Journal of Visual Communication and Image Representation 50 (2018) 1–8

Contents lists available at ScienceDirect

Journal of Visual Communication and Image Representation

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

Optimal feature combination analysis for crowd saliency prediction☆

Guangyu Gao[a,*], Cen Han[a], Kun Ma[a], Chi Harold Liu[a], Gangyi Ding[a], Erwu Liu[b]

[a] School of Software, Beijing Institute of Technology, Beijing 100081, China
[b] School of Electronics and Information, Tongji University, Shanghai 200092, China

ARTICLE INFO

Keywords:
Crowd
Saliency
Random forest
Visual attention
Face detection

ABSTRACT

Crowd saliency prediction refers to predicting where people look at in crowd scene. Humans have remarkable ability to rapidly direct their gaze to select visual information of interest when looking at a visual scene. Until now, research efforts are still focused on what type of feature is representative for crowd saliency, and which type of learning model is robust for crowd saliency prediction. In this paper, we propose a Random Forest (RF) based crowd saliency prediction approach with optimal feature combination, i.e., the Feature Combination Selection for Crowd Saliency (FCSCS) framework. More specifically, we first define three representative crowd saliency features, namely, FaceSizeDiff, FacePoseDiff and FaceWhrDiff. Next, we adopt the Random Forest (RF) algorithm to construct our saliency learning model. Then, we evaluate the performance of FCSCS framework with different feature combinations (fifteen combinations in our experiments). Those selected features include low-level features (i.e., color, intensity, orientation), four crowd features (i.e., face size, face density, frontal face, profile face) and three new defined features (i.e., FaceSizeDiff, FacePoseDiff and FaceWhrDiff). We use FCSCS framework to obtain the optimal feature combination that is most suitable for crowd saliency prediction and further train the saliency model based on the optimal feature combination. After that, we evaluate the performance of the crowd saliency prediction classifiers. Finally, we conduct extensive experiments and empirical evaluation to demonstrate the satisfactory performance of our approach.

1. Introduction

For a given image or a video, one way to find if human interested in or not is to give it semantic tags, i.e. [1,2]. However, these semantic tags are summarized with natural language and labeled with human labor. More naturally, given a visual scene, human has the ability to selectively locate eye fixations on some informative contents to present what they are interested in, namely fixation prediction, also known as saliency prediction. More specifically, the meaning of saliency is that regions or objects stand out from their neighbors. Saliency prediction is always a basic technique for many applications. For example, Nguyen et al. [3] proposed a spatial-temporal attention-aware pooling for action recognition. While egocentric videos analysis methods are very popular with the universal wearable devices [4], the saliency prediction will be a good candidate assistant in the area egocentric videos analysis [5].

Liu et al. [6] proposed a computational framework to learn visual features from raw images with multiresolution convolutional neural network. However, most of the existing saliency models focus on regular density scenarios. Here, a particular scenario is the crowd scene, where a relatively large amounts of people existed in the image. In other words, saliency in crowd scene may be very different from that in the regular density scenarios. Therefore, except for the conventional methods and saliency features, more specific crowd characteristics should be considered for crowd saliency prediction. Crowd is very prevalent in most of the vision images, and saliency in crowd is very important for various significant problems, such as population monitoring, urban planning, and public security. In many scenarios, crowd scenes may be more important than regular density scenarios, because criminal or terrorist attacks often happen within a crowd scene. Jiang et al. [7] have done some pioneering research efforts on saliency in crowd and got some good results. However, it is still a very challenge task on that: (1) which feature has more influence on crowd saliency? (2) how to combine different saliency features to achieve the best performance? (3) which kind of saliency prediction model is more suitable and robust for saliency prediction in crowd?

Meanwhile, with consideration of different types of feature together, it always got a more satisfactory performance, i.e., the multi-

☆ This work was supported by the National Natural Science Foundation of China under Grant 61401023.
* Corresponding author.
E-mail addresses: guangyugao@bit.edu.cn (G. Gao), mysayalhan@gmail.com (C. Han), kunma@bit.edu.cn (K. Ma), chiliu@bit.edu.cn (C.H. Liu), dgy@bit.edu.cn (G. Ding), erwuliu@tongji.edu.cn (E. Liu).

https://doi.org/10.1016/j.jvcir.2017.11.002
Received 21 February 2017; Received in revised form 29 October 2017; Accepted 1 November 2017

modality analysis in multimedia area [8]. Based on all these considerations, in this paper, we construct a robust approach to explore the optimal feature combination for crowd saliency prediction. In particular, we first define three novel crowd saliency features, namely, FaceSizeDiff, FacePoseDiff and FaceWhrDiff. Then, a novel framework, namely, the Feature Combination Selection for Crowd Saliency (FCSCS) framework, is constructed to find the optimal features for saliency prediction. After that, we evaluate the performance of fifteen prediction classifiers with different feature combinations for saliency in crowd. Finally, we obtain the optimal feature combination for crowd saliency prediction. Actually, in order to demonstrate the effectiveness of the FCSCS framework, we adopt the "wrapper for feature subset selection" [9] to obtain the optimal feature combinations, and the result of that is consistent with our FCSCS framework. In order to effectively integrate information from multiple features at both low- and high-level, we adopt the Random Forest (RF) algorithm to learn a more robust discrimination model between salient and non-salient regions.

In addition, this paper extensively extends our perviously published conference paper [10] in WCSP2016, by improving in terms of technical content, theoretical analysis, performance evaluation and presentation compared to the conference version. In this paper, we further do some research efforts, the main innovation points are summarized as:

(1) A new crowd feature, i.e., FaceWhrDiff, is defined and used for optimal feature combination. This feature is proved to be beneficial on improving the predicting performance.
(2) We explain the reasons why we select fifteen feature combinations to train our saliency model. We add a "Feature Classification" Module in the FCSCS framework.
(3) We demonstrate the satisfactory performance of FCSCS framework by using "wrapper approach for feature subset selection" [9] to obtain the optimal feature.

The rest of the paper is organized as follows. In Section 2, we describe the related research efforts about saliency prediction, especially crowd saliency prediction. In Section 3, we propose three new crowd features, used for optimal feature combination selection, and the proposed framework for saliency prediction. Section 4 discusses the experiments and performance evaluation, and also Section 5 presents the conclusions of this paper.

2. Related work

Related research efforts about predicting saliency can be classified into two categories, namely: (a) visual saliency in regular density scenario, (b) visual saliency in crowd scenario.

2.1. Visual saliency in regular density scenario

Xu et al. [11] presented a method based on Gaussian mixture model to predict saliency. Jiang et al. [12] predicted saliency based on neurophysiological and psychophysical studies of peripheral vision. Zhao et al. [13] proposed a multi-context deep learning framework for salient object prediction. In [14], Wang et al. learned a combined model of visual saliency for fixation prediction. Besides, Parkhurst et al. [15] demonstrated that stimulus-driven, bottom-up mechanisms contribute significantly to attentional guidance under natural viewing conditions. And also, Zhang et al. [16] proposed a Boolean Map based Saliency model (BMS) to demonstrate the usefulness of surroundedness for eye fixation prediction.

Meanwhile, some literatures also used the bayesian model to predict where human look at. For example, Torralba [17] and Oliva [18] proposed a bayesian framework in view of the visual search task. Zhang et al. [19] proposed a definition about visual significance, naming SUN (Saliency Using Natural Statistics). Qin et al. [20] presented an integration algorithm in the Bayesian framework to take advantage of multiple saliency maps.

In addition, the decision theory models were also used to predict saliency region. For example, Bruce et al. [21] proposed the AIM (Attention based on Information Maximization) model to compute the salience value in image. Seo and Milanfar [22] proposed a predicting method on saliency region based on decision theory.

Others used the pattern classification models to predict visual saliency. In these models, machine learning algorithms are used to construct saliency prediction models. For example, Judd et al. [23] learned the saliency with a set of low-, mid-, and high-level features using SVM algorithm. Zhao et al. [24] adopted a least square technique for saliency prediction.

Except for that, Lang et al. [25] also focused on introduction the eye fixation dataset NUS-3DSaliency compiled from 600 images for both 2D and 3D scenes. In [26], Nguyen et al. gave a comprehensive study and analysis on dynamic saliency and static saliency together.

2.2. Saliency in crowd scenario

Many related research efforts have done for saliency prediction. However, firstly, most of these methods focused on regular density scenario, which can't predict the saliency in the crowd scene well. Secondly, the performance of saliency models in crowd scenario are not good enough. That is to say, visual saliency has been extensively studied, but only a few efforts have been spent in crowd scene. Compared to saliency detection in regular density scenario, saliency in crowd will be more diversity and difficult.

Nevertheless, many researchers always pay attentions to this area with consideration of more particular knowledge in crowd scenario. For example, Lim et al. [27] proposed to identify and localize salient regions in a crowd scene, by transforming low-level features extracted from crowd motion field into a global similarity structure. In [28], an adaptive inductive reasoning mechanism was presented for saliency extraction and information reconstruction in a distributed camera sensors network. Not only that, the authors of [29] presented a efficient unsupervised learning method on video analysis for abnormal crowd activity detection based on spatiotemporal saliency detector. Kok et al. [30] analyzed the crowd behavior with the review on where physics meets biology.

In the scene with many human faces, faces detection are also used in saliency prediction. For example, Cerf et al. [31] predicted human gaze using low-level saliency combined with face detection and demonstrate the importance of faces in gaze deployment. Cerf et al. [32] demonstrated that faces attract attentions strongly and rapidly. Mathialagan et al. [33] proposed a method to find the important people in images. That means, faces in crowd scenario is very crucial, and saliency prediction performance can been significantly improved with the use of face detection. For example, Jiang et al. [7] have done some pioneering research efforts on saliency in crowd. They proposed several features related with faces which is used to predict saliency in the crowd scene and demonstrate that crowd density affects saliency.

In this paper, in order to enhance the performance of saliency models in the crowd scenario, we first defined three new crowd features, and next proposed a FCSCS framework. Finally, we constructed a crowd saliency prediction model to evaluate the performance.

3. Feature definition and saliency model

In this section, we define three crowd features and propose a framework of Feature Combination Selection for Crowd Saliency (FCSCS).

3.1. Data set

We use the eye tracking dataset proposed by Jiang et al. [7], for saliency analysis in crowd scenes. The dataset consists of 500 natural