Neurocomputing 267 (2017) 475–488

Contents lists available at ScienceDirect

Neurocomputing

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

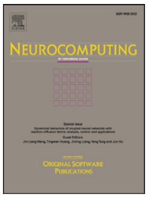

Image-based action recognition using hint-enhanced deep neural networks☆

Tangquan Qi[a], Yong Xu[a,*], Yuhui Quan[a], Yaodong Wang[a], Haibin Ling[a,b]

[a] School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, China
[b] Center for Information Science and Technology, Computer and Information Science Department, Temple University, Philadelphia, PA, USA

ARTICLE INFO

Article history:
Received 22 August 2016
Revised 15 June 2017
Accepted 20 June 2017
Available online 29 June 2017

Communicated by Jiwen Lu

Keywords:
Action recognition
Pose hints
Convolutional neural networks

ABSTRACT

While human action recognition from still images finds wide applications in computer vision, it remains a very challenging problem. Compared with video-based ones, image-based action representation and recognition are impossible to access the motion cues of action, which largely increases the difficulties in dealing with pose variances and cluttered backgrounds. Motivated by the recent success of convolutional neural networks (CNN) in learning discriminative features from objects in the presence of variations and backgrounds, in this paper, we investigate the potentials of CNN in image-based action recognition. A new action recognition method is proposed by implicitly integrating pose hints into the CNN framework, i.e., we use a CNN originally learned for object recognition as a base network and then transfer it to action recognition by training the base network jointly with inference of poses. Such a joint training scheme can guide the network towards pose inference and meanwhile prevent the unrelated knowledge inherited from the base network. For further performance improvement, the training data is augmented by enriching the pose-related samples. The experimental results on three benchmark datasets have demonstrated the effectiveness of our method.

1. Introduction

Human action recognition aims at recognizing human actions in videos or still images, which is an active topic in computer vision and has a wide range of applications, such as surveillance and human computer interaction [1–5]. Despite of the efforts made in the past decades, action recognition remains a very challenging task, where the difficulties arise from the cluttered backgrounds, human pose variations, occlusions, illumination changes, and appearance changes in videos. Such difficulties are aggravated for still images, as the motion cues, which play important roles in expressing human actions in videos [6–10], are completely lost in the images. See Fig. 1 for an illustration of the difficulties in image-based action recognition.

1.1. Motivation

To address the aforementioned challenges, we use the-state-of-art deep learning model, convolutional neural network (CNN), to deal with action recognition. Our motivation is that CNN has shown its success in learning discriminative features from objects, even in the presence of cluttered backgrounds or large variations in the appearances and poses of objects. However, traditional CNNs cannot be directly applied to action recognition due to two obstacles:

- Data insufficiency. It is well known that CNN need to be trained on a huge number of images for satisfactory performance. Nevertheless, unlike object recognition, most existing action datasets like Stanford-40 contain a limited number of training images.
- Overfitting. A simple CNN used for action recognition is likely to overfit the appearance of objects as it is not equipped with any prior on human action. For instance, an overfitting CNN might distinguish the action of playing volleyball only via detecting the volleyball.

☆ Yuhui Quan would like to thank the support by National Natural Science Foundation of China (Grant no. 61602184), Science and Technology Program of Guangzhou (Grant no. 201707010147), and Educational Reform Project of South China University of Technology (j2jwY9160960). Yong Xu would like to thank the support by the National Natural Science Foundation of China (U1611461616722416160218 4 and 61528204), the Cultivation Project of Major Basic Research of NSF-Guangdong Province 2016A030308013 and Science and Technology Program of Guangzhou (201707010147).

* Corresponding author.

E-mail addresses: qi.tangquan@mail.scut.edu.cn (T. Qi), yxu@scut.edu.cn (Y. Xu), yuhui.quan@scut.edu.cn (Y. Quan), w.yaodong@mail.scut.edu.cn (Y. Wang), haibin.ling@gmail.com (H. Ling).

http://dx.doi.org/10.1016/j.neucom.2017.06.041
0925-2312/© 2017 Elsevier B.V. All rights reserved.

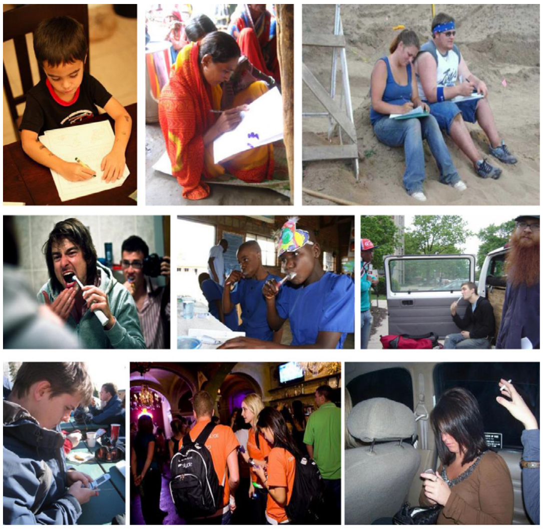

Fig. 1. Examples of action images in the Stanford-40 dataset. It can be observed that human performing the same action may look very different, and cluttered backgrounds, human pose variations, occlusions, illumination changes and appearance changes are often presented in the images. Our task is to recognize the actions of the people in the images.

To deal with the problem of data insufficiency, in this paper, we investigate the transfer of CNN from object recognition to action recognition and design an effective data augmentation scheme. This work is inspired by the fact that the training dataset for object recognition is significantly more than that of action recognition. However, the CNN learned from objects emphasizes the appearance of objects and thus using such a CNN as the base network in transfer is likely to aggravate the aforementioned overfitting problem. To alleviate the overfitting, we resort to the hints given by poses, which are very important for recognizing actions, and then we develop a hint-enhanced CNN that can simultaneously and effectively utilize the hints from both the appearance and pose for action recognition from still images.

1.2. Contribution

In this paper, we develop a new CNN for utilizing pose hints in action recognition from still images, which incorporates a task of pose inference into the base CNN that originates from object recognition. By exploiting the pose hints, the proposed CNN can encode pose cues for action recognition and reduce the unrelated knowledge inherited from the base network. To improve the performance of the transferred network, we augment the data with a pose-sensitive sampling strategy, where image patches are cropped within or around the human bounding box and then used as samples. We evaluated the performance of the proposed method on three widely-used benchmark datasets, including the Stanford-40 Actions dataset [11], the PPMI dataset [12] and the VOC 2012 Actions Dataset [13]. The results show the effectiveness of the proposed method as well as its superior performance to the base network.

1.3. Organization

The rest of this paper is organized as follows. The related work is described in Section 2.1. In Section 3, we present the details of the proposed method. The experimental results are discussed in Section 5, and Section 6 concludes the paper.

2. Preliminaries

2.1. Related work

Many existing methods for action recognition extract high-level action representations by exploiting the cues from human–object interaction or human poses. The methods based on human–object interaction usually describe the relative position, the relative size, as well as the overlap between the person and object. See [12,14] for the examples of such kind of methods.

Human pose can be viewed as the spatial configuration of body parts, which is discriminative to a broad spectrum of actions. For instance, using computer and climbing can be distinguished by the hand poses. The part-based methods (e.g. [15–19]) are one of most effective methods built upon human pose, which describe a human pose by the corresponding parts of human body. Yang et al. [17] build up a graphical model to represent the relations between different body parts, including the upper-body, legs, left arm and right arm. In the similar spirit, Yao et al. [18] proposed a 2.5D graph for the representation of action, where the nodes corresponding to the key-points of the human body are represented based on their 3D positions and 2D appearances. Then a minimum set of dominating images is selected to cover all possible cases for each action class. A recent popular part-based method is the so-called *poselets* [20] method. Briefly speaking, a poselet is a detector for some specific body part, which in fact is a linear SVM trained on the clustered image patches that share a salient pattern of local pose from a viewpoint. Based on the poselets, Maji et al. [16] proposed the poselet activation vector (PAV) for representing human actions, which calculates the distribution over the poselets.

In recent years, inspired by the success of deep learning in a wide range of applications, many researchers have started to investigate the application of convolutional neural network (CNN) to action recognition. Oquab et al. [21] used an 8-layer CNN for action classification. Hoai [22] proposed an effective pooling method for CNN in action recognition based on a geometrical distribution of regions placed in bounding boxes of images. Gkioxarie et al. [23] trained body part detectors on ‘Pool5’ features in a sliding-window manner and combined them with the bounding boxes to jointly train a CNN. They also applied contextual cues to build an action recognition system [24]. Simonyan and Zisserman [25] combined a 16-layer CNN with a 19-layer CNN and trained multiple linear SVMs on ‘FC7’ features from images with bounding boxes.

Different from the above methods [21,22,25], we introduce poselets into CNN by integrating an auxiliary pose-inference task into the the training of CNN, which is for learning features related to human poses. One closely-related work to ours is the [23] which used deep version of poselets to capture parts of the human body under a distinct set of poses. The difference between our method and [23] is that, poses are directly used as features in [23], while they are indirectly utilized as hints in our method to regularize the network.

2.2. Convolutional neural network

The framework of convolutional neural network (CNN) is first introduced by LeCun et al. [26] with impressive performance in digit recognition, and soon, its variants (e.g., VGG [25], AlexNet [27]) have emerged in multitude. In general, a CNN is forward network with multiple layers, each of which can be a convolutional layer, a nonlinear activation layer, a pooling layer, or a fully-connected layer. The fully-connected layer is a classic layer in neural network which extracts the global information from its input. It is often used as the final layer to be the classifier in the task. The convolutional layer can be viewed as the weight-sharing