Saliency in Crowd

Ming Jiang, Juan Xu, and Qi Zhao*

Department of Electrical and Computer Engineering
National University of Singapore

Abstract. Theories and models on saliency that predict where people look at focus on regular-density scenes. A crowded scene is characterized by the co-occurrence of a relatively large number of regions/objects that would have stood out if in a regular scene, and what drives attention in crowd can be significantly different from the conclusions in the regular setting. This work presents a first focused study on saliency in crowd. To facilitate saliency in crowd study, a new dataset of 500 images is constructed with eye tracking data from 16 viewers and annotation data on faces (the dataset will be publicly available with the paper). Statistical analyses point to key observations on features and mechanisms of saliency in scenes with different crowd levels and provide insights as of whether conventional saliency models hold in crowding scenes. Finally a new model for saliency prediction that takes into account the crowding information is proposed, and multiple kernel learning (MKL) is used as a core computational module to integrate various features at both low- and high-levels. Extensive experiments demonstrate the superior performance of the proposed model compared with the state-of-the-art in saliency computation.

Keywords: visual attention, saliency, crowd, multiple kernel learning

1 Introduction

Humans and other primates have a tremendous ability to rapidly direct their gaze when looking at a visual scene, and to select visual information of interest. Understanding and simulating this mechanism has both scientific and economic impact [21, 36, 7, 31].

Existing saliency models are generally built on the notion of "standing out", i.e., regions [17, 25] or objects [8, 26] that stand out from their neighbors are salient. The intuition has been successfully validated in both the biological and computational domains, yet the focus in both communities is regular-density scenarios. When a scene is crowded, however, there is a relatively large number of regions/objects of interest that would compete for attention. The mechanism that determines saliency in this setting can be quite different from the conventional principles, and saliency algorithms that completely ignore the crowd information may not be the optimal in crowded scenes.

There is hardly any work that explicitly models saliency in crowd, yet the problem has remarkable societal significance. Crowd is prevalent [24, 22] and saliency in crowd has direct applications to a variety of important problems like security, population monitoring, urban planning, and so on. In many applications, automatic systems to monitor

* Corresponding author. Email: eleqiz@nus.edu.sg

crowded scenes can be more important than regular scenes as criminal or terrorist attacks often happen with a crowd of people. On the other hand, crowded scenes are more challenging to human operators, due to the limited perceptual and cognitive processing capacity.

This paper presents a focused study on saliency in crowd. Given the evolutionary significance as well as prevalence in real-world problems, this study focuses on humans (faces). In particular, we identify key features that contribute to saliency in crowd and analyze their roles with varying crowd densities. A new framework is proposed that takes into account crowd density in saliency prediction. To effectively integrate information from multiple features at both low- and high-levels, we propose to use multiple kernel learning (MKL) to learn a more robust discrimination between salient and non-salient regions. We have also constructed a new eye tracking dataset for saliency in crowd analysis. The dataset includes images with a wide range of crowding densities (defined by the number of faces), eye tracking data from 16 viewers, and bounding boxes on faces as well as annotations on face features.

The main contributions of the paper are summarized as follows:

1. Features (on faces) are identified and analyzed in the context of saliency in crowd.
2. A new framework for saliency prediction is proposed which takes into account crowding information and is able to adapt to crowd levels. Multiple kernel learning (MKL) is employed as a core computational method for feature integration.
3. A new eye tracking dataset is built for crowd estimation and saliency in crowd computation.

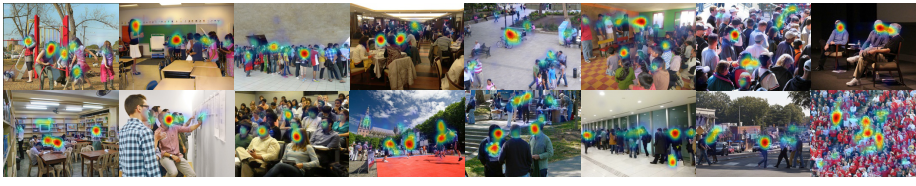

Fig. 1. Examples of image stimuli and eye tracking data in the new dataset. Note that despite the rich (and sometimes seemingly overwhelming) visual contents in crowded scenes, fixations between subjects are quite consistent, indicating a strong commonality in viewing patterns.

2 Related Work

2.1 Visual Saliency

There is an abundant literature in visual saliency. Some of the models [17, 5, 28] are inspired by neural mechanisms, e.g., following a structure rooted in the Feature Integration Theory (FIT) [35]. Others use probabilistic models to predict where humans look at. For example, Itti and Baldi [16] hypothesized that the information-theoretical

concept of spatio-temporal surprise is central to saliency, and computed saliency using Bayesian statistics. Vasconcelos *et al.* [11, 23] quantified saliency based on a discriminant center-surround hypothesis. Raj *et al.* [30] derived an entropy minimization algorithm to select fixations. Seo and Milanfer [32] computed saliency using a "self-resemblance" measure, where each pixel of the saliency map indicates the statistical likelihood of saliency of a feature matrix given its surrounding feature matrices. Bruce and Tsotsos [2] presented a model based on "self-information" after Independent Component Analysis (ICA) decomposition [15] that is in line with the sparseness of the response of cortical cells to visual input [10]. In Harel *et al.*'s work [13], an activation map within each feature channel is generated based on graph computations.

A number of recent models employed data-driven methods and leveraged human eye movement data to learn saliency models. In these models, saliency is formulated as a classification problem. Kienzle *et al.* [20] aimed to learn a completely parameter-free model directly from raw data (13×13 patches) using support vector machine (SVM) [3] with Gaussian radial basis functions (RBF). Judd *et al.* [19] learned saliency with a set of low-, mid-, and high-level features using liblinear SVM [9]. Zhao and Koch [39, 40] employed least-square regression and AdaBoost to infer weights of biologically-inspired features and to integrate them for saliency prediction.

Among all the methods, also relevant to the proposed work is the role of faces in saliency prediction. In 2007, Cerf *et al.* [5] first demonstrated quantitatively the importance of faces in gaze deployment. It has been shown that faces attract attention strongly and rapidly, independent of tasks [5, 4]. In their works as well as several subsequent models [13, 19, 39, 40], a face detector was added to saliency models as a separate visual cue, and combined with other low-level features in a linear or nonlinear manner. Saliency prediction performance has been significantly boosted with the face channel, though only frontal faces with reasonably large sizes were detected [37].

2.2 Saliency and Crowd Analysis

While visual saliency has been extensively studied, few efforts have been spent in the context of crowd. Given the specialty of crowded scenes, the vast majority works in saliency are not directly applicable to crowded scenes. The most relevant works relating to both topics (i.e., *saliency* and *crowd*) are those which applied bottom-up saliency models for anomaly detection in crowded scenes. For example, Mancas *et al.* [24] used motion rarity to detect abnormal events in crowded scenes. Mahadevan *et al.* [22] used a spatial-temporal saliency detector based on a mixture of dynamic textures for the same purpose. The model achieves state-of-the-art anomaly detection results and also works well in densely crowded scenes.

Note that although similar in name (with key words of *saliency* and *crowd*), the works mentioned above are inherently different from the proposed one. They applied saliency models to crowded scenes for anomaly detection while we aim to find key features and mechanisms in attracting attention in crowd. In a sense the previous models focused on the application of suitable bottom-up saliency algorithms to crowd while ours aims to investigate mechanisms underlying saliency in crowd and develop new features and algorithms for this topic. Furthermore, previous works relied heavily on motion and have no or limited predictability power with static scenes, while we aim to

look at underlying low- and high-level appearance features, and the model is validated with static images.

3 Dataset and Statistical Analysis

3.1 Dataset Collection

A large eye tracking dataset was constructed for saliency in crowd study (examples shown in Fig. 1). In particular, we collected a set of 500 natural crowd images with a diverse range of crowd densities. The images comprised indoor and outdoor scenes from Flickr and Google Images. They were cropped and/or scaled to a consistent resolution of 1024×768. 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.

Sixteen students (10 male and 6 female, between the ages of 20 and 30) with corrected or uncorrected normal eyesight free-viewed the full set of images. These images were presented on a 22-inch LCD monitor (placed 57cm from the subjects), and eye movements of the subjects were recorded using an Eyelink 1000 (SR Research, Osgoode, Canada) eye tracker, at a sample rate of 1000Hz. The screen resolution was set to 1680×1050, and the images were scaled to occupy the full screen height when presented on the display. Therefore, the visual angle of the stimuli was about $38.8° \times 29.1°$, and each degree of visual angle contained about 26 pixels in the 1024×768 image.

In the experiments, each image was presented for 5 seconds and followed by a drift correction. The images were separated into 5 blocks of 100 each. Before each block, a 9-point target display was used for calibration and a second one was used for validation. After each block subjects took a 5 minute break.

3.2 Statistics and Observations

The objective of the work and the dataset is to provide a benchmark for saliency studies in crowded scenes. Due to the significant role of faces, we define "crowd" based on the number of faces in a scene, and the dataset includes a wide range of crowding levels, from a relatively low density ($3 - 10$ faces per image) to a very high density (up to 268 faces per image). The varying levels of crowding in the dataset allows an objective and comprehensive assessment of whether and how eye movement patterns are modulated by crowd levels. Fig. 2(a) shows the distribution of the numbers of faces per image. To better quantify the key factors with respect to crowd levels, we sorted the images by their numbers of faces, and evenly partitioned all images into 4 crowd levels (namely, low, mid, high and very high, Fig. 2(b)).

With eye tracking experiments, we collected 15.79 ± 0.97 (mean$\pm$SD) eye fixations from each subject for each image. To analyze the fixation distribution, we constructed a fixation map of each image, by convolving a fovea-sized (i.e. $\sigma = 26$ pixels) Gaussian

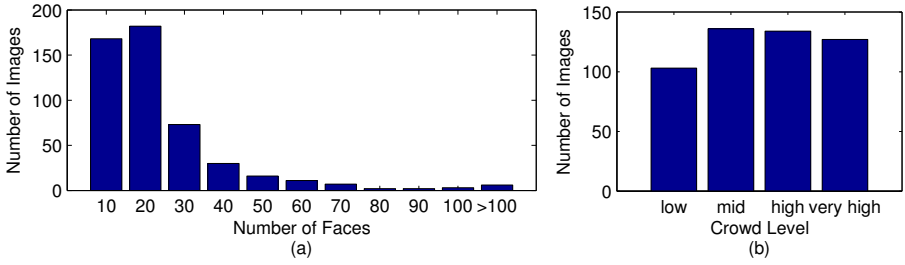

Fig. 2. (a) Histogram of face numbers per image. (b) Number of images for each crowd level.

kernel over the successive fixation locations of all subjects and normalizing it to sum 1, which can be considered as a probability density function of eye fixations.

In the following, we report key observations on where people look at in crowd:

Observation 1: Faces attract attention strongly, across all crowd levels. Furthermore, the importance of faces in saliency decreases as crowd level increases.

Consistent with previous findings [33, 19, 4, 39], the eye tracking data display a center bias. Fig. 3(a) shows the distribution of all human fixations for all the 500 images, where 40.60% of the eye fixations are in the center 16% area, and 68.99% fixations are in the center 36% area. Note that 68.58% fixations are in the upper half of the images, in line with the distribution of the labeled faces (see Fig. 3(b)), suggesting that humans consistently fixate at faces despite the presence of whole bodies.

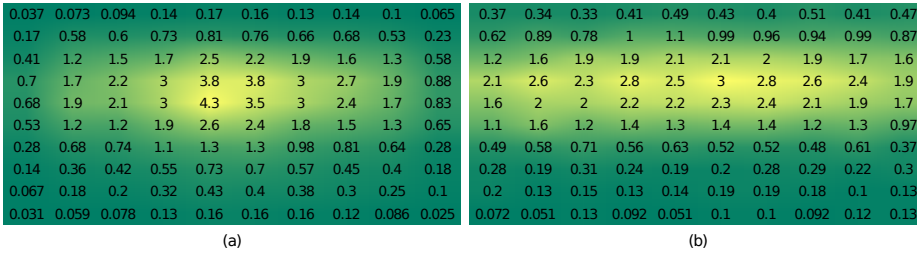

Fig. 3. Distributions of (a) all eye fixations, and (b) all faces in the dataset. The number in each histogram bin represents the percentage of fixations or faces.

We further investigated the importance of faces by comparing the mean fixation densities on faces and on the background. From Fig. 4, we observe that (1) faces attract attention more than non-face regions, consistent across all crowd levels, and (2) the importance of faces in saliency decreases with the increase of crowd densities.

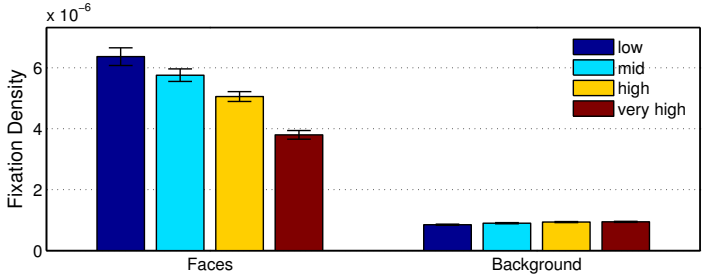

Fig. 4. Fixation densities averaged over the stimuli under the four crowd levels. Error bars indicate the standard error of the mean.

Observation 2: The number of fixations do not change (significantly) with crowd density. The entropy of fixations increases with the crowd level, consistent with the entropy of faces in a scene.

We then analyzed two global eye fixation parameters (i.e., *number* and *entropy*). Fig. 5(a) illustrates that the number of fixations does not increase with the crowd level, indicating that only a limited number of faces can be fixated at despite the larger number of faces in a crowded scene. Similarly we measured the entropy of the face as well as fixation distributions to analyze their randomness in different crowd densities. Formally, entropy is defined as $S = -\sum_{i=1}^{n} p_i log_2(p_i)$ where the vector $\mathbf{p} = (p_1, \ldots, p_n)$ is a histogram of $n = 256$ bins representing the distribution of values in each map. To measure the entropy of the original image in terms of face distributions, we constructed a face map for each image, i.e., plotting the face centers in a blank map and convolving it using a Gaussian kernel the same way as generating the fixation map. Fig. 5(b) shows that as a scene gets more crowded, the randomness of both the face map and the fixation map increases.

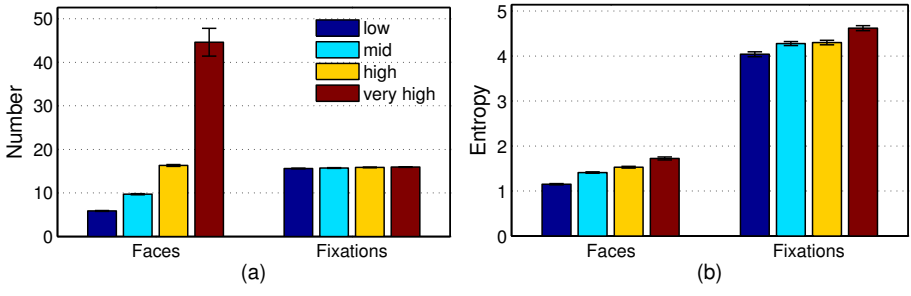

Fig. 5. (a) Numbers of faces and fixations averaged over the stimuli under the four crowd levels. (b) Entropies of faces and fixations averaged over the stimuli under the four crowd levels. Error bars indicate the standard error of the mean.

Observation 3: Crowd density modulates the correlation of saliency and features.

From Observations 1 and 2, we know that faces attract attention, yet in crowding scenarios, not all faces attract attention. There is a processing bottleneck that allows only a limited number of entities for further processing. What, then, are the driving factors in determining which faces (or non-face regions) are the most important in crowd? Furthermore, are these factors change with crowd density? While there is no previous works that systematically study these problems in the context of saliency in crowd, we aim to make a first step in this exploration. In particular, we first define a number of relevant features in the context of crowd.

Face Size. Size describes an important object-level attribute, yet it is not clear how it affects saliency - whether large or small objects tend to attract attention. In this work, we measure the face size with $d_i = \sqrt{h_i \times w_i}$, where h_i and w_i are the height and width of the i-th face.

Face Density. This feature describes the local face density around a particular face. Unlike regular scenes where faces are sparse and mostly with low local density, in a crowded scene, local face density can vary significantly in a same scene. Formally, for each face, its local face density is computed as follows:

$$f_i = \sum_{k \neq i} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\left((x_k - x_i)^2 + (y_k - y_i)^2\right)/2\sigma^2\right), \quad (1)$$

where (x_i, y_i) is the center coordinate of the i-th face, and σ is set to 2 degrees of visual angle.

Face Pose. Several recent works showed that faces attract attention [4, 39, 40], yet they all focused on frontal faces. While frontal faces are predominantly important in many regular images due to for example, photographers' preference; in a crowding setting, faces with various poses frequently appear in one scene, and to which extent pose affects saliency is.

Face Occlusion. In crowded scenes, occlusion becomes more common. While studies [18] show that humans are able to fixate on faces even though they are fully occluded, the way occlusion affects saliency has not been studied.

We then analyzed how each of the features affects saliency with varying crowd levels. Fig. 6 illustrates the saliency values (ground truth, from fixation maps) of faces with different feature values for all 4 crowd groups, and the following observations were made:

Observation 3.1 Saliency increases with face size, across all crowd levels. Intuitively in natural images, a larger size suggests a closer distance to the viewer thus is expected to be more salient. For faces of similar sizes, saliency decreases as crowd density increases.

Observation 3.2 Saliency decreases with local face density, across all crowd levels, suggesting that isolated faces attract more attention than locally crowded ones. In the same local density category, saliency decreases with global crowd density.

Observation 3.3 Generally frontal faces attract attention most strongly, followed by profile faces, and then back-view faces. Note that the "difference" of saliency values

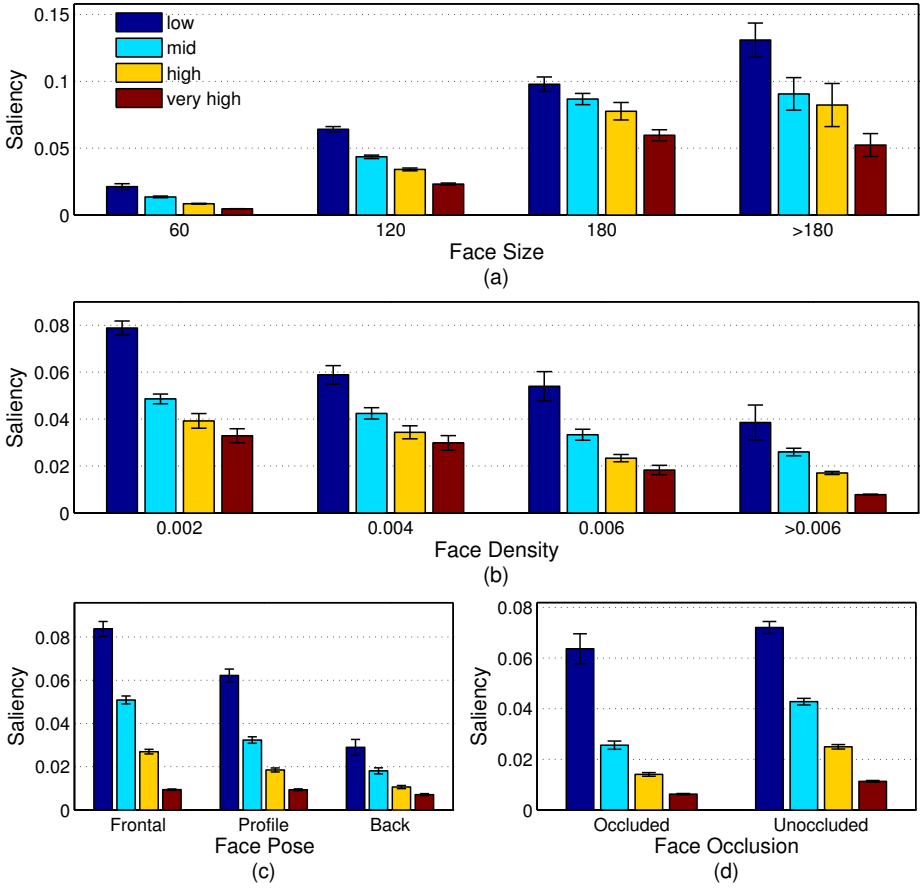

Fig. 6. Average saliency values (ground truth, from fixation maps) of faces change with (a) size, (b) density, (c) pose and (d) occlusion, modulated by the crowd levels. Error bars indicate the standard error of the mean.

of the three face categories drop monotonically with the crowd density, and for the highly crowded group, saliency with different poses are similar indicating little contribution of pose in determining saliency there. In addition, within each pose category, saliency decreases with crowd levels.

Observation 3.4 Although humans still fixate consistently on (partially) occluded faces, unoccluded faces attract attention more strongly, across all crowd levels. The saliency for both occluded and unoccluded categories decreases with crowd density.

To summarize, for all individual features, saliency on face regions decreases as crowd density increases, in consistent with Observation 2. In addition, crowd density modulates the correlation between saliency and features. The general trend is that larger faces are more salient; frontal faces are more salient than profile ones, and back-view

ones are the least salient (though saliency with different poses are similar in the most crowded group); and occluded faces are less salient than unoccluded ones. The varying importance with respect to the features is largely due to the details contained in the face regions as well as ecological reasons like experiences and genetic factors. Note that, however, the ways/parameters that characterize the trends vary significantly with different crowd levels.

4 Computational Model

In this section, we propose a computational model based on Support Vector Machine (SVM) learning to combine features automatically extracted from crowd images for saliency prediction at each pixel.

4.1 Face Detection and Feature Extraction

Section 3 suggests an important role of various face features in determining saliency, especially in the context of crowd. Despite the success in face detection, automatically detecting all the face features remains challenging in the literature. We in this work employ a part-based face detector [41] that is able to provide pose information besides the location and size of the faces. In particular, with its output face directions $\alpha \in [-90°, 90°]$, we consider faces with $|\alpha| \leq 45°$ as frontal faces and the others as profile faces. We expect that with the constant progress in face detection, more attributes like back-view, occlusion, can also be incorporated in the computational model.

With a wide range of sizes and different poses in crowd, the number of detected faces is always smaller than the ground truth, thus the partition of the crowd levels needs to be adjusted to the face detection results. As introduced in Section 3, the way we categorize crowd levels is data-driven and not specific to any number of faces in a scene, thus the generalization is natural.

Our model combines low-level center-surround contrast and high-level semantic face features for saliency prediction in crowd. 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) following Itti et al.'s approach [17]. Moreover, while Observation 1 emphasizes the importance of face in saliency prediction, Observation 2 implies that a single feature map on faces is not sufficient since only a limited number of faces are looked at despite a larger number of faces present in crowded scenes. It points to the importance of new face features that can effectively distinguish salient faces from the many faces. According to Observation 3 and the availability of face features from current face detectors (detailed below), we propose to include in the model four new feature maps on faces (size, density, frontal faces and profile faces). The face feature maps are generated by placing a 2D Gaussian component at the center of each face, with a fixed width of ($\sigma = 1°$ of visual field, 24 pixels). For size and density maps, the magnitude of each Gaussian component is the corresponding feature value computed as described in Observation 3, while for the two maps of frontal and profile faces, all Gaussian components have the same magnitude.

4.2 Learning a Saliency Model with Multiple Kernels

To predict saliency in crowd, we learn a classifier from our images with eye-tracking data, using a 10-fold cross validation (i.e. 450 training images and 50 test images). From the top 20% and bottom 70% regions in a ground truth saliency map, we randomly sample 10 pixels respectively, yielding a training set of 4500 positive samples and 4500 negative samples. The values at each selected pixel in the seven feature maps are concatenated into a feature vector. All the training samples are normalized to have a zero mean and a unit variance. The same parameters are used to normalize the test images afterwards. This sampling and normalization approach is consistent with the implementation in the MIT model [19] that learns a linear support vector machine (SVM) classifier for feature integration.

In this paper, instead of learning an ordinary linear SVM model, we propose to use multiple kernel learning (MKL) [6] that is able to combine features at different levels in a well founded way that chooses the most appropriate kernels automatically. The MKL framework aims at removing assumptions of kernel functions and eliminating the burdensome manual parameter tuning in the kernel functions of SVMs. Formally, the MKL defines a convex combination of m kernels. The output function is formulated as follows:

$$s(\boldsymbol{x}) = \sum_{k=1}^{m} \left[\beta_k \left\langle \boldsymbol{w}_k, \Phi_k(\boldsymbol{x}) \right\rangle + b_k\right] \tag{2}$$

where $\Phi_k(\boldsymbol{x})$ maps the feature data $\boldsymbol{x}$ using one of m predefined kernels including Gaussian ($\sigma = 0.05, 0.1, 0.2, 0.4$) and polynomial kernels (degree $= 1, 2, 3$), with an L1 sparsity constraint. The goal is to learn the mixing coefficients $\boldsymbol{\beta} = (\beta_k)$, along with $\boldsymbol{w} = (\boldsymbol{w}_k)$, $\boldsymbol{b} = (b_k)$, $k = 1, \ldots, m$. The resulting optimization problem becomes:

$$\min_{\boldsymbol{\beta}, \boldsymbol{w}, \boldsymbol{b}, \xi} \frac{1}{2}\Omega(\boldsymbol{\beta}) + C \sum_{i=1}^{N} \xi_i \tag{3}$$

$$\text{s.t.} \quad \forall i : \xi_i = l\left(s(\boldsymbol{x}^{(i)}), y^{(i)}\right) \tag{4}$$

where $(\boldsymbol{x}^{(i)}, y^{(i)}), i = 1, \ldots, N$ are the training data and N is the size of the training set. Specifically, $\boldsymbol{x}^{(i)}$ is the feature vector concatenating all feature values (from the feature maps) at a particular image pixel, and the training label $y^{(i)}$ is $+1$ for a salient point, or -1 for a non-salient point.

In Eq. 4, C is the regularization parameter and l is a convex loss function, and $\Omega(\boldsymbol{\beta})$ is an L1 regularization parameter to encourage a sparse $\boldsymbol{\beta}$, so that a small number of crowd levels are selected. This problem can be solved by iteratively optimizing $\boldsymbol{\beta}$ with fixed $\boldsymbol{w}$ and $\boldsymbol{b}$ through linear programming, and optimizing $\boldsymbol{w}$ and $\boldsymbol{b}$ with fixed $\boldsymbol{\beta}$ through a generic SVM solver.

Observation 3 provides a key insight that crowd level modulates the correlation between saliency and the features. To account for this, we learn an MKL classifier for each crowd level, and the use of MKL automatically adapts both the feature weights and

the kernels to each crowd level. In this work, the crowd levels are categorized based on the number of faces detected. In practice, the number of detected faces is normally smaller than the ground truth due to a wide range of sizes/poses in crowd, thus the partition of the crowd levels needs to be adjusted to the face detection results. The way we categorize crowd levels is data-driven and not specific to any number of faces in a scene, thus the generalization is natural.

5 Experimental Results

Extensive and comparative experiments were carried out and reported in this section. We first introduce experimental paradigm with the choice of face detection algorithms and implementation details, followed by metrics to evaluate and compare the models. Qualitative as well as quantitative comparative results are then shown to demonstrate the effective of the algorithm to predict saliency in crowd.

5.1 Evaluation Metrics

In the saliency literature, there are several widely used criteria to quantitatively evaluate the performance of saliency models by comparing the saliency prediction with eye movement data. One of the most common evaluation metrics is the area under the receiver operator characteristic (ROC) curve (i.e. AUC) [34]. However, the AUC as well as many other metrics are significantly affected by the center bias effect [33], so the Shuffled AUC [38] is then introduced to address this problem. Particularly, to calculate the Shuffled AUC, negative samples are selected from human fixational locations from all images in a dataset (except the test image), instead of uniformly sampling from all images.

In addition, the Normalized Scanpath Saliency (NSS) [29] and the Correlation Coefficient (CC) [27] are also used to measure the statistical relationship between the saliency prediction and the ground truth. NSS is defined as the average saliency value at the fixation locations in the normalized predicted saliency map which has zero mean and unit standard deviation, while the CC measures the linear correlation between the saliency map and the ground truth map. The three metrics are complementary, and provide a relatively objective evaluation of the various models.

5.2 Performance Evaluation

We perform qualitative and quantitative evaluation of our models with a single MKL classifier (SC-S) and a combination of multiple classifiers (SC-M) for different crowd-levels, in comparison with six classic/state-of-the-art saliency models that are publicly available.

Two of the comparative models are bottom-up ones combined with object detectors (i.e. MIT [19] and SMVJ [5], while the others are purely bottom-up, including the Itti et al.'s model implemented by Harel, the Graph Based Visual Saliency (GBVS) model [13], the Attention based on Information Maximization (AIM) model [2], the Saliency Using Natural statistics (SUN) model [38], the Adaptive Whitening Saliency (AWS)

model [12], and the Image Signature (IS) model [14]. For a fair comparison, the Viola-Jones face detector used in the MIT and SMVJ models is replaced with [41]. We also compare with the face detector as a baseline saliency model. Moreover, since the MIT saliency model and our models are both data-driven, we test them on the same training and test image sets, and the parameters used for data sampling and SVM learning are also the same. In addition, the "distance to center" channel in the MIT model is discarded to make it fair with respect to this spatial bias. Finally, all the saliency maps are smoothed with the same Gaussian kernel.

Fig. 7 shows the quantitative evaluation following Borji's implementations [1]. Further, in Fig. 8, we illustrate the ROC curves for the Shuffled AUC computation of the compared models. Four key observations are made:

1. Models with face detectors perform generally better than those without face detectors.
2. The face detector itself does not perform well enough. It only predicts a small region in the images (where the faces are detected) as salient, and the saliency of non-faces is considered to be zero. Since most of the predictions are zero, in the ROC curve for the face detector, both true positive rate and false positive rate are generally low, and there are missing samples in the right side of the curve.
3. The proposed models outperform all other models in predicting saliency in crowd (with all three metrics), suggesting the usefulness of the new face related features. The comparative models (i.e. SMVJ and MIT) use the same face detector. By combining low-level features and the face detector, SMVJ and MIT perform better than most low-level models.
4. The better performance of SC-M compared with SC-S demonstrates the effectiveness of considering different crowd levels in modeling. In fact, besides the richer set of face features, the proposed models use only three conventional low-level features, so there is still a large potential in our models to achieve higher performance with more features.

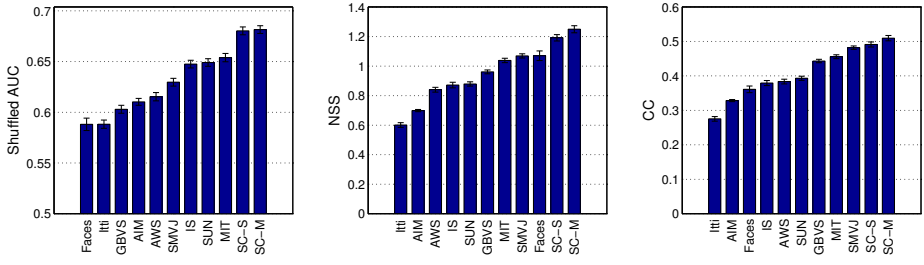

Fig. 7. Quantitative comparison of models. The prediction accuracy is measured with Shuffled AUC, NSS and CC scores. The bar values indicate the average performance over all stimuli. The error bars indicate the standard error of the mean.

For a qualitative assessment, Fig. 9 illustrates saliency maps from the proposed models and the comparative ones. First, as illustrated in the human fixation maps (2^{nd}

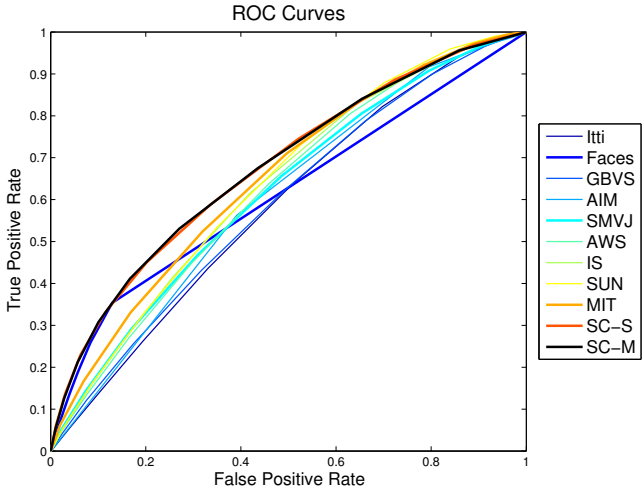

Fig. 8. ROC curves of the compared models. Bold lines represent models incorporating the face detector.

column), faces consistently and strongly attract attention. Models with face detectors (SC-M, SC-S, MIT and SMVJ) generally outperform those without face detectors (GBVS, IS, SUN, AIM and Itti). Compared with the MIT model that performs the best among all comparative models, our models use fewer low-, mid-, and high-level features, yet still perform better, demonstrating the importance of face related features in the context of crowd. Second, in scenes with relatively high crowd densities, e.g. images (f-h), there is a large variance in face size, local density, and pose, so the proposed models are more powerful in distinguishing salient faces from non-salient ones. Third, by explicitly considering crowding information in modeling, the SC-M model adapts better to different crowd densities, compared with SC-S. For example saliency prediction for faces with certain poses is more accurate with SC-M (e.g., images (c) and (i)).

6 Conclusions

The main contribution of the paper is a first focused study on saliency in crowd. It builds an eye tracking dataset on scenes with a wide range of crowd levels, and proposes a computational framework that explicitly models how crowd level modulates gaze deployment. Comprehensive analyses and comparative results demonstrate that crowd density affects saliency, and incorporating this factor in modelling boosts saliency prediction accuracy.

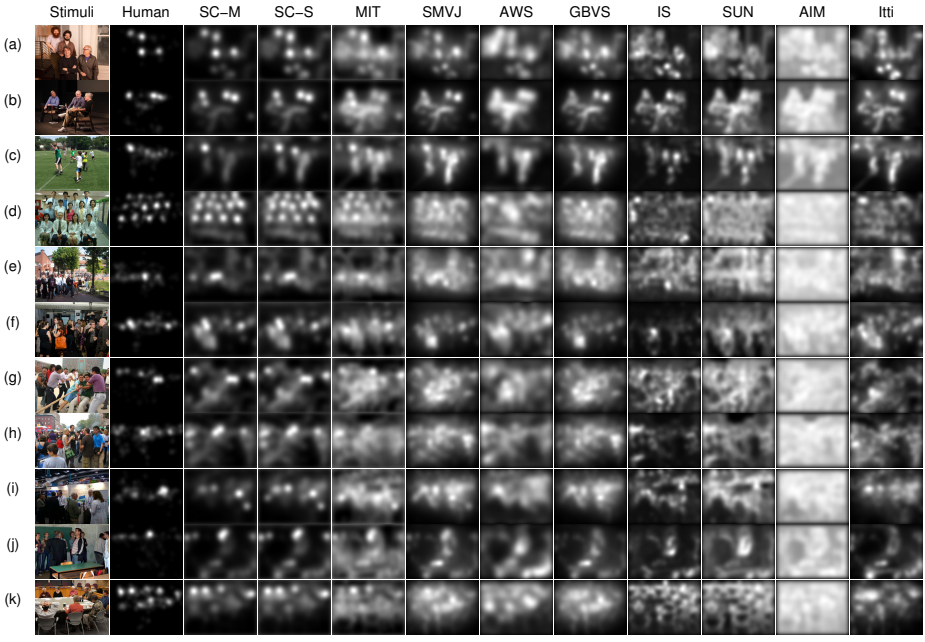

Fig. 9. Qualitative results of the proposed models and the state-of-the-art models over the crowd dataset.

Acknowledgement

This research was supported by he Singapore Ministry of Education Academic Research Fund Tier 1 (No.R-263-000-A49-112) and the Singapore NRF under its IRC@SG Funding Initiative and administered by the IDMPO.

References

1. Borji, A.: Evaluation measures. https://sites.google.com/site/saliencyevaluation/evaluation-measures
2. Bruce, N., Tsotsos, J.: Saliency, attention, and visual search: An information theoretic approach. Journal of Vision 9(3), 5 (2009)
3. Burges, C.: A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery 2(2), 121–167 (1998)
4. Cerf, M., Frady, E., Koch, C.: Faces and text attract gaze independent of the task: Experimental data and computer model. Journal of Vision 9(12), 10 (2009)
5. Cerf, M., Harel, J., Einhäuser, W., Koch, C.: Predicting human gaze using low-level saliency combined with face detection. In: NIPS (2008)
6. Chapelle, O., Vapnik, V., Bousquet, O., Mukherjee, S.: Choosing multiple parameters for support vector machines. Machine Learning 46(1-3), 131–159 (2002)
7. Chikkerur, S., Serre, T., Tan, C., Poggio, T.: What and where: a bayesian inference theory of attention. Vision Research 50(22), 2233–2247 (2010)
8. Einhäuser, W., Spain, M., Perona, P.: Objects predict fixations better than early saliency. Journal of Vision 8(14), 18 (2008)
9. Fan, R., Chang, K., Hsieh, C., Wang, X., Lin, C.: Liblinear: A library for large linear classification. Journal of Machine Learning Research 9, 1871–1874 (2008)
10. Field, D.: What is the goal of sensory coding? Neural Computation 6, 559–601 (1994)
11. Gao, D., Mahadevan, V., Vasconcelos, N.: The discriminant center-surround hypothesis for bottom-up saliency. In: NIPS (2007)
12. Garcia-Diaz, A., Fdez-Vidal, X.R., Pardo, X.M., Dosil, R.: Saliency from hierarchical adaptation through decorrelation and variance normalization. Image and Vision Computing 30(1), 51–64 (2012)
13. Harel, J., Koch, C., Perona, P.: Graph-based visual saliency. In: NIPS (2007)
14. Hou, X., Harel, J., Koch, C.: Image signature: Highlighting sparse salient regions. T-PAMI 34(1), 194–201 (2012)
15. Hyvarinen, A., Oja, E.: Independent component analysis: algorithms and applications. Neural Networks 13(4-5), 411–430 (2000)
16. Itti, L., Baldi, P.: Bayesian surprise attracts human attention. In: NIPS (2006)
17. Itti, L., Koch, C., Niebur, E.: A model for saliency-based visual attention for rapid scene analysis. T-PAMI 20(11), 1254–1259 (1998)
18. Judd, T.: Learning to predict where humans look. http://people.csail.mit.edu/tjudd/WherePeopleLook/index.html
19. Judd, T., Ehinger, K., Durand, F., Torralba, A.: Learning to predict where humans look. In: ICCV (2009)
20. Kienzle, W., Wichmann, F., Scholkopf, B., Franz, M.: A nonparametric approach to bottom-up visual saliency. In: NIPS (2006)
21. Koch, C., Ullman, S.: Shifts in selective visual attention: towards the underlying neural circuitry. Human Neurobiology 4(4), 219–227 (1985)
22. Li, W., Mahadevan, V., Vasconcelos, N.: Anomaly detection and localization in crowded scenes. T-PAMI 36(1), 18–32 (2014)
23. Mahadevan, V., Vasconcelos, N.: Spatiotemporal saliency in highly dynamic scenes. T-PAMI 32(1), 171–177 (2010)
24. Mancas, M.: Attention-based dense crowds analysis. In: WIAMIS (2010)
25. Margolin, R., Tal, A., Zelnik-Manor, L.: What makes a patch distinct? In: CVPR (2013)
26. Nuthmann, A., Henderson, J.: Object-based attentional selection in scene viewing. Journal of Vision 10(8), 20 (2010)

27. Ouerhani, N., Von Wartburg, R., Hugli, H., Muri, R.: Empirical validation of the saliency-based model of visual attention. Electronic Letters on Computer Vision and Image Analysis 3(1), 13–24 (2004)
28. Parkhurst, D., Law, K., Niebur, E.: Modeling the role of salience in the allocation of overt visual attention. Vision Research 42(1), 107–123 (2002)
29. Peters, R., Iyer, A., Itti, L., Koch, C.: Components of bottom-up gaze allocation in natural images. Vision Research 45(18), 2397–2416 (2005)
30. Raj, R., Geisler, W., Frazor, R., Bovik, A.: Contrast statistics for foveated visual systems: Fixation selection by minimizing contrast entropy. Journal of the Optical Society of America A 22(10), 2039–2049 (2005)
31. Rudoy, D., Goldman, D.B., Shechtman, E., Zelnik-Manor, L.: Learning video saliency from human gaze using candidate selection. In: CVPR (2013)
32. Seo, H., Milanfar, P.: Static and space-time visual saliency detection by self-resemblance. Journal of Vision 9(12), 15 (2009)
33. Tatler, B.: The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. Journal of Vision 7(14), 4 (2007)
34. Tatler, B., Baddeley, R., Gilchrist, I.: Visual correlates of fixation selection: Effects of scale and time. Vision Research 45(5), 643–659 (2005)
35. Treisman, A., Gelade, G.: A feature-integration theory of attention. Cognitive Psychology 12(1), 97–136 (1980)
36. Treue, S.: Neural correlates of attention in primate visual cortex. Trends in Neurosciences 24(5), 295–300 (2001)
37. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: CVPR (2001)
38. Zhang, L., Tong, M., Marks, T., Shan, H., Cottrell, G.: Sun: A bayesian framework for saliency using natural statistics. Journal of Vision 8(7), 32 (2008)
39. Zhao, Q., Koch, C.: Learning a saliency map using fixated locations in natural scenes. Journal of Vision 11(3), 9 (2011)
40. Zhao, Q., Koch, C.: Learning visual saliency by combining feature maps in a nonlinear manner using adaboost. Journal of Vision 12(6), 22 (2012)
41. Zhu, X., Ramanan, D.: Face detection, pose estimation, and landmark localization in the wild. In: CVPR (2012)