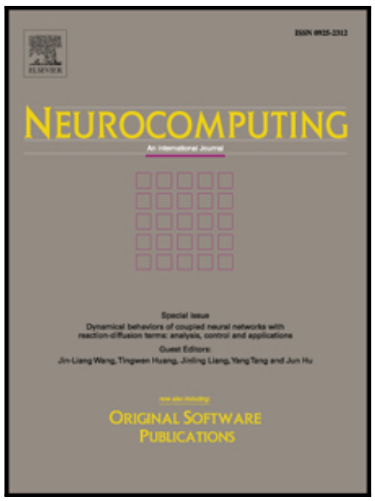

A Deep-Learning Based Feature Hybrid Framework for Spatiotemporal Saliency Detection inside Videos

Zheng Wang , Jinchang Ren , Dong Zhang , Meijun Sun , Jianmin Jiang

PII: S0925-2312(18)30109-7
DOI: 10.1016/j.neucom.2018.01.076
Reference: NEUCOM 19278

To appear in: *Neurocomputing*

Received date: 8 September 2017
Revised date: 4 January 2018
Accepted date: 29 January 2018

Please cite this article as: Zheng Wang , Jinchang Ren , Dong Zhang , Meijun Sun , Jianmin Jiang , A Deep-Learning Based Feature Hybrid Framework for Spatiotemporal Saliency Detection inside Videos, *Neurocomputing* (2018), doi: 10.1016/j.neucom.2018.01.076

A Deep-Learning Based Feature Hybrid Framework for Spatiotemporal Saliency Detection inside Videos

Zheng Wang[a], Jinchang Ren[b], Dong Zhang[c], Meijun Sun[c,]*, Jianmin Jiang[d]

[a]*Media Technology and System (MTS) Lab., School of Computer Software, Tianjin University, Tianjin 300350, China*
[b]*Centre for excellence in Signal and Image Processing, University of Strathclyde, Glasgow, U.K.*
[c] *School of Computer Science and Technology, Tianjin University, Tianjin, 300350, China*
[d] *Research Institute for Future Media Computing, College of Computer Science & Software Engineering, Shenzhen University, China*

**corresponding author, sunmeijun@tju.edu.cn, jianmin.jiang@szu.edu.cn*

Abstract: Although research on detection of saliency and visual attention has been active over recent years, most of the existing work focuses on still image rather than video based saliency. In this paper, a deep learning based hybrid spatiotemporal saliency feature extraction framework is proposed for saliency detection from video footages. The deep learning model is used for the extraction of high-level features from raw video data, and they are then integrated with other high-level features. The deep learning network has been found extremely effective for extracting hidden features than that of conventional handcrafted methodology. The effectiveness for using hybrid high-level features for saliency detection in video is demonstrated in this work. Rather than using only one static image, the proposed deep learning model take several consecutive frames as input and both the spatial and temporal characteristics are considered when computing saliency maps. The efficacy of the proposed hybrid feature framework is evaluated by five databases with human gaze complex scenes. Experimental results show that the proposed model outperforms five other state-of-the-art video saliency detection approaches. In addition, the proposed framework is found useful for other video content based applications such as video highlights. As a result, a large movie clip dataset together with labeled video highlights is generated.

Keywords*:* spatiotemporal saliency detection, Human gaze, convolutional neural networks, visual dispersion, movie highlight extraction.

1. Introduction

Visual saliency has been an important and popular research in image processing for decades with a sole purpose to mimic biological visual perception for machine vision applications. Substantial interests in the field as evidenced by the vast volume of publications, such as application of saliency concept for image/video compression and recognition[1]-[6], automatic image cropping[7], non-photorealistic rendering[8], adaptive image display on small devices[9], movie summarization[10], shot detection [11], human-robot interaction[12], and detection of multi-class geospatial targets[13][14] have been reported in the last two decades.

Historically, saliency detection research was first initiated by Treismanand Gelade in 1980[15] who proposed the "Feature Integration Theory", which illustrated how visual attention was attracted by features in the imagery. Itti and Koch's model triggered strong interests in this field of research, including the use of low-level features to map the saliency regions/objects in the image scene[16]. He et al[17] proposed a biologically inspired saliency model using high-level object and contextual features for saliency detection based on Judd's concept [18]. Further extension of research along this line was reported by Goferman et al [19] who emphasized that four important factors, including local low-level features, global consideration, visual organization and high-level factors could affect saliency detections strongly. The methodology for feature extractions has also been improved.

Despite of intensive research in the image based saliency detection, video saliency has not been addressed until recent years. In fact, video saliency is quite different from that of still images, mainly because of the very limited frame-to-frame interval time for the observers' attention to be drawn by features in the scene. Although there are extensions from the image-based saliency models for the video stream such as the temporal intensity and orientation contrasts as dynamic features[20][21], better frame work is needed for more efficient saliency detection from video footage.

While most work in the field has been focusing on low level features, human attention prediction is considered to be dominated by some high-level features, such as objects, actions and events. Rudoy et al[22] employed viewer's gazing direction and also to use their actions as cue to locate the saliency features, as opposed to the conventional image based pixel feature extraction method. Han et al.[23] proposed that meaningful objects were important to saliency detection. Based on visual attention and eye movement data, a video saliency detection model was trained and it was found to outperform all other state-of-the-art algorithms.

On one hand, conventional handcrafted features have proven their success in existing approaches and applications. On the other hand, deep learning networks has shown their great potential in computer vision such as coping with human perception especially for large-scale data and more complicated problems. It is our intension here to combine all of these approaches together to address the challenges for video saliency detection. Some papers [24][25] report that it is effective to combine deep learning based features and handcrafted features for saliency detection. However these methods use only single image and do not consider temporal