Computer Vision and Image Understanding 187 (2019) 102794

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

Attention with structure regularization for action recognition☆

Yuhui Quan[a,c], Yixin Chen[a], Ruotao Xu[a,*], Hui Ji[b]

[a] *School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, China*
[b] *Department of Mathematics, National University of Singapore, Singapore 119076, Singapore*
[c] *Guangdong Provincial Key Laboratory of Computational Intelligence and Cyberspace Information, China*

ARTICLE INFO

Communicated by: Nikos Paragios

Keywords:
Action recognition
Attention
Block-wise sparsity
Deep recurrent network

ABSTRACT

Recognizing human action in video is an important task with a wide range of applications. Recently, motivated by the findings in human visual perception, there have been numerous attempts on introducing attention mechanisms to action recognition systems. However, it is empirically observed that an implementation of attention mechanism using attention mask of free form often generates ineffective distracted attention regions caused by overfitting, which limits the benefit of attention mechanisms for action recognition. By exploiting block-structured sparsity prior on attention regions, this paper proposed an $\ell_{2,1}$-norm group sparsity regularization for learning structured attention masks. Built upon such a regularized attention module, an attention-based recurrent network is developed for action recognition. The experimental results on two benchmark datasets showed that, the proposed method can noticeably improve the accuracy of attention masks, which results in performance gain in action recognition.

1. Introduction

Human action recognition is a challenging yet important task which has been receiving increasing attention in recent years. Recognizing human actions is about identifying human activities (*e.g.* running, walking, or dancing) in video sequences or images. It is an essential tool that enables effective analysis on human behaviors as well as efficient interactions between humans and vision systems. Thus, human action recognition can see its usage in a wide range of applications, including surveillance, video retrieval, human activity prediction, content-based summarization, electronic entertainment, automated cinematography, and many others. See *e.g.* Moeslund et al. (2006) and Poppe (2010) for more discussions. Meanwhile, human action recognition is also a very challenging task due to significant variations in human actions, in terms of personal styles, human appearance, camera viewpoints, varying background and other environmental changes.

In the past decades, there has been an enduring effort on the development of effective action recognition systems, and they are quite successful under well-controlled environments (Poppe, 2010). However, for action recognition in a single video sequence taken under unconstrained scenarios, it remains a challenging problem with limited success. Over the past years, many manually-crafted features have been proposed for action recognition to exploit various types of cues. To name a few, human poses (Lv and Nevatia, 2007; Thurau and Hlavác, 2008; Raptis and Sigal, 2013), skeletons from depth cameras (Wang et al., 2012; Du et al., 2015), local space–time patterns (Niebles et al., 2008; Yeffet and Wolf, 2009; Scovanner et al., 2007), trajectories of interest points (Wang et al., 2011; Wang and Schmid, 2013), motion patterns from optical flow (Fathi and Mori, 2008; Laptev et al., 2008; Carreira and Zisserman, 2017), and additional features from external non-visual cues such as video attributes (Yao et al., 2011) and movie scripts (Laptev et al., 2008). To further improve the performance in complex scenarios, these features are often combined together to achieve a better representation; see *e.g.* Wang and Schmid (2013), Bilen et al. (2016), Wang et al. (2016), Feichtenhofer et al. (2016a) and Carreira and Zisserman (2017).

More recently, with great advance in deep learning, there has been rapid progress on applying deep neural network to solve the problem of action recognition. Most existing studies adopt two types of neural networks architectures. One is Convolutional Neural Network (CNN) (Ji et al., 2013; Simonyan and Zisserman, 2014; Feichtenhofer et al., 2016b), and the other is Recurrent Neural Network (RNN) (Baccouche et al., 2011; Wang et al., 2012; Du et al., 2015; Zhu et al., 2016). By replacing manually-crafted features using adaptive features learned from data, these neural-network-based approaches for action recognition showed impressive improvement over traditional approaches. Regarding action recognition in videos, the RNN with Long-Short Term Memory (LSTM) cells (Yeung et al., 2015; Du et al., 2015; Zhu et al., 2016) is particularly appealing, as it allows the NN to exploit one

☆ No author associated with this paper has disclosed any potential or pertinent conflicts which may be perceived to have impending conflict with this work. For full disclosure statements refer to https://doi.org/10.1016/j.cviu.2019.102794.

* Corresponding author.
E-mail addresses: csyhquan@scut.edu.cn (Y. Quan), chen.yixin@mail.scut.edu.cn (Y. Chen), xu.ruotao@mail.scut.edu.cn (R. Xu), matjh@nus.edu.sg (H. Ji).

https://doi.org/10.1016/j.cviu.2019.102794
Received 29 August 2018; Received in revised form 7 May 2019; Accepted 6 August 2019
Available online 12 August 2019
1077-3142/