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

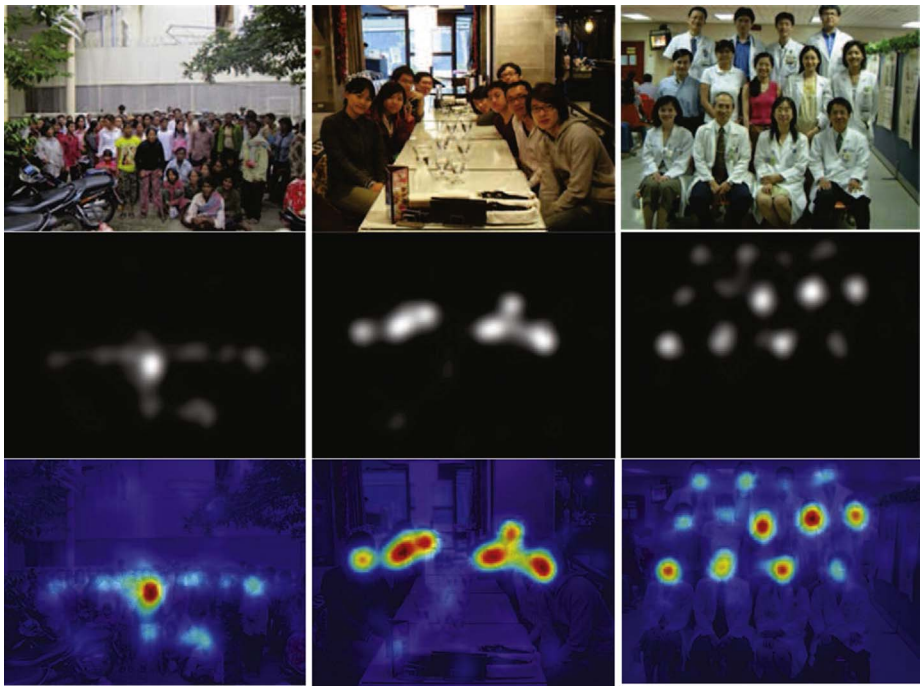

Fig. 1. Examples of crowd image stimuli for eye track dataset.

crowd images (1024 × 768) with a diverse range of crowd densities. In all images, human faces were manually labeled with rectangles, and two attributes were annotated on each face: *pose* and *partial occlusion*. Pose has three categories: *frontal* if the angle between the face's viewing and the image plane is roughly less than 45°, *profile* if the angle is roughly between 45° and 90°, and *back* otherwise. The second attribute was annotated as *partial occluded* if a face is partially occluded. Note that if a face is completely occluded, it is not labeled. For more details about this dataset, please refer to Jiang et al. [7]. Fig. 1 shows examples of the crowd image stimuli for eye track dataset.

3.2. Feature definition

From previous study, we know that faces attract attention, yet in crowding scenarios, not all faces attract attention. What are the driving factors in determining faces attract people's attention? Jiang et al. [7] have done some pioneering research efforts about this problem. In this paper, we make further steps in this exploration. Except for several traditional low-level features and crowd features presented in [7], we also develop three novel features based on face information. For face detection, we used the person head detection method proposed by Vu et al. [34], and also the OKAO face detection [35] together. In particular, we first define the three relevant features about crowd saliency as follows.

FaceSizeDiff. This feature describes the size difference of a specific face compared with other faces in the same image. For simplicity, this feature that we denoted as "FaceSizeDiff", is the size difference of faces, where let FSD_i be its value for the ith face.

$$\mathrm{FSD}_i = \frac{1}{\mathrm{nFace\text{-}1}} \sum_{k \neq i} \tanh[(x_i - x_k)^2], \tag{1}$$

where x_i,x_k are the size of the ith face and the kth face respectively. nFace is the number of faces in an image. If two faces have the approximate size, their FSD value will be similar. In fact, the bigger the face's FSD value is, the more likely that it is saliency in crowd. For example, in Fig. 2, the faces around the man in the center are relatively small size faces, while the man with scarf and hat has a big size face. The face of this man is obviously different from any other faces, and will attract more attentions intuitively. Thus, the FSD of the face of the man with scarf should be big. Meanwhile, in Fig. 3, the far and small faces' size is different from most of the other faces, thus the FSD of these faces should be bigger, and attract more attentions.

FacePoseDiff. This feature describes the pose difference of a specific face compared with other faces in the same image. There are three types of pose for a face, i.e. *frontal*, *profile*, and *back*. For simplicity, the feature of pose difference, is denoted as "FacePoseDiff", where let FPD_i be its value for the ith face.

$$\mathrm{FPD}_i = \exp\left(\frac{-\mathrm{nPose}_i}{\mathrm{nFace}}\right), \tag{2}$$

where nPose_i is the number of the pose of the ith face. If the pose of the ith face is a *frontal*, *profile*, or *back* faces, the term of nPose_i refers to the number of the corresponding type of face in the image. nFace is the total number of faces in the image. The smaller the proportion (nPose/nFace) is, the bigger the FPD is. A face will attract more attentions when it's FPD value increases. This embodies the principle of "when a thing is scarce, it is precious". For example, in Fig. 4, there are three types of face pose, i.e., *frontal*, *profile* and *back* faces. Intuitively, the left girl with yellow T-shirt attract more eye fixations, while she shows a profile face, which is different to others. Thus the FPD of she's face is the biggest, and she's face is more likely to be a saliency region.

FaceWhrDiff. This feature describes the aspect ratio difference of a specific face compared with other faces in the same image. For simplicity, the feature of face's aspect ratio differences, is denoted as "FaceWhrDiff", where let FWD_i be its value for the ith face.

$$\mathrm{FWD}_i = \frac{1}{\mathrm{nFace}-1} \sum_{k \neq i} \{1 - \exp[-(r_i - r_k)^2]\}, \tag{3}$$

where r_i,r_k are the aspect ratio of the ith face and the kth face respectively. nFace is the number of faces in an image. If two faces have the similar aspect ratio, the FWD is similar. In fact, big FWD always refers to saliency in crowd. For example, in Fig. 5, there are 8 faces. For example, in Fig. 5, the aspect ratio of the 6th face from left to right, is different from most of the other faces, thus the FWD of the 6th face should be bigger, and it attracts more attention.

Fig. 2. Image with different face sizes. The man with scarf and hat, in the image center, has bigger face, and catch more attentions.

Fig. 3. Image with different face sizes. The far and small faces will attract more attentions.

3.3. Saliency prediction model

As shown in Fig. 6, we propose the FCSCS framework, which combines low-level features and high-level semantic face features for crowd saliency prediction. For each image, we pre-compute feature maps for every pixel of the image resized to 256×192. In particular, we generate three simple biologically plausible low-level feature maps (i.e., intensity, color, and orientation) as Itti et al. [36] have done. Moreover, face features can effectively distinguish salient faces from other faces. We extract seven feature maps on face features (i.e., face size, face density, fontal face, profile face, FaceSizeDiff, FacePoseDiff, FaceWhrDiff). Among these face features, FaceSizeDiff, FacePoseDiff and FaceWhrDiff are new features defined by us. In our FCSCS framework, we select the best feature combination with the following seven steps:

Step 1: Ten feature maps are extracted from each natural crowd image as that introduced in [7].
Step 2: We randomly sample 10 pixels respectively, yielding a training set of 4500 positive samples and 4500 negative samples from the top 20% and bottom 70% regions in a ground truth saliency map. The values at each selected pixel in the ten feature maps are concatenated into a feature vector.
Step 3: We classify ten features into four primary groups. The four primary groups features are shown in Table 1.
Step 4: Fifteen feature combinations were generated based on the above mentioned four primary groups features, which are shown in Table 2.
Step 5: For each feature combination, a saliency classifier is constructed using random forest algorithm.
Step 7: Each saliency classifier is evaluate with evaluation metrics of Shuffled AUC, NSS and CC.
Step 8: The best feature combination is determined by comparing the fifteen evaluation results.

Actually, the ten features are classified into four primary groups based on feature relevance. The principles of the classification are listed as follows.

(1) Color, intensity and orientation are all low level feature. Thus, we classify color, intensity and orientation into a same primary group.
(2) Face size and face density have the impact on visual saliency based on local face density. Thus, we classify face size and face density into a same primary group.
(3) Frontal face and profile face have the impact on visual saliency based on face pose. Thus, we classify frontal face and profile face into a same primary group.

Fig. 4. Image with different face pose, including *frontal* faces, *back* faces, and *profile* faces. It seems that, the left girl on yellow T-shirt show a profile face, which is different to others, has attract more fixations.

Fig. 5. Image with different face aspect ratio. Intuitively, the person the center attracts more fixations, and he has different face aspect ratio compared to other person in this image.

(4) FaceSizeDiff, FacePoseDiff and FaceWhrDiff have the impact on visual saliency based on difference of face. Thus, we classify FaceSizeDiff, FacePoseDiff and FaceWhrDiff into a same primary group.

After that, we can see the four primary groups features in Table 1. Then, we generate the fifteen feature combinations based on this four primary groups features, and the fifteen feature combinations are shown in Table 2.

We use the random forest algorithm to train the saliency model and generate the classification of saliency and non-saliency with the extracted features. The Random Forest algorithm [37] can be summarized as:

(1) Draw n_{tree} bootstrap samples from the original data.
(2) For each of the bootstrap samples, grow an unpruned classification or regression tree, with the following modification: at each node, rather than choosing the best split among all predictors, randomly sample m_{try} of the predictors and choose the best split from all of those variables.
(3) Predict new data by aggregating the predictions of the n_{tree} trees (i.e., majority votes for classification, average for regression).

In our model, we denote n_{tree} as the number of trees in the random forest algorithm, and M is the number of features for saliency train in each classification. For each tree, we sample training objects as our bootstrapped training data set, and use $\max(\lfloor M/3 \rfloor,1)$ features to fit the training set, i.e., $m_{\text{try}} = \max(\lfloor M/3 \rfloor,1)$. For each tree, the output is $+1$ for salient, or -1 for non-salient. Finally, we select saliency value of most of n_{tree} trees as our predicted saliency value.

In the experiments, we use 9000 objects in 500 images as the training set, where $n_{\text{tree}} = 500$. We find the Random Forest appropriate for two reasons: first, the generalization error is small; second, the Random Forest always converge so that overfitting is not a problem.

Table 1
Four primary groups features.

Primary group	Features
(1)	low-level Features (color, intensity, orientation)
(2)	face size, face density
(3)	frontal face, profile face
(4)	FaceSizeDiff, FacePoseDiff, FaceWhrDiff

Table 2
Fifteen feature combinations.

No.	Feature combinations
(1)	low-level Features
(2)	face size, face density
(3)	frontal face, profile face
(4)	FaceSizeDiff, FacePoseDiff, FaceWhrDiff
(5)	face size, face density, frontal face, profile face
(6)	face size, face density, FaceSizeDiff, FacePoseDiff, FaceWhrDiff
(7)	frontal face, profile face, FaceSizeDiff, FacePoseDiff, FaceWhrDiff
(8)	face size, face density, frontal face, profile face, FaceSizeDiff, FacePoseDiff, FaceWhrDiff
(9)	face size, face density, low-level Features
(10)	frontal face, profile face, low-level Features
(11)	FaceSizeDiff, FacePoseDiff, FaceWhrDiff, low-level Features
(12)	face size, face density, frontal face, profile face, low-level Features
(13)	face size, face density, FaceSizeDiff, FacePoseDiff, FaceWhrDiff, low-level Features
(14)	frontal face, profile face, FaceSizeDiff, FacePoseDiff, FaceWhrDiff, low-level Features
(15)	face size, face density, frontal face, profile face, FaceSizeDiff, FacePoseDiff, FaceWhrDiff, low-level Features

In the saliency model, there are many features affect the crowd saliency. We concatenate different features together to train the crowd saliency prediction classifier. We have ten different features, i.e., color,

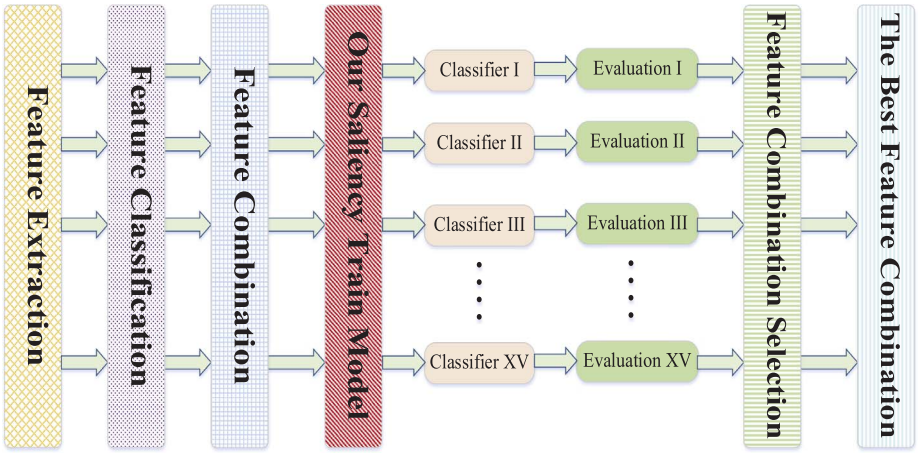

Fig. 6. Feature Combination Selection for Crowd Saliency (FCSCS) framework.

intensity, orientation, face size, face density, frontal face, profile face, FaceSizeDiff, FacePoseDiff, FaceWhrDiff. Firstly, we classify the ten features into four primary groups, and then we use these four primary groups features to constitute fifteen kinds of feature combinations. For each feature combination, the *M* value is different, for example, the *M* value is 3 for *No. (1)* at Table 2, and the *M* value is 2 for *No. (2)* at Table 2.

4. Performance evaluation

In this section, we evaluate the performance of FCSCS framework and our saliency model for crowd saliency prediction.

In literature, there are several widely used criteria that can quantitatively evaluate the performance of saliency models by comparing the saliency prediction with eye movement data. These evaluation metrics are:

- The Normalized Scanpath Saliency (NSS) [38].
- The Correlation Coefficient (CC) [39].
- The area under the receiver operator characteristic (ROC) curve (i.e. AUC) [40].

By definition, NSS evaluates salience values at fixated location in the normalized predicted saliency map which has zero mean and unit standard deviation. A larger NSS implies a greater correspondence between fixation locations and the saliency predictions. While CC is defined to measure the linear correlation between the saliency map and the ground truth map. AUC is also a good metric to evaluate the performance of saliency models. However, AUC is significantly affected by the center bias. In order to eliminate the center bias effect, the Shuffled AUC is presented in [19]. Thus, in this paper, we use the Shuffled AUC as the evaluation metric. The three metrics (i.e. NSS, CC and the Shuffled AUC) provide a relatively objective evaluation of various models.

4.1. Performance evaluation on FCSCS framework

In this section, we perform the quantitative evaluation of our FCSCS framework. We divide the feature combinations into three kinds, they are shown as follows:

(1) Low-level features, i.e., *No. (1)* in Table 2.
(2) Crowd features without combining low-level features, i.e., *No. (2)* to *No. (8)* in Table 2.
(3) Crowd features combining low-level features, i.e., *No. (9)* to *No. (15)* in Table 2.

We train the fifteen saliency prediction classifiers according to those feature combinations, then evaluate the classifier's performance by the Shuffled AUC, NSS, and CC metrics. Table 3 shows the performance of this fifteen feature combinations with our prediction model.

As shown in Table 3 (with our model), the Shuffled AUC, NSS, and CC value of the 2nd to the 8th row are greater than the Shuffled AUC, NSS, and CC value of the 9th to the 15th row respectively. In other words, feature combinations without low-level features are better than the counterparts combining low-level features. Therefore, that is to say, low-level features are not absolute good for saliency prediction in crowd scenarios. The Shuffled AUC, NSS, and CC metric of *No. (6)* has the greatest value, which means that the optimal feature combination is "face size, face density, FaceSizeDiff, FacePoseDiff, and FaceWhrDiff".

In order to demonstrate the effectiveness of the FCSCS framework, we also use the wrapper approach for feature subset selection [9] to obtain the optimal feature combination. The wrapper approach for feature subset selection is summarized as Algorithm 1.

In Algorithm 1, *i* denotes the *i*th feature, *F* denotes the set which includes features. From Algorithm 1, we can see that when the performance of the set *F* no longer improves, the feature subset includes the optimal feature combinations. The feature subset are shown at Table 4. In Algorithm 1, the Shuffled AUC, NSS, CC are used as evaluation metric to evaluate the performance of the set *F*, and the performance results of feature subsets at Table 4 are shown at Table 5. The performance of the set *F* improves as the number of features in the set *F* increases. As shown at Table 5, the Shuffled AUC, NSS, CC value of No. (5) is greatest compared with the other four feature sets. Feature subset of No. (5) at Table 5 includes five features, i.e., "FaceWhrDiff, density, size, FaceSizeDiff, and FacePoseDiff".

In Table 5, we find the best features subset for saliency prediction in crowd with the wrapper approach, i.e. FaceWhrDiff, density, size, FaceSizeDiff, and FacePoseDiff. It seems to be consistent with the claim with our FCSCS framework, that the low-level features are not so good for saliency prediction. In order to make this claim to be more convince, we added the proposed three low-level features one by one to the selected best features subset, to examine if the performance improved. Based on feature subset of *No. (5)* at Table 4, the feature subsets with six features are shown at Table 6. And the results in Table 7 show that all the feature subset with low-level features nearly get the same performance, which means that the newly added low-level features don't make any obvious contributions to the performance.

The result of "wrapper approach for feature subset selection" is consistent with the result of FCSCS framework, that is, the optimal feature combinations is "FaceWhrDiff, density, size, FaceSizeDiff, and FacePoseDiff". Thus, the experiment demonstrates the effectiveness of our FCSCS framework.

Table 3
Performance results of different feature combinations with our model.

No.	Shuffled AUC	NSS	CC
(1)	0.5991	0.8870	0.4157
(2)	0.6702	1.1575	0.5132
(3)	0.6646	1.1452	0.5060
(4)	0.6699	1.1593	0.5171
(5)	0.6693	1.1552	0.5120
(6)	0.6761	1.2062	0.5285
(7)	0.6720	1.1662	0.5225
(8)	0.6715	1.1659	0.5227
(9)	0.6637	1.1523	0.5058
(10)	0.6541	1.1093	0.5057
(11)	0.6693	1.1586	0.5209
(12)	0.6613	1.1470	0.5029
(13)	0.6737	1.1573	0.5201
(14)	0.6709	1.1591	0.5239
(15)	0.6719	1.1575	0.5253

Table 4
Feature subsets.

No.	Feature subsets
(1)	FaceWhrDiff
(2)	FaceWhrDiff, face density
(3)	FaceWhrDiff, face density, face size
(4)	FaceWhrDiff, face density, face size, FaceSizeDiff
(5)	FaceWhrDiff, face density, face size, FaceSizeDiff, FacePoseDiff

Table 5
Performance evaluation results on feature subsets.

No.	Shuffled AUC	NSS	CC
(1)	0.6696	1.1664	0.5107
(2)	0.6743	1.1725	0.5211
(3)	0.6750	1.1848	0.5276
(4)	0.6751	1.1858	0.5279
(5)	0.6761	1.2062	0.5285

Table 6
Feature subsets with six features.

No.	Feature subsets with six features
(1)	FaceWhrDiff, face density, face size, FaceSizeDiff, FacePoseDiff, color
(2)	FaceWhrDiff, face density, face size, FaceSizeDiff, FacePoseDiff, intensity
(3)	FaceWhrDiff, face density, face size, FaceSizeDiff, FacePoseDiff, orientation
(4)	FaceWhrDiff, face density, face size, FaceSizeDiff, FacePoseDiff, frontal
(5)	FaceWhrDiff, face density, face size, FaceSizeDiff, FacePoseDiff, profile

Table 7
Performance evaluation results on feature subsets with six features.

No.	Shuffled AUC	NSS	CC
(1)	0.6750	1.1846	0.5266
(2)	0.6583	1.1382	0.5180
(3)	0.6751	1.1883	0.5269
(4)	0.6746	1.1814	0.5245
(5)	0.6743	1.1341	0.5249

Algorithm 1. The wrapper approach for feature subset selection.

Require: $F = \emptyset, k = 0$
repeat
  for $i = 1$ **to** n
    if $i \notin F$
      $F_i = F \cup i$
      Evaluate the performance of F_i
    end if
  end for
  $F =$ the best F_i
until $k > n$ or the performance of F no longer improves

4.2. Performance evaluation on our saliency model

In order to prove the good performance of our model, we further perform qualitative and quantitative evaluation, in comparison with four state-of-the-art saliency models that are publicly available.

The four comparative models are MIT model [23], SMVJ model [31], Itti et al.'s model [36], and SC-M model [7]. MIT model, SMVJ model and SC-M model are bottom-up ones combined with object detectors. Itti et al.'s model is purely bottom-up.

Figs. 7–9 show the quantitative comparison of our model and the other four state-of-the-art saliency models. In these experiments, the quantitative evaluation is following Borji's implementations [41].

In Fig. 10, we illustrate the ROC curves for the Shuffled AUC

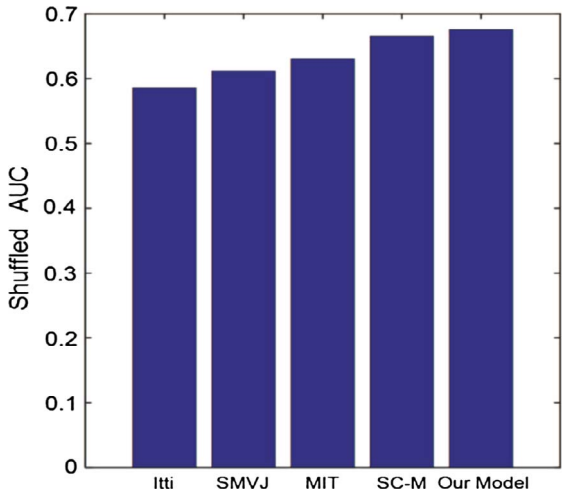

Fig. 7. Quantitative comparison of models. The bar values indicate the average performance over all stimuli. The prediction accuracy is measured with Shuffled AUC scores.

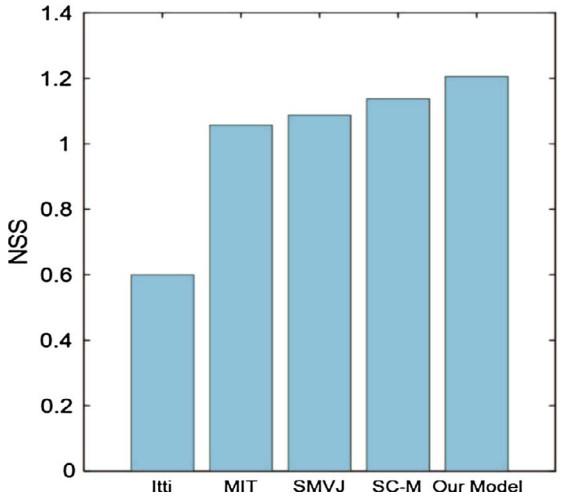

Fig. 8. Quantitative comparison of models. The bar values indicate the average performance over all stimuli. The prediction accuracy is measured with NSS scores.

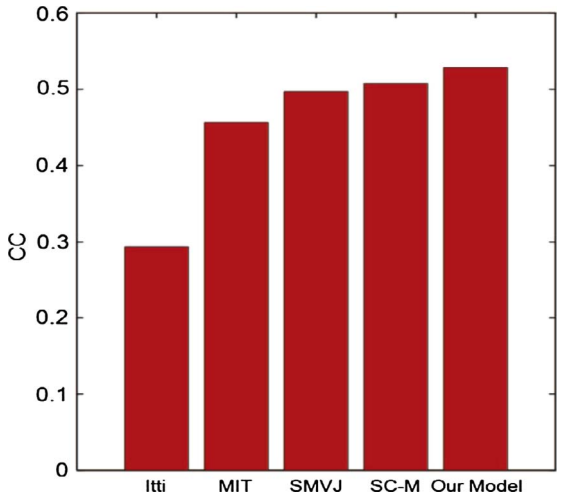

Fig. 9. Quantitative comparison of models. The bar values indicate the average performance over all stimuli. The prediction accuracy is measured with CC scores.

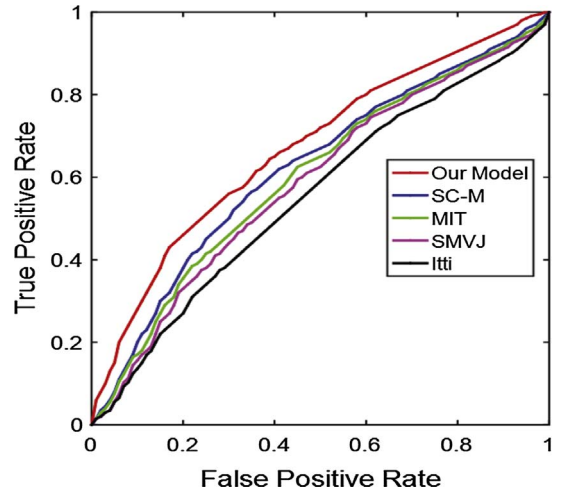

Fig. 10. ROC curves for compared model. Our model, SC-M, MIT, SMVJ models incorporate face features.

computation of the compared models. Four key observations are made as follows. First, our proposed model outperform all other models in predicting crowd saliency (with all three metrics), demonstrating that the three new defined face features (i.e., FaceSizeDiff, FacePoseDiff, and FaceWhrDiff) are useful. Second, models with face features perform better than those without face features. Third, FaceWhrDiff is a important feature which do improve the performance of crowd saliency model. Finally, face size, face density, FaceSizeDiff, FacePoseDiff, and FaceWhrDiff are the key features for saliency prediction in crowd scenarios.

5. Conclusions

In this paper, we define three new crowd features and propose a Random Forest (RF) based crowd saliency prediction approach with the optimal feature combinations, i.e., the Feature Combination Selection for Crowd Saliency (FCSCS) framework. In this framework, we adopt the Random Forest algorithm to construct our saliency learning model. We evaluate the performance of the FCSCS framework with different feature combinations (fifteen combinations in our experiments), and obtain the optimal feature combination which is most suitable for crowd saliency. Then, we further train the saliency model based on the optimal feature combination. Extensive experiments and empirical evaluations demonstrate the effectiveness and robustness of our model. In the future, we will develop more crowd features and further optimize our model to enhance the performance of our approach.

References

[1] M. Wang, X.S. Hua, R. Hong, J. Tang, G.J. Qi, Y. Song, Unified video annotation via multigraph learning, IEEE Trans. Circ. Syst. Video Technol. 19 (5) (2009) 733–746.
[2] J. Sang, C. Xu, J. Liu, User-aware image tag refinement via ternary semantic analysis, IEEE Trans. Multimedia 14 (3) (2012) 883–895.
[3] T.V. Nguyen, Z. Song, S. Yan, Stap: spatial-temporal attention-aware pooling for action recognition, IEEE Trans. Circ. Syst. Video Technol. 25 (1) (2015) 77–86.
[4] M. Wang, C. Luo, B. Ni, J. Yuan, J. Wang, S. Yan, First-person daily activity recognition with manipulated object proposals and non-linear feature fusion, IEEE Trans. Circ. Syst. Video Technol. PP (99) (2017) 1-1.
[5] Y.J. Lee, J. Ghosh, K. Grauman, Discovering important people and objects for egocentric video summarization, in: IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 1346–1353.
[6] N. Liu, J. Han, D. Zhang, S. Wen, T. Liu, Predicting eye fixations using convolutional neural networks, in: Computer Vision and Pattern Recognition, 2015, pp. 362–370.
[7] M. Jiang, J. Xu, Q. Zhao, Saliency in crowd, in: European Conference on Computer Vision. (ECCV'14), 2014, pp. 17–32.
[8] M. Wang, H. Li, D. Tao, K. Lu, X. Wu, Multimodal graph-based reranking for web image search, IEEE Trans. Image Process. 21 (11) (2012) 4649–4661.
[9] R. Kohavi, G.H. John, Wrappers for feature subset selection, Artif. Intell. 97 (1–2) (1997) 273–324.
[10] K. Ma, G. Gao, G. Ding, C.H. Liu, E. Liu, Crowd saliency prediction with optimal feature combinations, in: International Conference on Wireless Communications and Signal Processing, 2016, pp. 1–5.
[11] M. Xu, Y. Ren, Z. Wang, Learning to predict saliency on face images, in: IEEE ICCV, 2015, pp. 3907–3915.
[12] M. Jiang, S. Huang, J. Duan, Q. Zhao, Salicon: saliency in context, in: IEEE CVPR, 2015, pp. 1072–1080.
[13] R. Zhao, W. Ouyang, H. Li, X. Wang, Saliency detection by multi-context deep learning, in: IEEE CVPR, 2015, pp. 1265–1274.
[14] J. Wang, A. Borji, C.C.J. Kuo, L. Itti, Learning a combined model of visual saliency for fixation prediction, IEEE Trans. Image Process. 25 (4) (2016) 1566–1579.
[15] D. Parkhurst, K. Law, E. Niebur, Modeling the role of salience in the allocation of overt visual attention, IEEE Trans. Pattern Anal. Mach. Intell. 42 (1) (2002) 107–123.
[16] J. Zhang, S. Sclaroff, Exploiting surroundedness for saliency detection: a boolean map approach, IEEE Trans. Pattern Anal. Mach. Intell. 38 (5) (2016) 889.
[17] A. Torralba, Modeling global scene factors in attention, J. Opt. Soc. Am. A Opt. Image Sci. Vis. 20 (7) (2003) 1407–1418.
[18] A. Oliva, A. Torralba, M. Castelhano, J. Henderson, Top-down control of visual attention in object detection, in: International Conference on Image Processing, 2003, pp. 253–256.
[19] L. Zhang, M.H. Tong, T.K. Marks, H. Shan, G.W. Cottrell, Sun: a bayesian framework for saliency using natural statistics, J. Vis. 8 (7) (2008) 32-32.
[20] Y. Qin, H. Lu, Y. Xu, H. Wang, Saliency detection via cellular automata, in: IEEE CVPR, 2015, pp. 110–119.
[21] N. Bruce, J. Tsotsos, Saliency based on information maximization, in: International Conference on Neural Information Processing Systems, 2005, pp. 155–162.
[22] H. Seo, P. Milanfar, Static and space-time visual saliency detection by self-resemblance, J. Vis. 9 (12) (2009) 1–27.
[23] T. Judd, K. Ehinger, F. Durand, A. Torralba, Learning to predict where humans look, in: IEEE ICCV, 2009, pp. 2106–2113.
[24] Q. Zhao, C. Koch, Learning a saliency map using fixated locations in natural scenes, J. Vis. 11 (3) (2011) 74–76.
[25] C. Lang, T.V. Nguyen, H. Katti, K. Yadati, M. Kankanhalli, S. Yan, Depth matters: influence of depth cues on visual saliency, in: European Conference on Computer Vision, 2012, pp. 101–115.
[26] T.V. Nguyen, M. Xu, G. Gao, M. Kankanhalli, Q. Tian, S. Yan, Static saliency vs. dynamic saliency: a comparative study (2013) 987–996.
[27] K.L. Mei, V.J. Kok, C.L. Chen, C.S. Chan, Crowd saliency detection via global similarity structure, in: International Conference on Pattern Recognition, 2014, pp. 3957–3962.
[28] S. Chiappino, A. Mazzu, L. Marcenaro, C.S. Regazzoni, A bio-inspired logical process for saliency detections in cognitive crowd monitoring, in: IEEE International Conference on Acoustics, Speech and Signal Processing, 2015, pp. 2110–2114.
[29] Y. Wang, Q. Zhang, B. Li, Efficient unsupervised abnormal crowd activity detection based on a spatiotemporal saliency detector, in: IEEE Winter Conference on Applications of Computer Vision, 2016, pp. 1–9.
[30] V.J. Kok, K.L. Mei, C.S. Chan, Crowd behavior analysis: a review where physics meets biology, Neurocomputing 177 (2016) 342–362.
[31] M. Cerf, J. Harel, W. Einhuser, C. Koch, Predicting human gaze using low-level saliency combined with face detection, in: Neural Information Processing Systems, 2008, pp. 241–248.
[32] M. Cerf, E. Frady, C. Koch, Faces and text attract gaze independent of the task: experimental data and computer model, J. Vis. 9 (12) (2009) 10.1–1.
[33] C.S. Mathialagan, A.C. Gallagher, D. Batra, Vip: finding important people in images, in: Computer Vision and Pattern Recognition, 2015, pp. 4858–4866.
[34] T.H. Vu, A. Osokin, I. Laptev, Context-aware CNNs for person head detection, in: IEEE International Conference on Computer Vision, 2016, pp. 2893–2901.
[35] https://www.components.omron.com/components/web/webfiles.nsf$/FILES/AOT_OKAO_HVC.html.
[36] L. Itti, C. Koch, E. Niebur, A model of saliency-based visual attention for rapid scene analysis, IEEE Trans. Pattern Anal. Mach. Intell. 20 (11) (1998) 1254–1259.
[37] A. Liaw, M. Wiener, Classification and regression by randomforest, R News 2 (3) (2002) 18–22.
[38] R.J. Peters, A. Iyer, L. Itti, C. Koch, Components of bottom-up gaze allocation in natural images, Vis. Res. 45 (18) (2005) 2397–2416.
[39] N. Ouerhani, R. Von Wartburg, H. Hugli, R. Müri, Empirical validation of the saliency-based model of visual attention, ELCVIA: Electron. Lett. Comput. Vis. Image Anal. 3 (1) (2004) 13–24.
[40] B.W. Tatler, R.J. Baddeley, I.D. Gilchrist, Visual correlates of fixation selection: effects of scale and time, Vis. Res. 45 (5) (2005) 643–659.
[41] https://sites.google.com/site/saliencyevaluation/evaluation-measures.