harris3D detector, proposed by Laptev and Lindeberg [23] as an extension of the Harris corner detector [16]. Once the interest points are detected, each of them is described by a Histogram of Oriented Gradients (HOG) plus an Histogram of Optical Flow (HOF). The HOG is a 72-bin histogram of local appearance, whereas the HOF is a 90-bin histogram accounting for local motion. These are both concatenated into a 162-feature HOG-HOF descriptor, as proposed in [24].

2.2. Extending BoW with temporal information (t-BoW)

The proposed temporal extension consists of counting pairs of co-occurring words at particular time separations. Let $H(w_i, l)$ be the histogram of word $w_i, i \in \{1, \ldots, K\}$ of the initial codebook co-occurring with any other word at $l \in \{1, \ldots, L\}$ time units before or later. In this work we use the frame numbers as the time units, but other time granularities would be possible. This amount of possible pairs can be expressed as

$$H(w_i, l) = \sum_{r=l+1}^{N-l} \sum_{j=1}^{K} h_r(w_i) \cdot \left(h_{r+l}(w_j) + h_{r-l}(w_j)\right), \quad i \in \{1, \ldots, K\},\ l \in \{1, \ldots, L\}, \tag{1}$$

where $h_r(w_i)$ denotes the frequency of the word w_i in the frame r, and N is the number of frames in the video sequence. Expression (1) quantifies the number of possible pairwise combinations between a word w_i in frame r and all possible words w_j that appear in a subsequent frame $r + l$ or in a preceding frame $r - l$ in the video sequence, and can be rewritten as

$$H(w_i, l) = \sum_{r=l+1}^{N-l} h_r(w_i) \cdot S_{r,l}, \tag{2}$$

with

$$S_{r,l} = \sum_{j=1}^{K} h_{r+l}(w_j) + h_{r-l}(w_j). \tag{3}$$

This alternative expression shows more explicitly the idea that word counts at a given frame r are modulated by word counts at a previous frame $r - l$ or a subsequent frame $r + l$. The procedure is illustrated in Fig. 1 where it can be observed that *t*-BoW is not just a weighted version of BoW; an histogram in *t*-BoW has L times more bins than the regular BoW histogram, to account for the L time differences considered.

Notice that the different word pairs combinations are *not* separately kept, but this information is simply aggregated for different time differences. Although the information on particular word pairs co-occurrence is certainly lost, the proposed representation has proven to be rich enough to improve action recognition performance, while still being computationally affordable. The size of the $H(w_i, l)$ histograms is $O(LK)$, with $L \ll K$, whereas a quadratic cost, $O(K^2)$, would be required for a full bi-gram representation, or the even bigger cost, $O(LK^2)$, to account for *all* word pairs at *all* time differences considered. Besides this computational benefit, the temporal pair-wise word aggregation tends to estimate more robustly the underlying distribution of the $H(w_i, l)$ histograms because a larger sample is used to calculate them. Additionally, pair counts are computed both forwards $(h_{r+l}(\cdot))$ and backwards $(h_{r-l}(\cdot))$. This leads to a symmetric direction-less representation that might be arguably less discriminative, but contributes to a smoother and less sparse representation than a naive bi-gram choice.

Once the extended BoW histogram $H(w_i, l)$ is computed for a given video snippet, its conditional probability $P(l|w_i)$ can be estimated as

$$P(l|w_i) = \frac{H(w_i, l)}{\sum_{s=1}^{L} H(w_i, s)}, \quad i \in \{1, \ldots, K\},\ l \in \{1, \ldots, L\}.$$

This probability stands for the possibility that the word w_i is related with any other possible word at a temporal difference of l frames within the video sequence. Thus, the feature vector ($\mathbf{x}$) to