Contents lists available at ScienceDirect

Neurocomputing

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

A novel hierarchical Bag-of-Words model for compact action representation

Qianru Sun [a], Hong Liu [a,*], Liqian Ma [a], Tianwei Zhang [b]

[a] *Key Laboratory of Machine Perception, Shenzhen Graduate School, Peking University, Beijing 100871, China*
[b] *Nakamura Lab, Department of Mechano-Informatics, School of Information Science and Technology, The University of Tokyo, Tokyo 113-8685, Japan*

ARTICLE INFO

Article history:
Received 19 June 2015
Received in revised form
30 July 2015
Accepted 20 September 2015
Communicated by X. Gao
Available online 9 October 2015

Keywords:
Action representation
Bag-of-Words
Vector of Locally Aggregated Descriptors
Fisher Vectors

ABSTRACT

Bag-of-Words (BoW) histogram of local space-time features is very popular for action representation due to its high compactness and robustness. However, its discriminant ability is limited since it only depends on the occurrence statistics of local features. Alternative models such as Vector of Locally Aggregated Descriptors (VLAD) and Fisher Vectors (FV) include more information by aggregating high-dimensional residual vectors, but they suffer from the problem of high dimensionality for final representation. To solve this problem, we novelly propose to compress residual vectors into low-dimensional residual histograms by the simple but efficient BoW quantization. To compensate the information loss of this quantization, we iteratively collect higher-order residual vectors to produce high-order residual histograms. Concatenating these histograms yields a hierarchical BoW (HBoW) model which is not only compact but also informative. In experiments, the performances of HBoW are evaluated on four benchmark datasets: HMDB51, Olympic Sports, UCF Youtube and Hollywood2. Experiment results show that HBoW yields much more compact action representation than VLAD and FV, without sacrificing recognition accuracy. Comparisons with state-of-the-art works confirm its superiority further.

1. Introduction

Human action recognition has shown its significance in a large amount of applications from video surveillance to human-machine interaction [1]. Effective action representation is crucial for high recognition accuracy. Recently, successful representation models are mostly based on local space-time features such as 3D SIFT [2], HOG-HOF [3], 3D Gradients [4], and improved Dense Trajectory (iDT) [5], see [6,7] for more evaluation studies. Once local features are extracted from action videos, typically they are quantized by clustering algorithms to generate a visual codebook, then each of them is assigned to the nearest codeword. The statistic of word assignments yields the Bag-of-Words (BoW) histogram [8,9] which is very compact and robust for human action representation [6,10–13].

To improve BoW, researchers have recently developed many successful alternative models, such as Locality-constrained Linear Coding (LLC) [14], Fisher Vectors (FV) [15,16], Super Vector (SV) encoding [17], kernel codebook encoding [19], and Vector of Linearly Aggregated Descriptors (VLAD) [20,23]. Among them, VLAD and FV show outstanding performances for human action recognition [5,24,25,27–29]. Compared with BoW in Fig. 1, VLAD records the 1st-order difference between local features and codewords, i.e., the residual vectors generated by hard assignment. FV includes not only the 1st-order mean residual vectors but also the 2nd-order covariance residual vectors, which are generated by more complicated soft assignment.[1] Both of them support the potential of using residual information to get more efficient models.

Generally, model efficiency includes the time and storage costs of computing representations, learning classifiers on these representations and recognizing new videos. VLAD and FV are superior to BoW for computing representations. Taking VLAD [20] as an example, it includes high-dimensional information in each codeword, therefore when to reach a given level of performance, a small number of codewords are sufficient for VLAD. The cost of computing VLAD representation is thus greatly lower than that of BoW histogram. However, for *D-dim* local features and *K* codewords, BoW histogram is only *K-dim*, while VLAD representation is *KD-dim* because it aggregates *D-dim* residual vectors in an element-wise manner. Meanwhile, local space-time features are high-dimensional, e.g., $D = 396$ for iDT [5], so $dim_{VLAD} \gg dim_{BoW}$. High-dimensional VLAD representation results in high costs on time and storage for both training classifiers and recognizing new videos, especially on large-scale datasets like HMDB51 [39]. This

* Corresponding author.
E-mail addresses: liuh@szpku.edu.cn, hongliu@pku.edu.cn (H. Liu).

[1] Hard assignment quantizes the feature into the only codeword, while soft assignment enables the feature to be represented by multiple codewords [10].

http://dx.doi.org/10.1016/j.neucom.2015.09.074
0925-2312/© 2015 Elsevier B.V. All rights reserved.

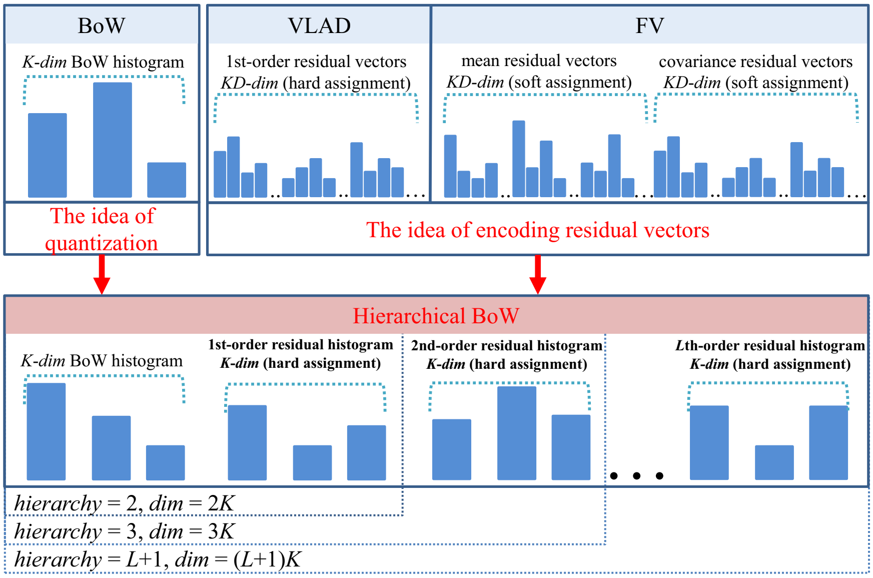

Fig. 1. The basic idea of hierarchical BoW. It is inspired by the quantization idea of BoW and the residual encoding idea of VLAD and FV. The final HBoW histogram is the concatenation of BoW histogram and multiple orders of residual histograms. K is the size of codebook and D is the dimension of local features. Hard assignment means using K-means, while soft assignment means using Gaussian Mixture Models (GMM) for probabilistic assigning weights. Note that HBoW with hierarchy$=1$ is same to BoW.

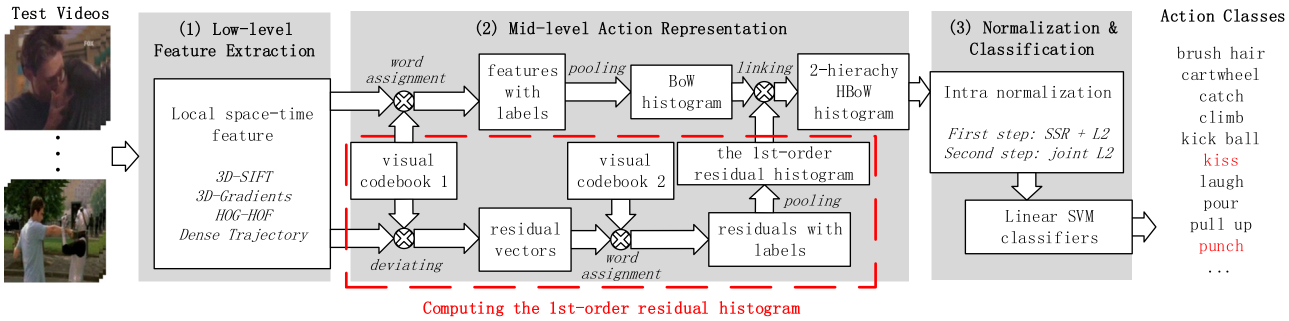

Fig. 2. The flow chart of our action recognition in videos. The 1st-order residual histogram is generated and linked to BoW histogram to form the 2-hierarchy HBoW histogram. Normalized HBoW histogram serves as the final action representation, and linear SVM classifiers are used for recognition. Note that all visual codebooks in (2) are pre-learned offline, and the normalization step in (3) is elaborated in Section 2.4.

problem is much worse for FV representation which has the dimension of $2KD$ [16,25,30].

In this paper, we show that this problem can be effectively solved by the means of BoW, for which we develop a new model called hierarchical BoW (HBoW). Its motivations are illustrated in Fig. 1. Specifically, it aims at (1) compressing residual vectors into compact residual histograms by simple but efficient BoW quantization, and (2) utilizing multiple orders of residual histograms to compensate for the quantization loss and make the final representation, called HBoW histogram, strongly informative for action recognition.

The flow chart of our action recognition method is illustrated in Fig. 2. It contains three main steps: (1) low-level feature extraction, (2) mid-level action representation, (3) normalization and classification. Our proposal of HBoW works in the second step, and an example process of generating the 1st-order residual histogram is given in the red frame of Fig. 2. The innovation lies in that the residual vectors between original local features and codewords are regarded as new features to execute word assignment again. The resulted 1st-order residual histogram is concatenated to BoW histogram to form the 2-hierarchy HBoW histogram. If we iterate this process for L times, then we get $(L+1)$-hierarchy HBoW histogram, see more details in Sections 2.2 and 2.3. If all codebooks are assumed to have K codewords, each iteration produces a K-dim residual histogram. The dimension of $(L+1)-$hierarchy HBoW histogram is therefore $K(L+1)$, which is significantly smaller than KD since $L+1 \ll D$. Then, using low-dimensional HBoW histogram for action representation saves a lot of time and storage for training classifiers and recognizing new videos.

In summary, HBoW is inherently derived from iterative BoW quantization with high-order residual vectors. It has two advantages that (1) It shares high compactness and efficiency of the original BoW; (2) It yields strongly discriminative representation by using high-order statistic information.

1.1. Related works

As we discussed, BoW model with local features has become very popular for visual understanding researches, such as image classification and object/action recognition. Alternative encoding models [14,17,18,20–22] based on BoW framework obtain the state-of-the-art performances in many visual tasks.

Wang et al. [31] and Peng et al. [10] evaluated most of these models for human action recognition, and observed that FV encoding outperforms others. FV combines the benefits of generative and discriminative approaches, and usually leverages Gaussian Mixture Model (GMM) as its dictionary. Wang et al. [5] adopted FV encoding with iDT features, and obtained generally good results on frequently-used action datasets, e.g., Olympic Sports [42], UCF Youtube [33] and