Expert Systems With Applications 78 (2017) 259–272

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Mining human movement evolution for complex action recognition

Yang Yi[a,b,1], Yang Cheng[a,*], Chuping Xu[a]

[a] *School of Data & Computer Science, Sun Yat-sen University, Guangzhou, China*
[b] *Xinhua College of Sun Yat-sen University, Guangzhou, China*

ARTICLE INFO

Article history:
Received 9 November 2016
Revised 5 January 2017
Accepted 9 February 2017
Available online 10 February 2017

Keywords:
Action recognition
Dense trajectory
Motion compensation
Feature representation
Hierarchical encoding

ABSTRACT

In this paper, a novel and efficient system is proposed to capture human movement evolution for complex action recognition. First, camera movement compensation is introduced to extract foreground object movement. Secondly, a mid-level feature representation called trajectory sheaf is proposed to capture the temporal structural information among low-level trajectory features based on key frames selection. Thirdly, the final video representation is obtained by training a sorting model with each key frame in the video clip. At last, the hierarchical version of video representation is proposed to describe the entire video with higher level representation. Experimental results demonstrate that the proposed method achieves state-of-the-art performance on UCF Sports, and comparable results on several challenge benchmarks, such as Hollywood2 and HMDB51 dataset.

1. Introduction

It has been noticed that video analysis has attracted increasing interest due to the exponential growth of video data over the recent years. The researches mostly focus on human action recognition in consideration of many relevant real-life applications, such as smartphone sensors (Ronao & Cho, 2016), video surveillance system (Kim et al., 2016), assisted living (Chaaraoui, Climent-Pérez, & Flórez-Revuelta, 2012; Olivieri, Gómez Conde, & Sobrino Vila, 2012), video retrieval (Gómez-Conde & Olivieri, 2015), and smart home applications (Banos, Damas, Pomares, Prieto, & Rojas, 2012; Diraco, Leone, & Siciliano, 2013; Wen, Zhong, & Wang, 2015).

The goal of human action recognition is to classify the unlabeled video clips into basic and complex human activities such as walking, boxing, and hand waving. Unlike videos in constrained environment, videos sourced in the wild contain massive background movements, which make the task of human action recognition more complicated. Furthermore, there are wide variations in appearance and motion within an action class due to motion speed variation.

1.1. *Background and motivation*

In the field of human action recognition, a multitude of remarkable works and approaches have been proposed. Over the last decades, most of the existing works has been related to (i) the development of local spatio-temporal descriptors (e.g. space-time interest points (Laptev, 2005), Dollar interest points (Dollár, Rabaud, Cottrell, & Belongie, 2005), spatiotemporal Hessian detector (Willems, Tuytelaars, & Van Gool, 2008), and dense sampling strategy (Wang, Kläser, Schmid, & Liu, 2013), (ii) the adoption of powerful encoding schemes with an already proven track record in object recognition (e.g. Fisher Vectors (FV) (Perronnin, Sánchez, & Mensink, 2010), Locality-constrained Linear Coding (LLC) (Wang et al., 2010)), (iii) the introduction of action classification models (e.g. Linear Discriminant Analysis (LDA) (Fisher, 1936), K-Nearest Neighbor (K-NN) (Cover & Hart, 1967), Support Vector Machine (SVM) (Vapnik, 1999), deep neural network (Simonyan & Zisserman, 2014), and Extreme Learning Machine (ELM) (Varol & Salah, 2015)). Fig. 1 shows the general flowchart of human action recognition. We refer the readers to the recent surveys (Onofri, Soda, Pechenizkiy, and Iannello, 2016 and Peng, Wang, Wang, and Qiao, 2015) for a detailed review of spatio-temporal feature descriptors, video encoding techniques and classification approaches for action recognition.

Apparently, all these stages of action recognition are important, but we want to emphasize the importance of obtaining reliable foreground object movement from a video containing immense background motions to improve the robustness of action recognition systems.

1.1.1. *Camera motion compensation*

Foreground object movement extraction is a vital step in the standard action recognition pipeline adopted in most of previous works. In particular, foreground-background separation based on

[*] Corresponding author.
E-mail addresses: issyy@mail.sysu.edu.cn (Y. Yi), chengy35@mail2.sysu.edu.cn (Y. Cheng), xuchp3@mail2.sysu.edu.cn (C. Xu).
[1] She is also a research fellow of Guangdong Province Key Laboratory of Big Data Analysis and Processing.

http://dx.doi.org/10.1016/j.eswa.2017.02.020
0957-4174/© 2017 Elsevier Ltd. All rights reserved.

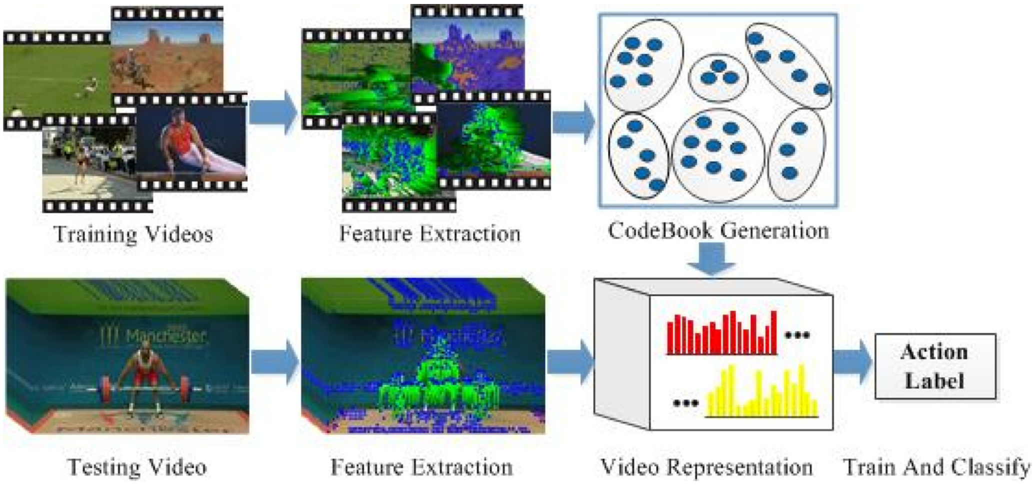

Fig. 1. The flowchart of human action recognition computation. Human action recognition consists of feature extraction, codebook generation, feature representation, as well as action classification.

camera motion compensation is one of the most efficient methods to get reliable foreground object movement. We can benefit from explicit motion compensation because camera motion generates many irrelevant background movements in realistic video clips while camera motion compensation can eliminate the interference to some extends. Yilmaz and Shah (2005), claiming to be the first one to deal with camera motion, exploits multi-view geometry in the field of action recognition. However, this solution requires very similar actions captured from different viewpoints or multiple camera setup, which limits the potential capability of action recognition. Jiang, Dai, Xue, Liu, and Ngo (2012) combines dense trajectory features (Wang, Kläser, Schmid, & Liu, 2011) and an extended BoV (Bag of Visual words) encoding method over pairs of local features to explicitly cancel dominant (camera) motion patterns which may discard the temporal information between local features.

1.1.2. Key frames selection

Key frames selection is one of the most important techniques in video summarization, browsing, searching and recognition. Discriminative approaches to identify key frames have also been used in last decades. Early methods try to group frames with similar features and select the frame closest to each cluster centroid as a key frame (Zhuang, Rui, Huang, & Mehrotra, 1998). Zhao and Elgammal (2008) select the key frames by their discriminative power and represent them by the local motion features detected in them and their temporal neighbors. However, those methods are not effective enough, because they are inconsistent with human visual perception. In other words, there is a gap between semantic interpretation of the video and its low-level features (color, contexture, and etc.). Raptis and Sigal (2013) consider key frames as latent variables and cast the learning of key frames in a max-margin discriminative framework to jointly learn a set of most discriminative key frames while also learning the local temporal context between them. However, this method relay heavily on the performance of spatio-temporal localization. While Alfaro, Mery, and Soto (2016) introduce a loss function to identify a sparse set of representative key frames capturing both, relevant particularities arising in the input video, as well as relevant generalities arising in the complete class collection. As a relevant advantage, their method to select key frames explicitly focuses on an effective mining of relevant intra-class variations. Lai and Yi (2012) extract the most attractive key frames based on saliency-based visual attention model. In the light of Lai & Yi's work (Lai & Yi, 2012), a novel key frames selection method based on visual attention and saliency detection is introduced to select the most attractive key frames.

1.1.3. Mid-level feature representation

Although low-level features perform well in earlier action recognition problems, they directly observe visual appearance and local motion feature vectors as video representation while temporal information is ignored. In the literature, mid-level feature representation exploits temporal information of video clips using different techniques. Raptis, Kokkinos, and Soatto (2012) propose mid-level features representation by forming clusters of dense trajectories that serve as candidates for action parts, and the pairwise relations between the parts is conveyed by a graphical model. Yuan, Xia, Sahbi, and Prinet (2012) hierarchically extract features from a video to comprise a set of mid-level components that contains consistent structure and motion information in spatial and temporal domain. Song, Morency, and Davis (2013) exploits a series of complex heuristics to present a hierarchical sequence temporal summarization method learning multiple layers of discriminative feature representations at different temporal granularities. While Wang, Qiao, and Tang (2013a) presents a mid-level method for video representation named motionlet which greedily select effective motionlet candidates that clustered by 3D regions with high motion saliency. Considering that the area occupied by human body parts in a video frame provides complementary information for action recognition, Varol and Salah (2015) concatenate 6 types of mid-level features to encode information about presence of humans in the videos, as well as color distributions. Yi and Lin (2016) propose a new trajectory clustering algorithm based on trajectory spectral embedding and density discontinuity detection. They decompose an action into a collection of semantically salient spatio-temporal (i.e. two spatial dimensions and one temporal dimension, 2D+t) action parts, in order to construct the mid-level feature representation for action videos.

1.1.4. Hierarchical structural representation

Hierarchical structural representation method is popular in action recognition field due to its capability in capturing the multi-level granularity of human action. Kostavelis and Gasteratos (2012) optimize the original Hierarchical Temporal Memory (HTM) method which stems from the memory prediction theory of the human brain. Their HTM comprises two different modules, viz. the spatial and the temporal module, which comply with the human vision system. Inspired by the HTM notion, Charalampous and Gasteratos (2014) propose an unsupervised on-line deep learning