Computer Vision and Image Understanding 000 (2016) 1–17

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice

Xiaojiang Peng[a,d,c,*], Limin Wang[b,c], Xingxing Wang[c], Yu Qiao[c]

[a] College of Computer Science and Technology, Hengyang Normal University, Hengyang, China
[b] Computer Vision Lab, ETH Zurich, Zurich, Switzerland
[c] Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China
[d] LEAR Team, INRIA, Grenoble, France

ARTICLE INFO

Article history:
Received 21 September 2015
Revised 8 January 2016
Accepted 21 March 2016
Available online xxx

Keywords:
Action recognition
Bag of visual words
Fusion methods
Feature encoding

ABSTRACT

Video based action recognition is one of the important and challenging problems in computer vision research. Bag of visual words model (BoVW) with local features has been very popular for a long time and obtained the state-of-the-art performance on several realistic datasets, such as the HMDB51, UCF50, and UCF101. BoVW is a general pipeline to construct a global representation from local features, which is mainly composed of five steps; (i) feature extraction, (ii) feature pre-processing, (iii) codebook generation, (iv) feature encoding, and (v) pooling and normalization. Although many efforts have been made in each step independently in different scenarios, their effects on action recognition are still unknown. Meanwhile, video data exhibits different views of visual patterns , such as static appearance and motion dynamics. Multiple descriptors are usually extracted to represent these different views. Fusing these descriptors is crucial for boosting the final performance of an action recognition system. This paper aims to provide a comprehensive study of all steps in BoVW and different fusion methods, and uncover some good practices to produce a state-of-the-art action recognition system. Specifically, we explore two kinds of local features, ten kinds of encoding methods, eight kinds of pooling and normalization strategies, and three kinds of fusion methods. We conclude that every step is crucial for contributing to the final recognition rate and improper choice in one of the steps may counteract the performance improvement of other steps. Furthermore, based on our comprehensive study, we propose a simple yet effective representation, called *hybrid supervector*, by exploring the complementarity of different BoVW frameworks with improved dense trajectories. Using this representation, we obtain impressive results on the three challenging datasets; HMDB51 (61.9%), UCF50 (92.3%), and UCF101 (87.9%).

1. Introduction

Human action recognition (Aggarwal and Ryoo, 2011; Turaga et al., 2008) has become an important area in computer vision research, whose aim is to automatically classify the action ongoing in a temporally segmented video. It is one of the challenging problems in computer vision for several reasons. Firstly, there are large intra-class variations in the same action class, caused by various motion speeds, viewpoint changes, and background clutter. Secondly, the identification of an action class is related to many other high-level visual clues, such as human pose, interacting objects, and scene class. These related problems are very difficult themselves. Furthermore, although videos are temporally segmented, the segmentation of an action is more subjective than a static object, which means that there is no precise definition about when an action starts and finishes. Finally, the high dimension and low quality of video data usually adds difficulty to develop robust and efficient recognition algorithms.

Early approaches interpret an action as a set of space-time trajectories of two-dimensional or three-dimensional points of human joints (Campbell and Bobick, 1995; Niyogi and Adelson, 1994; Webb and Aggarwal, 1981; Yacoob and Black, 1999). These methods usually need dedicated techniques to detect body parts or track them at each frame. However, the detection and tracking of body part is still an unsolved problem in realistic videos. Recently, local spatiotemporal features (Laptev, 2005; Laptev et al., 2008; Wang et al., 2013a, 2014) with the follow-mentioned bag-of-visual-words pipeline have become the main stream and obtained the state-of-the-art performance on many datasets (Wang and Schmid, 2013a). These methods do not require algorithms to detect human

* Corresponding author. Tel.: +330763117277.
E-mail address: xiaojiang.peng@inria.fr (X. Peng).

http://dx.doi.org/10.1016/j.cviu.2016.03.013
1077-3142/

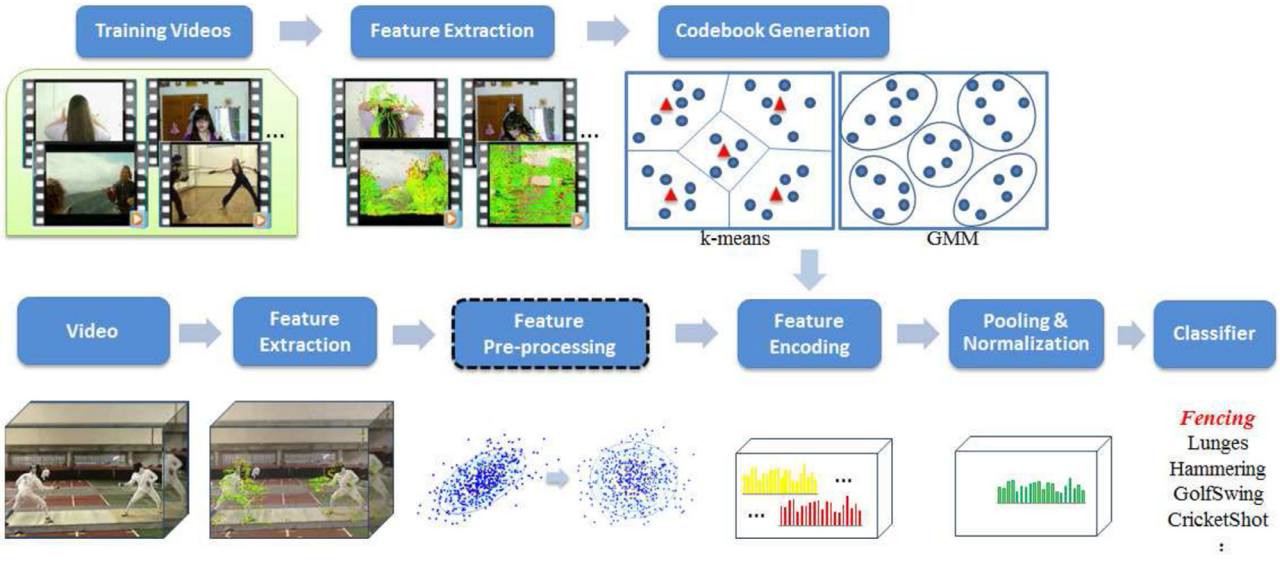

Fig. 1. The pipeline of obtaining BoVWs representation for action recognition. It is mainly composed of five steps; (i) feature extraction, (ii) feature pre-processing, (iii) codebook generation, (iv) feature encoding, and (v) pooling and normalization.

bodies, and are robust to background clutter, illumination changes, and noise.

More recently, with the progress of pose estimation (Yang and Ramanan, 2011) and deep learning (Krizhevsky et al., 2012; Simonyan and Zisserman, 2014), several works focus on how to combine local features with high-level information (e.g., pose information) and learned features. Xu et al. (2012) applied a popular pose estimator (Yang and Ramanan, 2011) and extracted HoG3D features (Klaser et al., 2008) based on the detected poses. Simonyan et al. designed the well-known two-stream convolutional networks based on raw RGB frames and pre-computed optical flows. Wang et al. (2015) combined the two-stream convolutional networks with dense trajectories, Wang et al. (2013a). Chéron et al. (2015) proposed P-CNN (Pose Convolutional Neural Networks) which extracts CNN features based on poses.

BoVW framework and its variants (Karaman et al., 2013; Murthy and Goecke, 2013; Peng et al., 2013; Wang and Schmid, 2013b; Wu, 2013) have dominated the research work of action recognition for a long time. It is necessary to overview the details and uncover the good practice of each step in BoVW pipeline for beginners or other researchers. As shown in Fig. 1, the pipeline of BoVW for video based action recognition consists of five steps; (i) feature extraction, (ii) feature pre-processing, (iii) codebook generation, (iv) feature encoding, and (v) pooling and normalization. Regarding local features, many successful feature extractors (e.g. STIPs (Laptev, 2005), Dense Trajectories (Wang et al., 2013a)) and descriptors (e.g. HOG (Laptev et al., 2008), HOF (Laptev et al., 2008), MBH (Wang et al., 2013a)) have been designed for representing the visual patterns of cuboid. Feature pre-processing technique mainly de-correlates these descriptors to make the following representation learning more stable. For codebook generation, it aims to describe the local feature space and provide a partition (e.g. *k*-means (Bishop, 2006)) or generative process (e.g. GMMs (Bishop, 2006)) for local descriptor. Feature encoding is a hot topic in image classification and many alternatives have been developed for effective representation and efficient implementation (see good surveys Chatfield et al. (2011) and Huang et al. (2014)). Max pooling (Yang et al., 2009) and sum pooling (Zhang et al., 2007) are usually used to aggregate information from a spatiotemporal region. For normalization methods, typical choices include ℓ_1-normalization (Zhang et al., 2007), ℓ_2-normalization (Wang et al., 2010), power normalization (Perronnin et al., 2010), and intra normalization (Arandjelovic and Zisserman, 2013). *How to make the best decision in each step for action recognition* still remains unknown and needs to be extensively explored.

Meanwhile, unlike static image, video data exhibits different views of visual pattern, such as appearance, motion, and motion boundary, and all of them play important roles in action recognition. Therefore, multiple descriptors are usually extracted from a cuboid and each descriptor corresponds to a specific aspect of the visual data (Laptev et al., 2008; Wang et al., 2013a). BoVW is mainly designed for a single descriptor and ignores the problem of fusing multiple descriptors. Many research studies have been devoted to fusing multiple descriptor for boosting performance (Cai et al., 2014; Gehler and Nowozin, 2009; Tang et al., 2013; Vedaldi et al., 2009a; Wang and Schmid, 2013a). Typical fusion methods include descriptor level fusion (Laptev et al., 2008; Wang et al., 2012), representation level fusion (Wang et al., 2013a; Wang and Schmid, 2013b), and score level fusion (Myers et al., 2014; Tang et al., 2013). For descriptor level fusion, multiple descriptors from the same cuboid are concatenated as a whole one and fed into a BoVW framework. For representation level fusion, the fusion is conducted in the video level, where each descriptor is firstly fed into a BoVW framework independently and the resulting global representations are then concatenated to train a final classifier. For score level fusion, each descriptor is separately input into a BoVW framework and used to train a recognition classifier. Then the scores from multiple classifiers are fused using arithmetic mean or geometric mean. In general, these fusion methods are developed in different scenarios and adapted for action recognition by different works. *How these fusion methods influence the final recognition of a BoVW framework and whether there exists a best one for action recognition* is an interesting question and well worth of a detailed investigation.

Several related study works have been performed about encoding methods for image classification (Chatfield et al., 2011; Huang et al., 2014). But these study works are with image classification task or lacking full exploration of all steps in BoVW framework. This paper is an extension of our previous work (Wang et al., 2012). We extend (Wang et al., 2012) from the following aspects:

- We explore pre-processing step for all the encoding methods (not only for Fisher vectors as Wang et al. (2012)).

Please cite this article as: X. Peng et al., Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice, Computer Vision and Image Understanding (2016), http://dx.doi.org/10.1016/j.cviu.2016.03.013