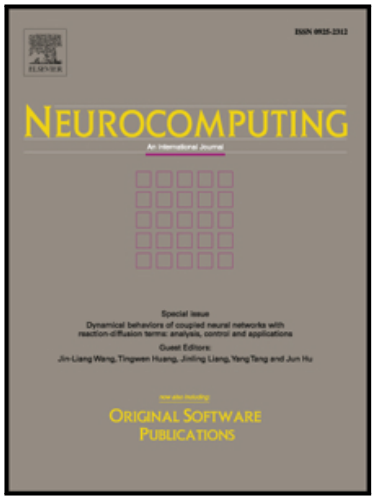
Communicated by Dr Zhiyong Wang

Accepted Manuscript



Action Unit Detection and Key Frame Selection for Human Activity Prediction

Haoran Wang, Chunfeng Yuan, Jifeng Shen, Wankou Yang, Haibin Ling

PII: S0925-2312(18)30987-1
DOI: https://doi.org/10.1016/j.neucom.2018.08.037
Reference: NEUCOM 19885

To appear in: *Neurocomputing*

Received date: 27 November 2017
Revised date: 22 July 2018
Accepted date: 17 August 2018

Please cite this article as: Haoran Wang, Chunfeng Yuan, Jifeng Shen, Wankou Yang, Haibin Ling, Action Unit Detection and Key Frame Selection for Human Activity Prediction, *Neurocomputing* (2018), doi: https://doi.org/10.1016/j.neucom.2018.08.037

Action Unit Detection and Key Frame Selection for Human Activity Prediction

Haoran Wang, Chunfeng Yuan, Jifeng Shen, Wankou Yang, and Haibin Ling

Abstract—Human activity prediction aims to recognize an unfinished activity with limited appearance and motion information. In this paper, we propose to predict an incomplete activity by combining the mid-level action units and the discriminative key frames exploited from each activity class. Specifically, we extract a great deal of action-related volumes from activity videos. Based on a set of low-level powerful features, similar volumes are aggregated into a mid-level feature, named action unit. Then, we detect these action units in each activity video and generate the frame feature by computing the distribution of concurrent action units in a single frame. Notice that human can easily recognize an incomplete activity using scanty key frames composed of representative interrelated action units together. The key frames in each activity class are selected by computing the entropy of each single frame feature. Finally, a structured SVM is trained to recognize activities with different observation ratios. The proposed approach is evaluated on several publicly available datasets in comparison with state-of-the-art approaches. The experimental results and analysis clearly demonstrate the effectiveness of the proposed approach.

Index Terms—Activity Prediction, Key Frame Selection, Action Unit Detection, Structured SVM.

1 Introduction

Human activity prediction has been an active research topic [1], [2], [4], [6], [7], [14], [35], [44], [45]. The prediction problem mainly focuses on analyzing ongoing activities, and it is critical in many real-world scenarios, such as driving assistance, human-computer interaction, video surveillance, and so on. In those situations, the intelligent systems are usually required to comprehend and react before the behaviors are finished. For example, the system is more useful if it is able to predict a dangerous driving situation and prevent an accident than recognize it after the event. Similarly, the ability of predicting human behaviors makes the human-computer interaction systems more humanized to provide better user experience.

Low-level features usually deal with the pixel-level characteristic which contains little spatial and temporal context information, and they try to detect the local intensity variations such as the key points and edges. Mid-level features are built on the low-level features, and contain more information than low-level features. High-level features refer to the semantic concept that we can directly see and recognize, such as a book and a car. They can be directly interpreted by human. High-level features are more effective to describe an object, but it's very hard to extract them from images or videos. Mid-level features are more discriminative than low-level features, and easier to extract than high-level features. Therefore, mid-level features are usually a good choice. Recently, mid-level features obtain outstanding performance in human activity analysis [8], [9], [10], [11], [12], [13], [15]. In this way, an activity is decomposed into a collection of spatial-temporal volumes. These volumes capture action features of multiple scales from body parts to the whole human body. Researchers are making efforts to exploit the interconnected relationship between adjacent mid-level features in order to discover more discriminative characters. Lan et al. [10] and Liu et al. [13] both adopt structured SVM to model the co-occurrence of mid-level pairs. Wu et al. [11] construct two kinds of graphs, named *video cooccurrence graph* and *video successiveness graph* respectively, to characterize the spatial and temporal relationship between adjacent local features. Furthermore, Jones

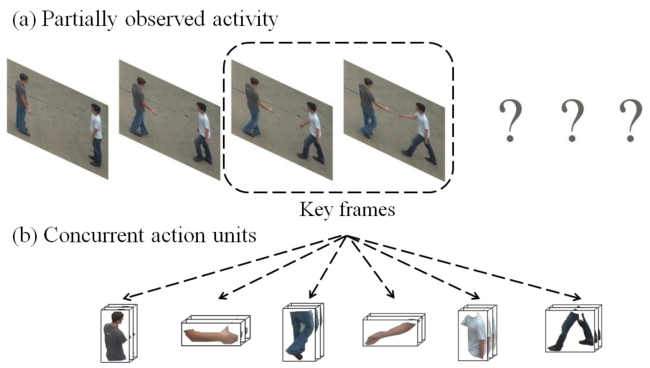


Fig. 1. Illustration of the proposed method. We use the key frames composed of concurrent action units to predict the partially observed activity class.

- *H. Wang is with the College of Information Science and Engineering, Northeastern University, Shenyang 110819, China. E-mail: wanghaoran@ise.neu.edu.cn*
- *C. Yuan is with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China. cfyuan@nlpr.ia.ac.cn*
- *J. Shen is with the School of Electrical and Information Engineering, Jiangsu University, Zhenjiang 212013, China. shenjifeng@ujs.edu.cn*
- *W. Yang is with the School of Automation, Southeast University, Nanjing 210096, China. youngwankou@yeah.net*
- *H. Ling is with the Department of Computer and Information Sciences, Temple University, Philadelphia, PA 19122, USA. hbling@temple.edu*