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.

© 2017 Elsevier B.V. All rights reserved.

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 (U1611461616722416 1602184 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.