Pattern Recognition Letters 112 (2018) 161–167

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

First Person Action Recognition via Two-stream ConvNet with Long-term Fusion Pooling

Heeseung Kwon[a], Yeonho Kim[b], Jin S. Lee[a], Minsu Cho[b,*]

[a] Department of Creative IT Engineering, POSTECH, 77 Cheongam-Ro, Nam-Gu, Pohang 37673, South Korea
[b] Department of Computer Science and Engineering, POSTECH, 77 Cheongam-Ro, Nam-Gu, Pohang 37673, South Korea

ARTICLE INFO

Article history:
Received 6 January 2018
Available online 7 July 2018

MSC:
41A05
41A10
65D05
65D17

Keywords:
First person action recognition
Action recognition
Convolutional neural network
Feature pooling

ABSTRACT

First person action recognition is an active research area with increasingly popular wearable devices. Action classification for first person video (FPV) is more challenging than conventional action classification due to strong egocentric motions, frequent changes of viewpoints, and diverse global motion patterns. To tackle these challenges, we introduce a two-stream convolutional neural network that improves action recognition via long-term fusion pooling operators. The proposed method effectively captures the temporal structure of actions by leveraging a series of frame-wise features of both appearance and motion in actions. Our experiments validate the effect of the feature pooling operators, and show that the proposed method achieves state-of-the-art performance on standard action datasets.

1. Introduction

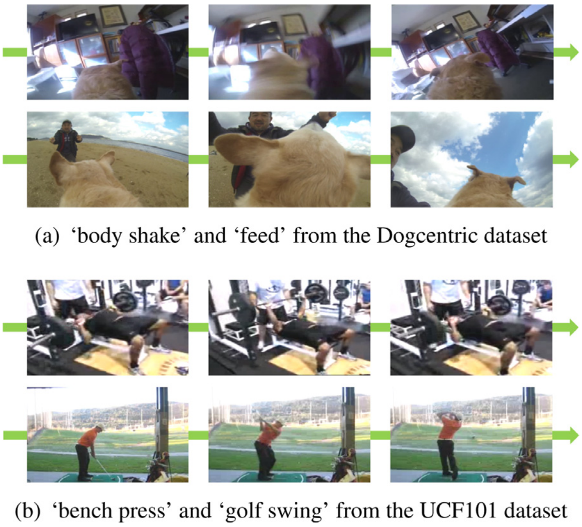
As wearable devices become more popular and diverse, there is a flood of video contents with first person view these days. In the past few years, people have been familiar with using handheld mobile cameras, action cameras with head or body mounted (e.g. GoPro). Increasing attention of wearable cameras has been drawn to research on first person video (FPV) such as video summarization [1–3], environmental mapping [4–6], and action recognition [7–9]. In particular, action recognition for first person videos is linked to numerous real-world applications. Contents of first person videos usually consist of life logging, sports, and surveillance which treat a large amount of data. Action recognition can be used to retrieve and edit certain events in these videos. In particular, first person action recognition would play a considerable part for augmented reality applications to provide context-aware information. Furthermore, it can be extended to automobile, drone and robot applications in the near future.

While action recognition is a widely studied topic in computer vision, first person action recognition is more challenging than conventional one for several reasons. First, first person actions contain strong egocentric motions causing severe motion blur and motion parallax (e.g. ‘body shake’ in Fig. 1(a)). Second, frequent and abrupt viewpoint changes in first person videos make it difficult to capture proper features from video frames (e.g. ‘feed’ in Fig. 1(a)). Third, unlike actions in conventional videos, actions in first person videos often involve diverse global motion patterns that become a critical factor for recognizing ambulatory actions such as walking, running, and jumping. To tackle such challenges, it is important to obtain features that reflect useful characteristics of first person actions. Appearance and motion features, extracted frame by frame, are important to recognize actions that contains particular objects or severe egocentric motions. However, they are not enough to analyze actions with frequent camera viewpoint changes or global motion patterns. In this paper, we introduce a two-stream convolutional neural network that effectively captures the temporal structure of actions via long-term fusion pooling operators. Our experiments validate the effect of the feature pooling operators, and show that the proposed method achieves state-of-the-art performance on standard action datasets.

The paper is organized as follows. Section 2 reviews related work and the proposed architecture is described in Section 3. Experimental results are presented in Section 4, and Section 5 concludes with a summary.

[*] Corresponding author.
E-mail addresses: aruno@postech.ac.kr (H. Kwon), beast@postech.ac.kr (Y. Kim), jsoo@postech.ac.kr (J.S. Lee), mscho@postech.ac.kr (M. Cho).

https://doi.org/10.1016/j.patrec.2018.07.011
0167-8655/© 2018 Elsevier B.V. All rights reserved.


(a) ‘body shake’ and ‘feed’ from the Dogcentric dataset

(b) ‘bench press’ and ‘golf swing’ from the UCF101 dataset

Fig. 1. Example frames of actions from the Dogcentric dataset [10] and the UCF101 dataset [11]. First person actions in (a) involve more severe egocentric motions, frequent viewpoint changes, and global motion patterns compared to conventional actions in (b).

2. Related work

Conventional action recognition: Most previous approaches have used hand-crafted features for conventional action recognition. For example, histogram of optical flow (HOF) and motion boundary histogram (MBH) are among the most successful handcrafted features for action recognition [12,13]. The improved dense trajectories (IDT) [14] is also widely used to extract short-term motion features. Recently, motivated by the success of Convolutional Neural Network (ConvNet) [15], many ConvNet-based methods have been proposed [16,17]. Ji et al. [16] design 3D ConvNets for action recognition to jointly capture spatial and temporal information. Simonyan and Zisserman [17] propose a two-stream ConvNet model that takes RGB images and stacked optical flow images as the network input. This network has outperformed handcrafted features in action recognition performance for the first time.

Since then, the two-stream ConvNet model has been popularized and used as a base architecture for action recognition [18–20]. Wu et al. [20] combine a two-stream ConvNet model with long short term memory (LSTM) to capture long-term temporal information. Wang et al. [18] propose novel features named trajectory pooled descriptors (TDD), which integrates the advantages of the two-stream ConvNet and IDT. Feichtenhofer et al. [21] analyze the two-stream ConvNet and suggest a novel two-stream ConvNet with convolutional fusion of streams. Wang et al. [22] successfully apply the idea of two-stream ConvNets to a deeper architecture and improve the performance. A few recent methods extend a two-stream ConvNet model with temporal encoding to analyze long-term temporal information [23,24]. Diba et al. [23] propose temporal linear encoding (TLE) with a two-stream ConvNet model, and Lan et al. [24] propose deep local video feature (DOVF) with a two-stream ConvNet model.

First person action recognition: First person action recognition methods can be divided roughly into two groups. The first group designs features using particular objects or short motions appearing in video sequences. Fathi et al. [7] exploit objects and hand movements for first person action recognition. Li et al. [25] use combination of diverse mid-level features from object, motion, hand pose, and head movement. A few recent methods propose ConvNet models to obtain deep features and increase the recognition performance [9,26]. Ma et al. [9] propose a novel two-stream ConvNet model that obtains advanced appearance information using hand segmentation and object localization. Singh et al. [26] design a novel ConvNet that uses hand mask, head motion, saliency map images as additional network inputs. The second group designs a feature representation that focuses on temporal encoding of frame-wise features. Ryoo et al. [27] propose a feature pooling method (PoT) that outperforms bag of features (BoF) and improved fisher vector (IFV), which motivates our pooling scheme. Piergiovanni et al. [28] suggest a feature representation using temporal attention filters (TAF) and LSTM. Zaki et al. [29] propose a sub-event modeling method with temporal pooling and encoding.

While leveraging both appearance and motion features in egocentric scenes by adopting a two-stream ConvNet, we propose a novel feature representation that effectively captures temporal structures of actions by long-term fusion pooling. The long-term fusion pooling has an advantage in reflecting feature variation across frames. The main contributions of this paper can be summarized as follows.

- We propose a two-stream ConvNet architecture with long-term fusion pooling that effectively captures temporal structures of actions with both appearance and motion.
- We introduce novel long-term pooling operators for appearance and motion information and analyze their effects compared to existing pooling operators.
- We evaluate our method on standard action recognition benchmarks and achieve the state-of-the-art performance on the Dogcentric dataset [10]

3. Methodology

3.1. Overall architecture

The proposed architecture (Fig. 2) consists of three parts: two-stream ConvNet, long-term fusion pooling, and feature classifier. (1) The two-stream ConvNet [17] is employed to extract both appearance and motion information from each frame. The extracted frame-wise features are stacked for each video segment. (2) The long-term fusion pooling layer aggregates appearance and motion features, and captures temporal structures of actions. Specifically, assuming N is the frame number of a video segment and D is the frame-wise feature dimension, we aggregate N by D frame-wise features to a D dimensional feature by the long-term fusion pooling layer. (3) The feature classification layer performs action classification using the global video representation that combines long-term fusion features.

3.2. Two-stream ConvNet

We use the two-stream ConvNet model [17] as our base architecture. The two-stream ConvNet is composed of an appearance stream and a motion stream. While the appearance stream generates appearance features from RGB images as in conventional networks, the motion stream generates motion features from stacked optical flows. Motion stream may play an important role in discriminating first person actions by recognizing motion patterns on different sizes and intensities. We choose the Inception network with Batch Normalization (BN-Inception) [30] as a backbone network for the two-stream ConvNet, which is deeper than that of the original two-stream model [17] Both streams use a cross entropy loss with softmax as a loss function. We extract the frame-wise features from the last convolution layer, which is the global average pooling layer of BN-Inception [30]. The last convolution layer may preserve spatial information better than the fully connected layers, and contain higher-level feature information than the other convolution layers [24].