Learning a Combined Model of Visual Saliency for Fixation Prediction

Jingwei Wang, Ali Borji, *Member, IEEE*, C.-C. Jay Kuo, and Laurent Itti

***Abstract*— A large number of saliency models, each based on a different hypothesis, have been proposed over the past 20 years. In practice, while subscribing to one hypothesis or computational principle makes a model that performs well on some types of images, it hinders the general performance of a model on arbitrary images and large-scale data sets. One natural approach to improve overall saliency detection accuracy would then be fusing different types of models. In this paper, inspired by the success of late-fusion strategies in semantic analysis and multi-modal biometrics, we propose to fuse the state-of-the-art saliency models at the score level in a para-boosting learning fashion. First, saliency maps generated by several models are used as confidence scores. Then, these scores are fed into our para-boosting learner (i.e., support vector machine, adaptive boosting, or probability density estimator) to generate the final saliency map. In order to explore the strength of para-boosting learners, traditional transformation-based fusion strategies, such as *Sum*, *Min*, and *Max*, are also explored and compared in this paper. To further reduce the computation cost of fusing too many models, only a few of them are considered in the next step. Experimental results show that score-level fusion outperforms each individual model and can further reduce the performance gap between the current models and the human inter-observer model.**

***Index Terms*— Saliency, bottom-up attention, regions of interest, eye movements, score level fusion, para-boosting learner, scene understanding.**

I. Introduction

PREDICTING where humans look in viewing natural scenes, known as saliency prediction or detection, has attracted a lot of interest in computer vision, robotics, and neurosciences. High saliency prediction accuracy is essential for many applications such as salient object detection and segmentation, content-aware media retargeting, content-based compression, scene understanding, robot navigation, and image/video quality assessment. Please see [1]–[12] for comprehensive reviews).

In existing literature, many saliency models have been proposed based on purely bottom-up image cues, top-down semantic cues and naturally the combination of both types of cues. These models have used various visual features including low-, middle- and high-level features. However, every single model has its own hypothesis and methodology focusing on a different aspect of human visual attention. Further, researchers have usually chosen few datasets to develop their models. Thus dataset bias [13] is often reflected in their models. Due to these assumptions and constraints, each model has its own favorite image category [14] and the corresponding weak cases or false positives (see some examples in Fig. 1). Therefore, one approach to construct a more predictive model is to combine weaker models. While a few studies have examined this in the past in limited cases, here we address it in a more systematic manner examining a larger number of combination methods and parameters.

A. Models Based on Low-Level Features

Bottom-up saliency models using biologically-plausible low-level features are mainly based on computational principles proposed by Itti et al. [15] and its predecessor (e.g., [16], [17]). Assuming that salient regions are conspicuous in terms of color, intensity, or orientation, Itti et al. derived bottom-up visual saliency using center-surround differences across multi-scale image features. Harel et al. [18] built a Graph Based Visual Saliency (GBVS) model based on the idea that local image patches highly dissimilar to their surrounding patches are salient. They measured the dissimilarity among local patches using a Markov chain method. Similarly, Liu et al. [19] used Conditional Random Fields (CRF) to identify regions of interest (ROI) using three types of features: *multi-scale contrast*, *center-surround histogram*, and *color spatial-distribution*. Based on the idea that fixations are attracted to rare local image patches in natural images, Bruce and Tsotsos [20] proposed the *Attention for Information Maximization (AIM)* saliency model. Hou and Zhang [21] and Zhang et al. [22] proposed models based on the same principle as in the AIM model utilizing sparse representation of local image patches. Kong et al. [23] proposed a multi-scale intergration strategy to combine various low-level saliency features at different levels. Fang et al. [24] derived a compressed domain saliency detection method using motion and color features. Assuming that visual attention area of interest has a shape prior, Jiang et al. [25] detected salient regions using contour energy computation. The above-mentioned models perform well, but with the limitation of not considering high-level object and semantic cues.

Manuscript received June 29, 2014; revised July 3, 2015, September 7, 2015, and November 17, 2015; accepted January 22, 2016. Date of publication January 27, 2016; date of current version February 23, 2016. This work was supported in part by the National Science Foundation under Grant CCF-1317433 and Grant CNS-1545089, in part by the Army Research Office under Grant W911NF-12-1-0433, and in part by the Office of Naval Research under Grant N00014-13-1-0563. *(Jingwei Wang and Ali Borji contributed equally to this work.)*

J. Wang and C.-C. Jay Kuo are with the Department of Electrical Engineering, University of Southern California, Los Angeles, CA 90089 USA (e-mail: wang1984wei@gmail.com; cckuo@sipi.usc.edu).

A. Borji is with the Department of Computer Science, University of Central Florida, Orlando, FL 32816 USA (e-mail: ali.borji@ucf.edu).

L. Itti is with the Department of Computer Science and the Department of Psychology, Neuroscience Graduate Program, University of Southern California, Los Angeles, CA 90089 USA (e-mail: itti@usc.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2016.2522380

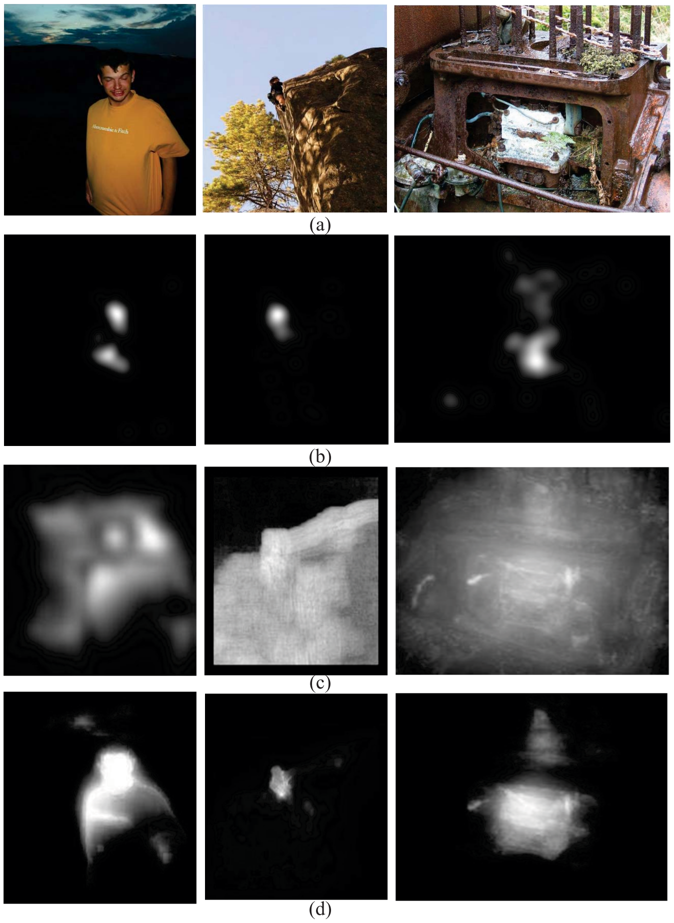

Fig. 1. Illustration of inaccurate cases of saliency models (image example from MIT dataset [28]). (a) Original images, (b) Ground-truth human fixation map, (c) Failed saliency prediction cases for three models, from left to right (Itti [15], AIM [20], and Judd [28], and (d) Our Para-boosting learner results. Compared to individual models, score-level fusion results are closer to human eye movement patterns.

As a result of different assumptions, each model has its own most suitable image category [14]. For example, focusing on local dissimilarities, Itti [15] and GBVS [18] models fall short in detecting global visual saliency, while the AIM [20] model shows weakness in detecting local dissimilarities (i.e., local vs. global saliency [26]). In general, these bottom-up models fail to detect salient regions with semantic information, such as humans, animals, faces, objects, texts, and signs. Such failure can be seen in the Fig. 1, row (c), where itti model cannot distinguish a human body from its surroundings.

B. Models Based on Low- and High-Level Features

Several researchers have addressed saliency detection from a top-down point of view by incorporating both low- and high-level features. Cerf et al. [27] showed performance improvements by adding a high-level factor, face detection, to the Itti's model. By adding more high-level features such as faces, people and text, in conjunction with other low- and middle-level features, Judd et al. [28] developed a saliency detection model by learning best weights for all combined features using Support Vector Machines (SVM). Similarly, based on the finding that observers tend to look at the center of objects [29], Chang et al. [30] proposed an object-based saliency model using an objectness measure [31]. Although top-down models highlight the importance of high-level and semantic features (e.g., faces, animals, cars, text, symmetry [32], object center-bias [33], and image center-bias [34]), they often fail to detect salient objects for which they have not been trained. Further, performance of the combined models in detecting non-semantic bottom-up salient regions may not be as as good as purely bottom-up models. For example, Judd [28] model fails to detect the local dissimilarity well in some cases (Fig. 1 row (c)). Considering that no individual model is able to fit to all images, fusion of different models has the potential of improving the overall saliency prediction performance.

C. Learning-Based Saliency Models

Saliency models usually, followed by Itti et al. [15], Koch and Ullman [16], and Feature Integration Theory (FIT) [17], first, extract a set of visual features such as contrast, edge content, intensity, and color for a given image. Then, they apply a spatial competition mechanism via a center-surround operation (e.g., using Difference of Gaussian filters) to quantify conspicuity in a particular feature dimension. Finally, they linearly (with equal weights) integrate conspicuity maps to generate a scalar master saliency map (e.g., [15]–[17], [26], [27], [35]). Some researchers have proposed "max" type of integration (e.g., [36]). Under the linear assumption, Itti and Koch [37] have proposed various ways to normalize the feature maps based on map distributions before linearly integrating them.

Instead of linear combination with equal weights, some models learn weights for different channels from a set of training data. Judd et al. [28] used low-level image features, a mid-level horizon detector, and two high-level object detectors (faces using [38] and humans using [39]) and learned a saliency model with liblinear SVM. Following Judd et al., Zhao and Koch [40] learned feature weights using constraint linear regression and showed enhanced results on different datasets using different sets of weights. Later, Borji [41] proposed an AdaBoost [42] based model to approach feature selection, thresholding, weight assignment, and integration in a principled, nonlinear learning framework. The AdaBoost-based method combines a series of base classifiers to model the complex input data.

Some models directly learn a mapping from image patches to fixated locations. Followed by Reinagel and Zadore's study [43] who proposed that fixated patches have different statistics than random patches, Kienzle et al. [44], learned a mapping from patch content to whether it should be fixated or not (i.e., +1 for fixated and −1 for not-fixated). They learned a completely parameter-free model directly from raw data using support vector machine with Gaussian radial basis functions (RBF). On the other hand, Zhang et al. [45] use a graphlet-based deep architecture to learn a saliency map from raw imge pixels to object-level graphlets (oGLs) and further to spatial-level graphlets (sGLs).

In [19], Conditional Random Field (CRF) that encodes interaction of neighboring pixels effectively in images, have been used for combining cues for salient object detection.[1] Similarly, Mai et al. [46] proposed a data-driven approach under a Conditional Random Field (CRF) framework to aggregate individual saliency maps, which focused on learning the interactive relationship between neighboring pixels by learning from groundtruth saliency map.

Borji et al. [14] considered two simple combination methods (sum and multiplication) using three normalization schemes (identity, exponential, and logarithmic) for combining saliency models. Bayesian integration, and linear weighted combination of different cues have been also used in some studies for salient object detection (e.g., [48]–[50]). Olivier et al. [47] investigated the supervised/unsupervised learning methods to aggregate 8 low-level feature(bottom-up) based saliency models, and they found that the aggregation performance is improved when there is a similarity in the training samples.

In the context of visual search or object detection, some researchers have proposed approaches for learning a set of parameters (gains or weights of saliency channels) to render an object of interest more salient (e.g., [35], [51]–[53]). Some researchers have used evolutionary optimization algorithms to find optimal set of combination weights [53].

D. Our Contributions

We present three major contributions. Firstly, we propose to fuse multiple saliency models' outputs, called para-boosting. Through exhaustive evaluation over challenging benchmark databases, combining both low-level and high-level feature-based saliency models, we show that para-boosting at the score level outperforms the individual state-of-the-art saliency prediction models, and achieves closer accuracy to the human inter-observer model. Secondly, several para-boosting strategies including transformation-based and learning-based fusion schemes (including joint density estimation based and linear/nonlinear classifier-based), are proposed and compared in this paper to investigate several possible fusing strategies, For example, according to the experimental results, our proposed learning-based schemes perform the best among different fusion schemes. Thirdly, we investigate the role of each individual model during the fusing procedure. Experimental results demonstrate that models do not play equal roles in the final decision, which provides the possibility of fusing fewer models while maintaining similar performance. Corresponding experimental results show that the integration of a few best of the individual models can sometimes outperform fusing all models.

II. Para-Boost for Saliency Learning

Inspired by the success of late fusion strategies in semantic analysis and multi-modal biometrics, we propose to use para-boosting for combining several saliency models at the confidence score level. Our choice is based on the two following reasons. First, the computation and combination flexibility of different saliency models can be maintained as much as possible using the late fusion strategy. Feeding low-level features directly into a black box machine learning method such as SVM for combining them (as in [28]) has the disadvantage that strength of the original computation on individual features will be eliminated. Also, in each individual model, feature subsets are manipulated differently to its own best in predicting saliency while early feature combination has the same manipulation schema on all features. Second, variation among different models can be maintained well in a late fusion stage. Even when sharing similar feature subsets, different computation models may generate different saliency maps. For example, although both Itti and HouCVPR models use illumination information, Itti model describes the saliency relationship in spatial domain when HouCVPR describes it in the FFT domain. This variation can be preserved to a large extent in a late fusion strategy.

Thus, to fill the performance gap between early feature-level fusion and the human inter-observer model, and to broaden applicable image ranges, we propose to fuse state-of-the-art saliency models at the score level. In our saliency detection scenario, saliency map generated by each model is regarded as a single score result, and score level fusing schema is then applied to boost saliency detection performance by taking different saliency detection models' outputs into consideration. With the aim of learning the influence of different learning choices, we further investigate various learning techniques (SVM, AdaBoosting, and PDE), and evaluate their prediction performance in this scenario.

Making use of the prediction results of different recognition systems, score level fusion strategies have already been broadly applied in a variety of biometric systems, and they can further improve the detection and recognition performance compared to a single system [55]. Here we present a brief overview of current score level fusion strategies in biometric systems. In general, there are three different types of score level fusion strategies described below:

- Transformation-Based Approaches: In transformation-based score fusion approaches, scores are first normalized to a common range for further combining. Choice of the normalization scheme depends highly on the input data itself ([55], [56]). Kittler et al. [57] discussed a fusion framework by evaluating the *sum rule, product rule, minimum rule, maximum rule, median rule, and majority voting rule* in their work. In their proposed scheme, scores are converted into posteriori probabilities through normalization. It has been experimentally shown that the sum rule outperformed other rules in biometric system applications.
- Classification-Based Approaches: In this scheme, scores from individual models are considered as feature vectors of a classifier, which are constructed to further improve detection accuracy ([58], [59]). Chen and Rao [60] used a neural network classifier to combine the scores from the face and iris recognition systems. Wang and Han [69] further proposed to apply a classification-based algorithm based on SVM to fuse the scores.

[1] Note that salient object detection models attempt to detect and segment the most salient object while fixation prediction models aim to predict sparse locations that observers may fixate. In this paper, we are interested in fixation prediction.

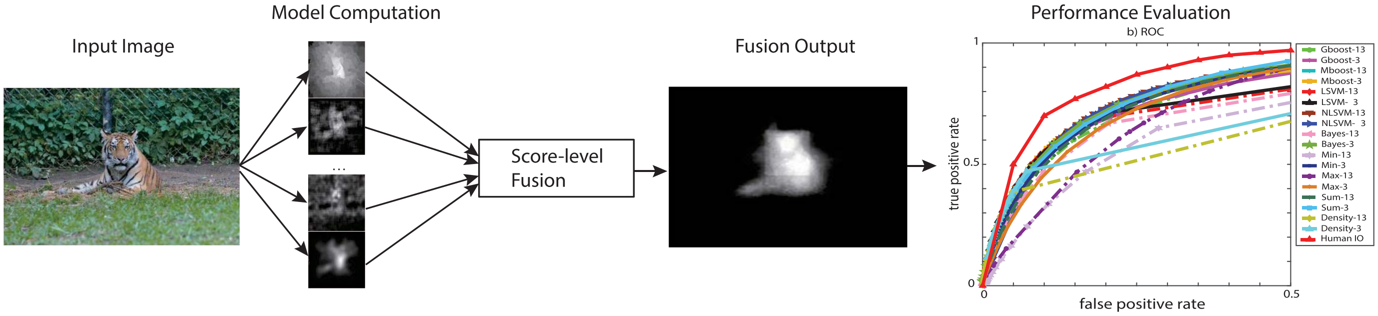

Fig. 2. Illustration of our saliency fusion framework (image example from MIT dataset [28]).

- Probability density-based approaches: Well-known probability models, such as naive Bayesian [61] and the Gaussian Mixture Models (GMM) [62], have been broadly applied for model fusion. Nandakumar et al. [62] proposed a score combination framework based on the likelihood ratio estimation. The input score vectors are modeled as a finite Gaussian mixture model. They show that this density estimation method achieved good performance on biometric databases with face, fingerprint, iris and speech modalities. Probability-density based score fusion methods highly depend on accuracy of score's probability density estimation.

Fig. 2 shows an illustration of our proposed para-boosting framework. We first retrieve the saliency maps of 13 state-of-the-art approaches, and these saliency maps are considered as input to our para-boosting system. Then, a final saliency map is predicted based on the score-level fusion schema. To investigate effectiveness of this framework, we employ two categories of score level fusion schemes including transformation-based and learning-based schemes. It is worth noting that the input to the para-boosting system consists of only 13 saliency probability maps, which is relatively low compared to other combination methods (e.g., [28]) which helps harness overfitting during learning.

A. Visual Features

Without loss of generality, following the discussions of [14], we choose 13 state-of-the-art models (low-level feature based: Itti [15], GBVS [18], AIM [20], HouNips [21], HouCVPR [54], AWS [63], SUN [22], CBS [25], SalientLiu [19] and SO [64]; High-level feature based: SVO [30], ST [65], and Judd [28]) based on the following two criteria.

- The selected models must achieve good prediction performance individually, and
- The features adopted by selected models must cover a wide range of features from bottom-up to top-down, so that our final fusion strategy can adjust to different image scenarios.

1) Low-Level Features: Based on their own space domain, these low-level saliency maps can be classified into three categories:

- Itti [15] and GBVS [18] models construct pixel-based saliency maps in the spatial domain. Specially, Itti [15] model calculates a saliency map using color, intensity, and orientation features in several scales (in total there are 42 feature maps used in this model). GBVS [18] models saliency as an activation map which uses up to 12 feature maps as input.
- HouCVPR [54], AIM [20], AWS [63], HouNips [21], and SUN [22] generate pixel-based saliency maps in the frequency domain. In particular, HouCVPR [54] and AWS [63] models compute saliency in the Fast Fourier transform (FFT) domain. The input features can be one single map (gray image) or three maps (RGB). AIM [20], HouNips [21], and SUN [22] models learn a dictionary of natural scene patches using RGB or DoG (Difference of Gaussian) channels. They then use RGB maps (3 feature maps) or DoG maps (12 feature maps) as input.
- SalientLiu [19], SO [64], and CBS [25] are region-based saliency models. SalientLiu [19] employs a broad range of features from local multi-scale contrast to global center-surrounding distributions. Similar to SalientLiu [19], CBS [25] applies several features such as color superpixels, closed shape, and center-bias.[2] Both of the two models use at least 3 feature maps as their input. SO [64] characterizes the spatial layout of image regions with respect to image boundaries, and optimized saliency map over multiple low-level cues including spatial layout of image regions.

2) High-Level Feature: As we mentioned in the Introduction section, some models use semantic information when building their saliency maps. Among 13 used models, SVO [30] builds a model that can describe the objectness [31]. This model uses semantic object detection results as well as saliency prior as input. Similarly, ST [65] uses object prior and other global measurement to estimate region similarities. Judd [28] model employs as many as 33 feature maps ranging from low-level to high-level features to learn a saliency map.

Note that our classification of models based on low- and high-level features is not crisp. Indeed, several models take advantage of both types of features (e.g., Judd, CBS).

B. Score-Level Fusion Strategies

Here we propose three types of score-level fusion strategies applied in para-boosting of saliency detection: 1) Non-learning

[2] Tendency of observers to preferentially look at the image center.

based fusing approaches such as transformation based fusion, 2) Learning based fusion approaches using pattern classifiers and 3) Density based approaches. These approaches fuse multiple score inputs from different viewpoints, hence it is worthwhile to test and compare their powers in score fusion.

1) Transformation-Based Fusion: As indicated in [55], the transformation based rule involves two steps: normalization and fusing rules. As shown in Fig. 2, our input individual saliency results is a probability map. We adopt the normalization method described in Eq. 1 to normalize each individual saliency map to have zero mean and unit standard deviation:

$$s' = \frac{s - \mu}{\sigma} \tag{1}$$

Where μ and σ are the mean and standard deviation (STD) of input map s. We adopt three different transformation-based fusion rules that have been reported to have good performance in biometric recognition systems.

- Sum rule: In this rule, the final score output is computed as the mean of input score sequence:

$$S = mean(s_1, \ldots, s_n) \tag{2}$$

  where s_k indicates each individual score and S is the final score output.
- Min rule: Here, the final score is the minimum value among all input scores:

$$S = min(s_1, \ldots, s_n) \tag{3}$$

- Max rule: Contrary to the min rule, the final score output here is the max value among all individual scores:

$$S = max(s_1, \ldots, s_n) \tag{4}$$

2) Classification-Based Fusion: We apply two different types of classifier: *linear* and *non-linear*. Linear classifiers are usually fast in computation while non-linear classifiers are usually slower but more powerful. We built a training set by sampling images at eye fixations (i.e., ground truth). Each sample contains 13 individual saliency probability at one pixel together with a $0/+1$ label (i.e., a 13D vector). Positive samples are taken from the top p percent salient pixels of the human fixation map and negative samples are taken from the bottom q percent. Following [28], we chose samples from the top 5% and bottom 30% in order to have samples that were strongly positive and strongly negative. Train vectors were normalized to have zero mean and unit standard deviation. The same parameters were used to normalize the testing data.

- SVM. Here we train two Support Vector Machines classifiers using publicly available Matlab versions of SVM: liblinear and libsvm. We adopted both the linear kernel (Eq. 5) and non-linear kernel Radial Basis Function (RBF) (Eq. 6) as they have been shown to perform well in a broad range of image applications. During testing, instead of predicting binary labels, we generate a label whose value is within the range of [0, 1], so that the final output is a saliency probability map:

$$K(x_i, x_j) = \alpha \cdot x_i^T x_j \tag{5}$$

$$K(x_i, x_j) = \exp(-\gamma \left\| x_i - x_j \right\|_2^2), \quad \gamma > 0 \tag{6}$$

  where α and γ are the parameters of the kernel function.
- AdaBoosting. To further investigate the non-linear classifiers' capability in fusion, we used AdaBoost algorithm [41], [42], which has been broadly applied in scene classification and object recognition. AdaBoost combines a number of weak classifiers h_t to learn a strong classifier:

$$H(x) = sign(f(x)); \quad f(x) = \sum\nolimits_{t=1}^{T} \alpha_t h_t(x) \tag{7}$$

  where α_t is the weight of the $t_t h$ classifier. Here, the number of weak classifiers T is set as 10 to balance the speed and accuracy. As in SVM, we consider the real value of $H(x)$ to create a saliency map (i.e., $f(x)$). We used the publicly available software for Gentle AdaBoost and Modest AdaBoost.[3]

3) Density-Based Fusion:

- Naive Bayesian. Assuming the independence among different saliency models M_k, then a Naive Bayesian accumulation model [14] can be built in Eq. 8:

$$p(x \,|\, M_1, M_2, \ldots, M_K) \propto \frac{1}{Z} \prod_{k=1}^{K} p(x \,|\, M_k) \tag{8}$$

  Here $p(x \,|\, M_1, M_2, \ldots, M_K)$ indicates the final fusion probability for each pixel, and $p(x \,|\, M_k)$ is the probability of each pixel observation from each individual model. K is the number of models and Z is a normalization factor. Since a very small value from a single model will suppress all other models, here we apply a modified Bayesian accumulation (Eq. 9) to damp the attenuation power of very small values such as 0:

$$p(x_f \,\middle|\, M_1, M_2, \ldots, M_K) \propto \frac{1}{Z} \prod_{k=1}^{K} (p(x_f \,\middle|\, M_k) + 1) \tag{9}$$

- General density estimation. Without assuming independence among saliency models, we propose to fuse different models based on join density estimation of their confidence outputs for the final saliency map [66]. Firstly, we classify the training samples into two classes: *non-salient* (c_0) and *salient* (c_1). Each sample has the d-dimensional feature vector if there are d different models. Then, for both the non-salient class c_0 and salient class c_1, we estimate their corresponding density function using Parzen-window density estimation as in Eq. 10:

$$\mathrm{P(x)} = \frac{1}{\mathrm{nh^d}} \sum_{i=1}^{n} K(\frac{x - x_i}{h}) \tag{10}$$

  where, n is the number of observations, h is the corresponding window width and $K(x)$ is a non-negative window function or kernel function in the d-dimensional space:

$$s.t. \int_{R^d} K(x) dx = 1 \tag{11}$$

[3] http://graphics.cs.msu.ru/en/science/research/machinelearning/adaboosttoolbox

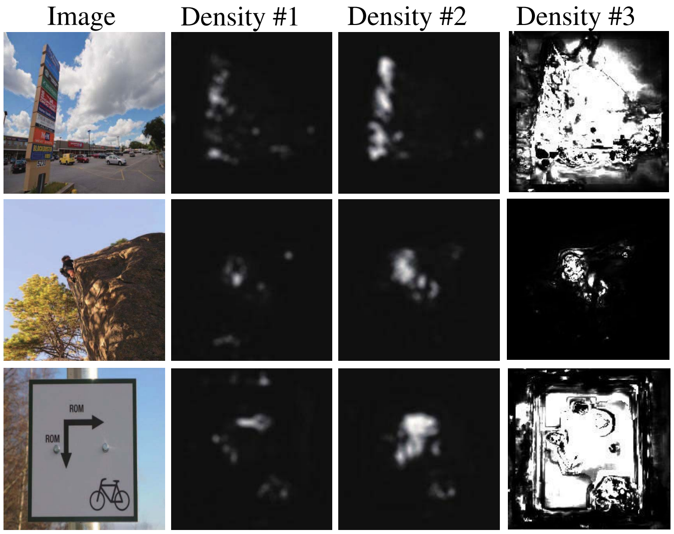

Fig. 3. Density results with different training sample set, from left to right, top 5%-bottom 30%, top 30%-bottom 30%, top 30%-bottom 70%.

Finally, the likelihood ratio $L = P(x|c_1)/P(x|c_0)$ is employed as the final confidence fusion score.

In the paper, we use the PRTools [67] to perform the Parzen-window density estimation. We used Gaussian Kernel and an optimum smoothing parameter h based on the observations to estimate the density. Furthermore, during our test trial (Fig. 3), our proposed density model performs better when we keep top 30%-bottom 30% samples for training. So, here we choose top 30%-bottom 30% for practical reasons.

III. Experimental Results

A thorough evaluation of different score fusion strategies is presented in this section. We also compute the human Inter-Observer (human IO) model as the upper-bound baseline [41]. In this model, we estimate the quality of each subject's saliency map result by using the ground truth saliency map generated by all other subjects. Then we average the individual measurement score of all subjects.

A. Datasets

We utilize two benchmark datasets. the MIT [28] and Imgsal [68], which have a broad range of images for the purpose of fair model comparison. The MIT dataset contains 1003 images collected from Flicker and LabelMe datasets. The ground truth saliency maps are generated using the eye fixation data collected from fifteen different human subjects. Several sample images from the MIT dataset [28] are shown in Fig. 4. The Imgsal dataset has 235 images collected from the web. The ground truth saliency maps are produced using eye fixation data from nineteen naive subjects. Several sample images from Imgsal dataset [68] are shown in Fig. 5. Images in this dataset are carefully collected with salient region size ranging from small, medium to large, including both repeated and random background clutter patterns.

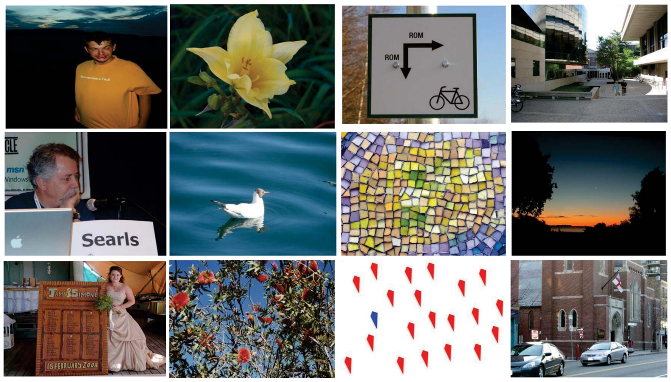

Fig. 4. Sample images from the MIT dataset.

Fig. 5. Sample images from the Imgsal dataset.

These two datasets cover many image scenarios, ranging from street view, human face, various objects to synthesized patterns. Due to this fact, saliency models which perform well over MIT and Imgsal datasets were expected to perform well for a large range of image types.

B. Evaluation Metrics

To compare the performance, we show two different measurements: ROC and Precision-Recall. The ROC curve and AUC score have been broadly applied in eye fixation prediction [41], and precision-recall is applied here as a pair-wise measurement with the ROC curve. A good fusion model should perform well over all measurement metrics.

- Precision-Recall: The final fusing saliency map is a probability map with values within [0,1]. Thus, to compare with ground truth eye fixation map, we generate a binary saliency map by comparing each value with a threshold. By varying a threshold within (0 : 0.1 : 1], different binary saliency maps can be produced. To avoid any bias over threshold, here we calculate the average precision-recall curve over all 10 threshold values as our final score. The calculation of precision and recall follows the Eq. 12 and Eq. 13:

$$precision = \frac{true\_positive}{true\_positive + false\_positive} \tag{12}$$

$$recall = \frac{true\_positive}{true\_positive + false\_negative} \tag{13}$$

- ROC and AUC: We calculate the Receiver Operating Characteristic (ROC) and Area Under ROC Curve (AUC) results in the form of true positive and false positive

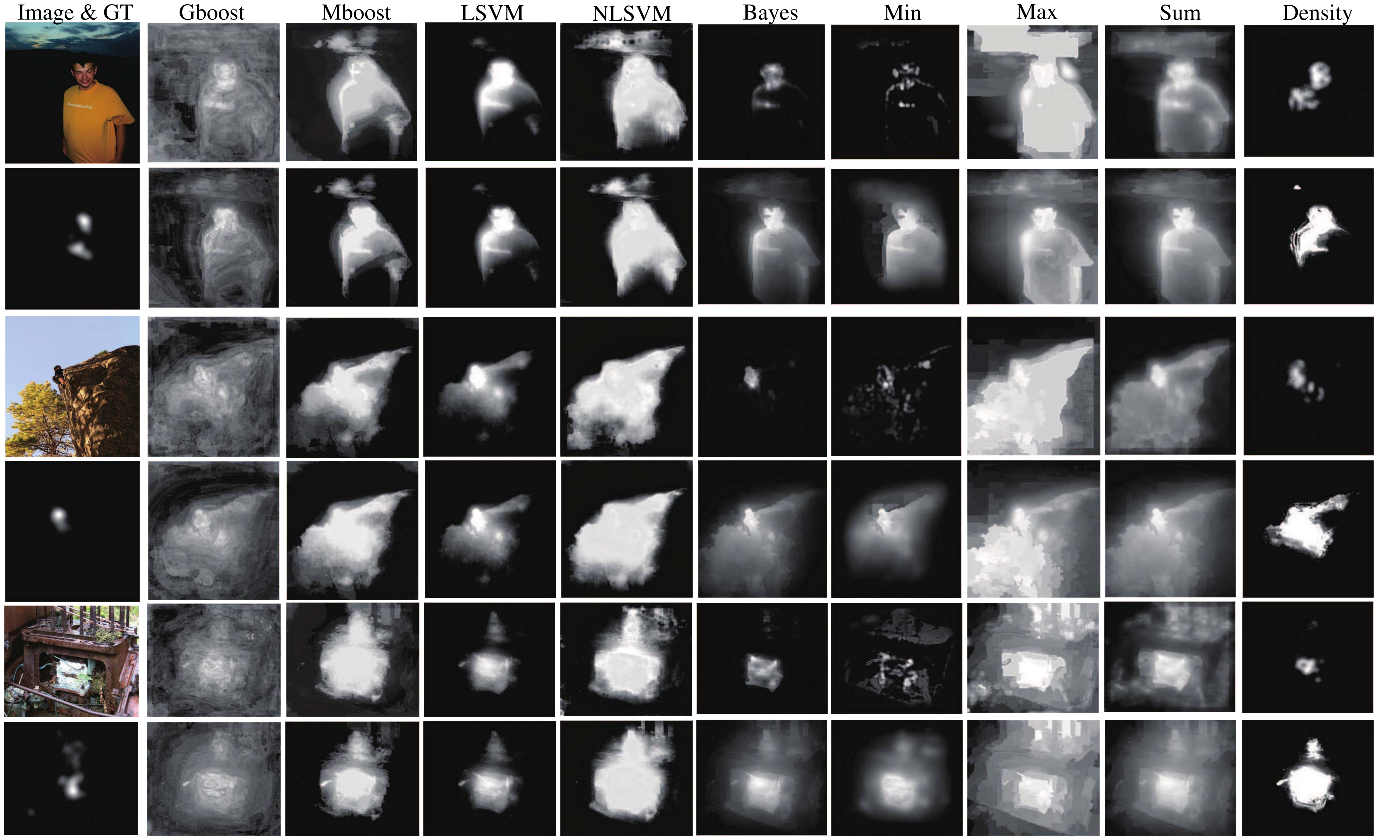

Fig. 6. Score-level fusion results (image examples from the MIT dataset [28]) and Imgsal [68]). For each image, the fist row shows original image and fusion results using top 13 models, and the second row shows ground truth and fusion results using top 3 models.

obtained during the calculation of precision-recall. The ROC curve is drawn as the true positive rate (TPR) vs. false positive rate (FPR), and the area under this curve (AUC) indicates how well the saliency map matches with human fixations:

$$TPR = \frac{true\_positive}{true\_positive + false\_negative} \quad (14)$$

$$FPR = \frac{false\_positive}{false\_positive + true\_negative} \quad (15)$$

1) Average Performance: For fusion methods involving training, such as the classifier based and general density based approaches, we followed a cross-validation procedure. Both the MIT [28] and Imgsal [68] datasets are combined first and then the combination is divided into 5 parts. Each time we trained the model over 4 parts and tested it on the remaining part. Results are then averaged over all partitions. First, we compare the performance of individual models and then fusion models.

C. Model Comparison

Figs. 6 and 7 illustrate sample images along with individual saliency maps and results of our fusion saliency model. Fig. 8 shows average Precision-Recall (PR) and ROC curves of different models over the MIT [28] and Imgsal [68] datasets. It is clear that most of our proposed para-boost strategies outperform individual models. To summarize and better highlight the gap between models and humans, we draw the AUC score on the MIT and Imgsal datasets in Fig. 9. In this figure, as different human subjects may have different opinions on salient regions, we further perform cross-validation on all human eye fixation data, resulting in HIO (human IO) model, which serves as our performance upper-bound as humans usually highly agree with each other.

In Fig. 9, fusion strategies such as gentle adaboost (Gboost), modest adaboost (Mboost), linear SVM (LSVM), nonlinear SVM (NLSVM), Naive Bayesian (Bayes), sum and general density (density) either outperform or strongly compete with any individual saliency model. Particularly, linear SVM and modest adaboost (Mboost) are top contenders among all the para-boosting strategies and are closest models to the Human IO model. It further proves that linear SVM and modest adaboost show their advantages in providing better generalization capability and resistance to overfitting compared to other prediction models. It is worthwhile to point out that linear SVM outperforms non-linear SVM. Rather than the non-linear relationship between input raw features (e.g. color, position) and fixation prediction, the relationship between the model-fused prediction and the individual prediction model should be linear, it is more reasonable to apply a linear regression prediction model to boost performance.

Among all score level fusion models, the learning based methods such as Gboost, Mboost, LSVM, NLSVM and density outperform Naive Bayes and Sum, which are non-learning

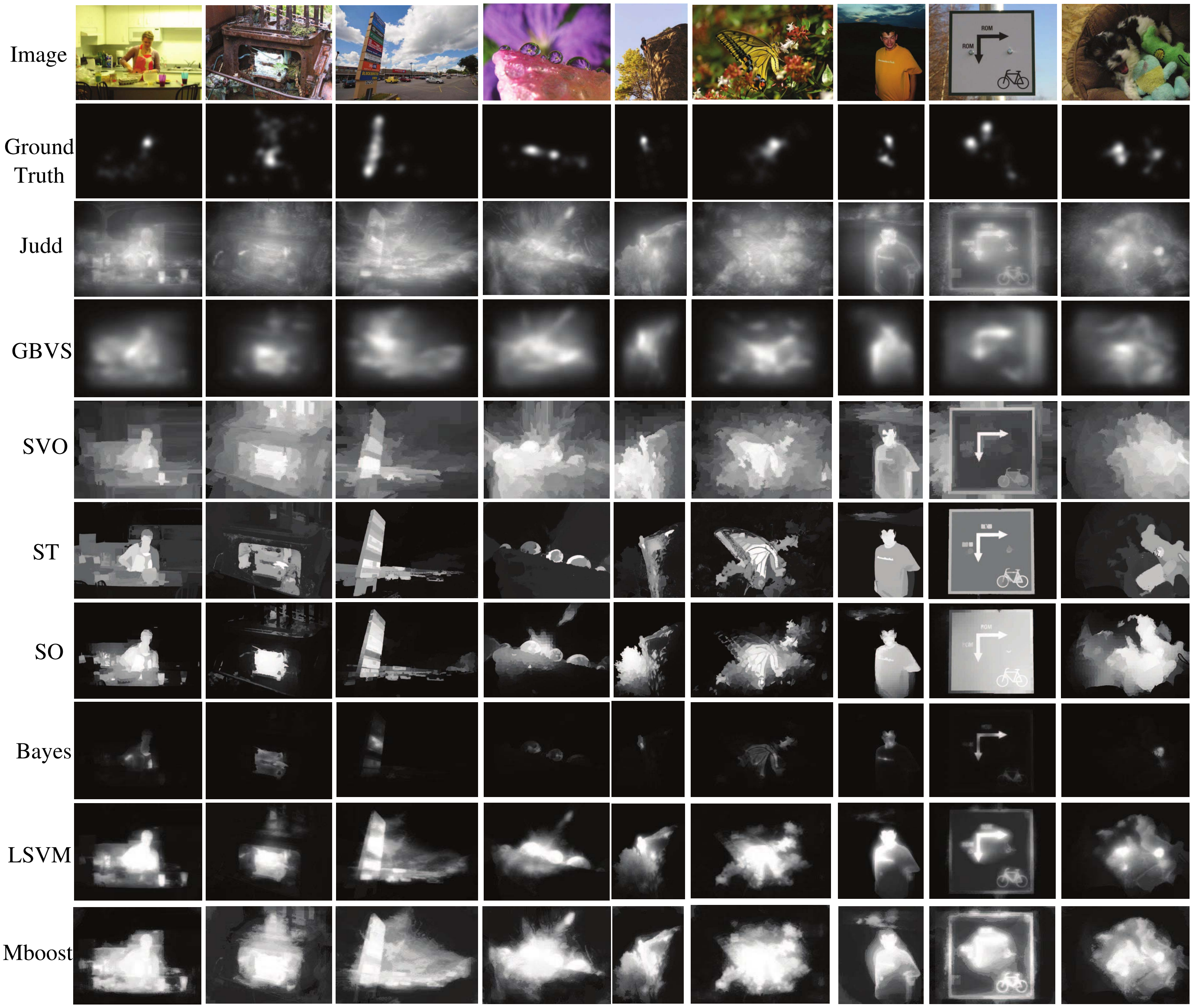

Fig. 7. Examples of saliency detection results of single models are shown in row 3 through 7 (image example from the MIT dataset [28]) and Imgsal [68]). Saliency maps generated by our proposed fusion strategies are given in row 8 through 10. (groundtruth denotes ground-truth eye fixation map).

fusing strategies. The reason is that both Bayes and Sum fusion rules assume independence among individual models, but infact independence assumption among models is hard to achieve as many saliency models share similar feature sets. Hence, a prediction error caused by certain features can be collapsed over several individual models and further corrupt final prediction results. Thus, a general integration strategy without independence assumption, such as density, outperforms Bayes and Sum.

Note that not all fusion approaches outperform the individual models. For example, Min and Max fusion strategies perform lower than several individual models. This is mainly because these fusion strategies introduce a bias regarding different model outputs. Specifically, Max fusion has strong bias on the model with a maximum confidence score while neglecting scores from other models. Moreover, the density fusing approach requires sufficient training data to obtain an accurate or reasonable density estimation function, hence its performance is not as stable as other learning based fusing approaches.

The fact that the learning based fusing schema outperforms others indicates the value of using human eye fixation data in guiding saliency map prediction. It is well known that saliency not only is related to objective features such as color and shape, but it is also influenced by different contexts [33]. To illustrate this, we show an example in Fig. 10 where the two images in the first row show the same clown in pink pants. The second row shows fixation heat maps as well as maps from several state-of-the-art models such as Judd, GBVS, and Saliency Tree algorithm. Our predicted saliency maps using the learning method are shown in the last row. From the second row, it is clear that, according to the context change, the same clown's saliency degree has increased. This kind of "context influence on object's saliency" can not be captured without learning from human fixation data. Furthermore, saliency detectors, such as GBVS, and Saliency Tree do not perform

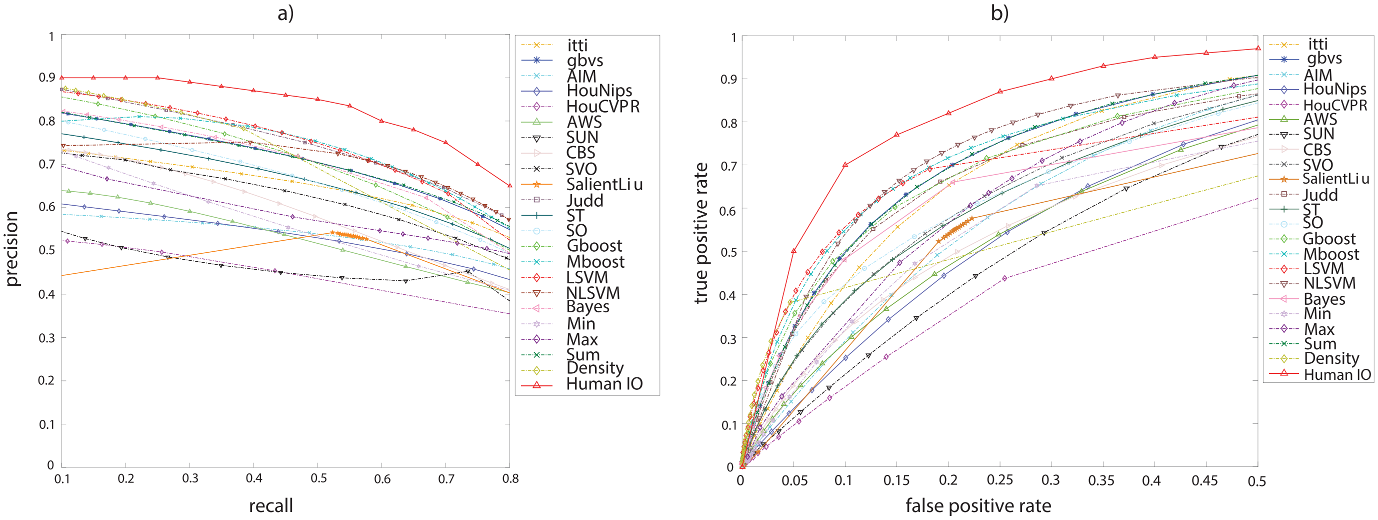

Fig. 8. Average precision-recall and ROC curves of all saliency fusion strategies (results collapsed on both the MIT and imgsal datasets). For better viewing purpose, here we show recall [0.1, 0.8] in (a), and false positive rate [0, 0.5] in (b). a) Precision-recall. b) ROC.

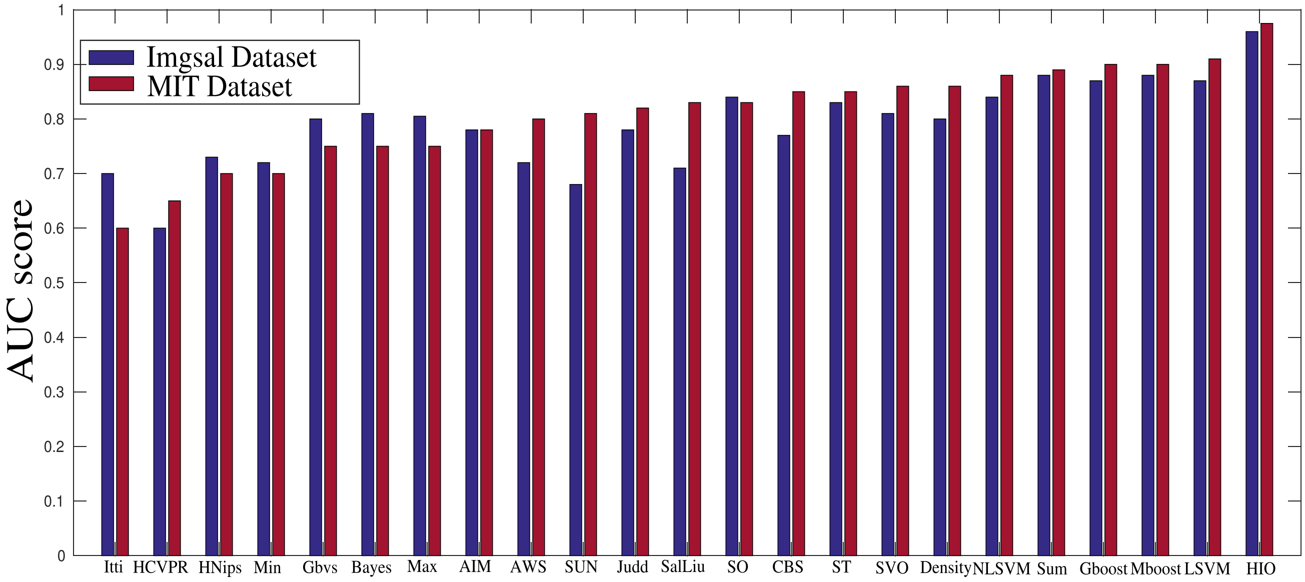

Fig. 9. Average AUC score on the MIT and Imgsal datasets.

very well if only given "low-level" features, such as color, location, gradient histogram. This is because similar objects produce highly correlated "low-level" features, which could only confuse a classifier in a different "context" scenario. On the other hand, "high-level" features, such as in our individual prediction models, are especially good at representing this "context" relationship as these models are designed to describe a local area with its surroundings. Hence, our fusion based strategies excel in capturing the "context" aspects. In Fig. 10, last row, it is clear that our learning based para-boosting model successfully captures the context influence, and adjusts the saliency degree of the clown in pink pants accordingly. This example provides the basic motivation behind our para-boosting model, and explains its good performance.

1) Model Choice and Comparison: A natural question regarding our late fusing schema is that how many individual models are needed to achieve good performance? Is it true that the more, the better? To explore the answer, we push our proposed para-boosting schema to its limit, that is, reducing as many individual models as possible while keeping similar performance. It turns out that decreasing the number of models to 3 does not affect learning based para-boost strategy very much while non-learning based schema are affected drastically. We thus choose the top 3 performing models among our 13 testing models: GBVS [18], SVO [30], and Judd [28].

The PR and ROC curves as performance comparison between fusing all models and the top 3 individual models are shown in Fig. 11. Generally speaking, non-learning based

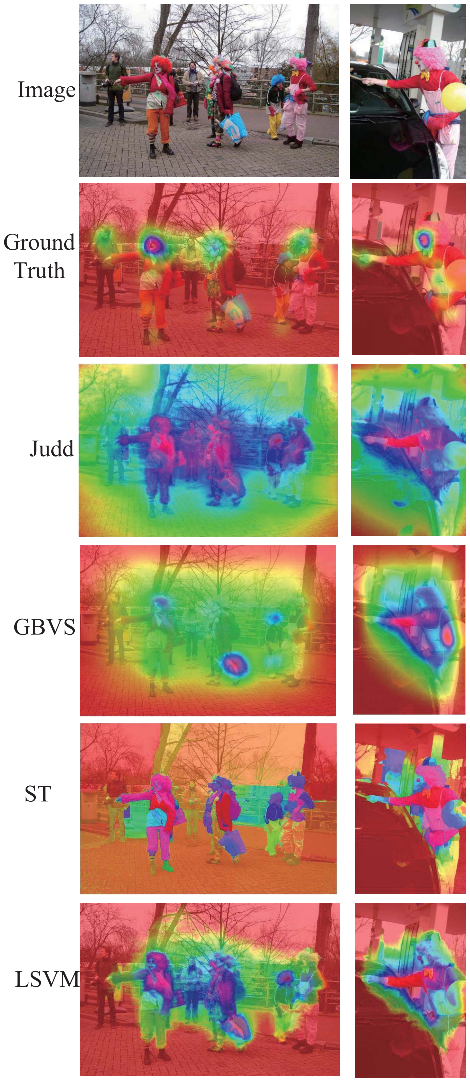

Fig. 10. Illustration of image context's influence on objects (image example from the MIT database). The clown in pink pants has different saliency degree as image context changes from left to right. first row: Original images, second row: Ground truth saliency map overlaid with original image, third, fourth and fifth row: Judd, GBVS and Saliency Tree saliency map overlaid on the original image, last row: Our learning based para-boost schema (linearSVM) saliency map overlaid on the original image. This result indicates that our proposed schema, using fixation data for learning, captures the influence of changing context on the same object.

fusing methods are less consistent than learning based methods. Non-learning based methods such as Bayes, Min, Max and Sum, using the top 3 models outperform fusing all models since selecting the top 3 models helps exclude the dominant influence brought in by relatively worse performed models. Unlike non-learning based fusing methods, learning based fusion techniques such as Gboost, Mboost, LSVM, NLSVM and Density demonstrate similar performance when fusing the top 3 models and all models as the learning procedure can block inferior individual models' influence. This comparison indicates that a dimensionality reduction operation is possible while retaining the performance.

Furthermore, the mean average precision (mAP) and AUC of 13 and 3 models are reported in Table I over the MIT dataset [28]. For non-learning based approaches, significant AUC improvement (from 0.7229 to 0.8759 for *Min rule*, from 0.8192 to 0.8664 for *Max rule*, and from 0.8251 to 0.8865 for *Naive Bayesian*) is observed, while learning based methods show no obvious difference. Besides, Linear SVM is more stable than other fusing methods in terms of mAP and AUC variance (LSVM-13: mAP as 78.57% and AUC as 0.8910; LSVM-3: mAP as 78.71% and AUC as 0.8728) and Modest Adaboost (MBoost-13: mAP as 65.27% and AUC as 0.8837; MBoost-3: mAP as 66.85% and AUC as 0.8860).

TABLE I

SALIENCY PREDICTION RESULTS OF FUSION METHODS USING ALL 13 (1ST ROWS) AND THE TOP 3 MODELS (2ND ROWS) USING mAP AND AUC SCORES ON THE MIT DATASET

Fusion method	Model Number	mAP	AUC
Gboost	13	0.6441	0.8856
	3	0.6326	0.8890
Mboost	13	0.6527	0.8837
	3	0.6685	0.8860
LSVM	13	0.7857	0.8910
	3	0.7871	0.8728
NLSVM	13	0.6272	0.8783
	3	0.6167	0.8533
Naive Bayesian	13	0.8129	0.8251
	3	0.6983	0.8865
Min rule	13	0.7269	0.7229
	3	0.6817	0.8759
Max rule	13	0.4810	0.8192
	3	0.5340	0.8664
Sum rule	13	0.6430	0.8651
	3	0.6231	0.8837
Density	13	0.8650	0.8428
	3	0.7962	0.8789

For more direct illustration, examples of fusing saliency map the top 3 and 13 models are shown in Fig.6. The individual model results are shown in Fig. 12 for comparison purposes. The saliency map results further echo our previous analysis that learning-based fusing approaches, such as linear SVM (LSVM), non-linear SVM (NLSVM), Gentle adaboost (Gboost) and Modest adaboost (Mboost) are more stable in saliency map generation while non-learning based fusing approaches show great variance.

It is worthwhile to point out that the performance of learning strategy remaining the same as the reduced set of selected models conveys important information. Fusion of our manually selected the top 3 models is supposed to be good since it discards the influence of weak models. Thus, the similar performance indicates that our score-level fusion model has the capability of distinguishing "good" and "bad" models through learning.

2) Model Consistency on Image Samples: To further investigate fusion models' strengths and weaknesses, in Fig. 13, we illustrate the most and the least consistent images for which our fusion models agree with Human IO. It can be seen that our fusion models work well when there is a clear salient object, and fall short at images without well-defined visual attention spot. This is reasonable as different models may produce different maps for the last case, hence fusing them may not achieve a more focused saliency region in the final saliency map.

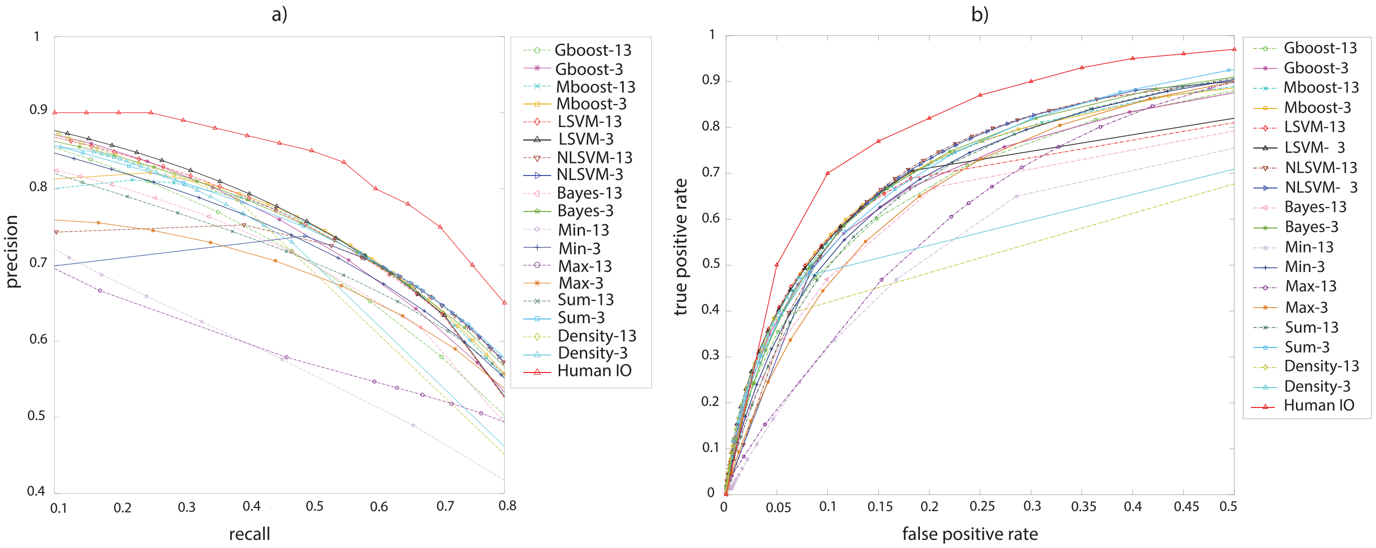

Fig. 11. Performance comparison of different choice of model selection (results collapsed on both the MIT and imgsal dataset). For better viewing purpose, here we show recall [0.1, 0.8] in (a), and false positive rate [0, 0.5] in (b). a) Precision-recall. b) ROC.

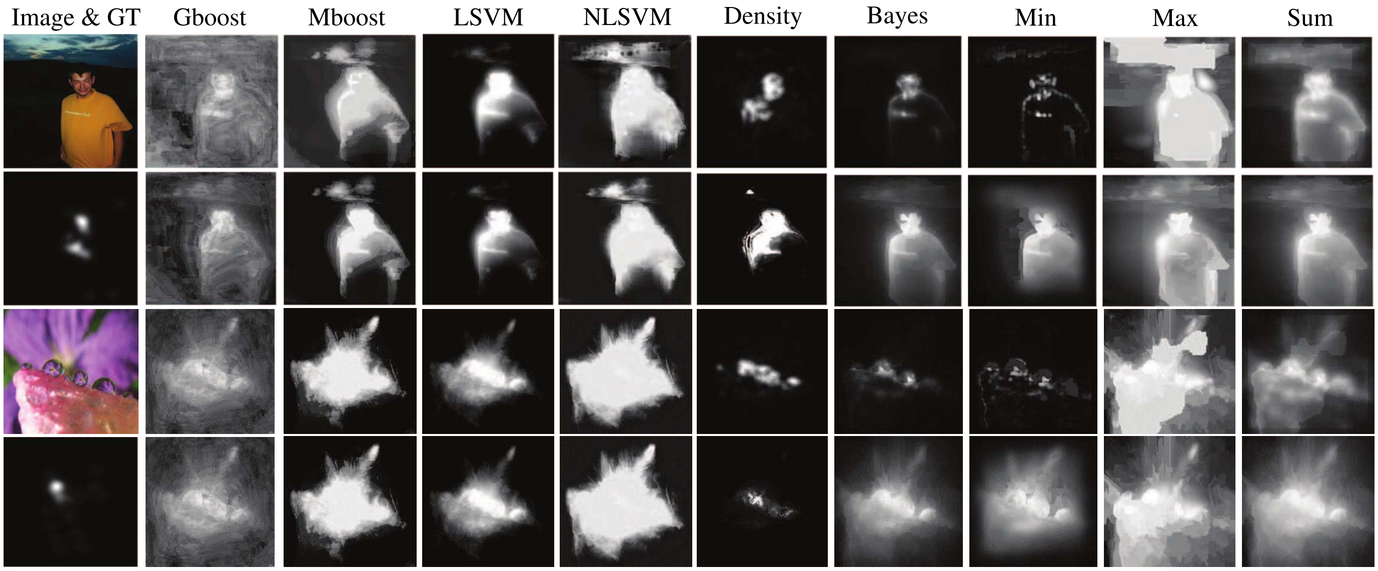

Fig. 12. Saliency map results of individual models (image example from the MIT dataset [28]).

TABLE II

AUC RANKING OF EXAMPLE IMAGES IN FIG. 7 (LSVM: LINEAR SVM, Mboost: MODEST AdaBoost)

Image number (top to bottom)	Img 1	Img 2	Img 3	Img 4	Img 5	Img 6	Img 7	Img 8	Img 9
Rank 1	LSVM	SO	LSVM	Bayes	Bayes	Bayes	LSVM	Bayes	LSVM
Rank 2	MBoost	Bayes	SO	ST	LSVM	LSVM	MBoost	LSVM	Judd
Rank 3	Bayes	LSVM	Bayes	LSVM	GBVS	MBoost	Bayes	MBoost	MBoost
Rank 4	SO	ST	SVO	SO	MBoost	GBVS	GBVS	SVO	GBVS
Rank 5	ST	MBoost	MBoost	MBoost	ST	SVO	Judd	GBVS	Bayes
Rank 6	Judd	Judd	ST	SVO	SVO	SO	SVO	Judd	SVO

Besides, sample saliency maps from top 5 individual models and fusion results of the top 3 methods are shown in Fig. 7, and the corresponding model rankings are shown in Table II. These sample images cover from street scene and people to simple patterns image. It can be seen that our proposed approach shows superiority over other models for different image types. Saliency maps from fusing models are more concentrated and focused comparing to results of each individual model. According to AUC and ROC curves reported for each individual image, the fusing results are closer to the ground truth, which indicates that our score level fusion model reduces the gap between the proposed model and Human IO model.

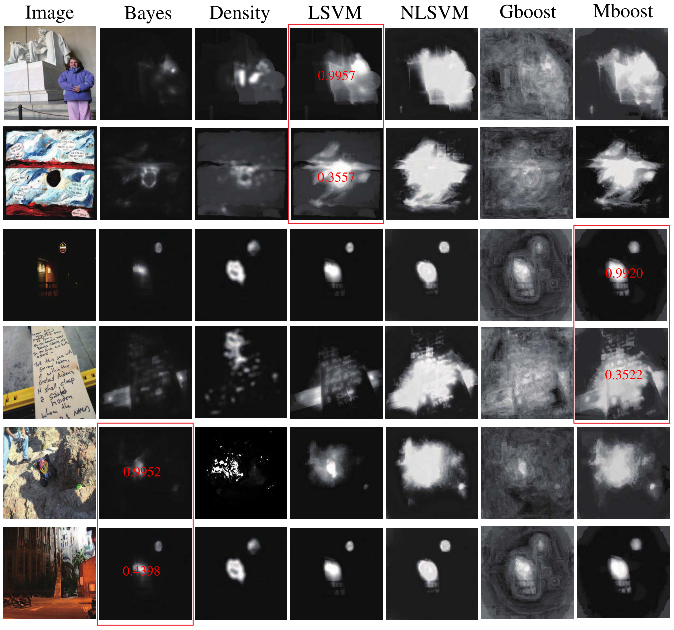

Fig. 13. Highest AUC score and lowest AUC score images for three fusion methods using all 13 models in the MIT database. Generally, least consistent images (lowest AUC score) are more cluttered.

IV. Conclusions and Future Work

In this work, we proposed different score-level fusion strategies including transformation based, classification based, and probability density based fusion schemes for combining state-of-the-art saliency prediction models. Experimental results indicate that our score-level fusion strategies outperform state-of-the-art individual models and score closeset to the performance of the human inter-observer model. Furthermore, through extensive comparison, we showed that our proposed fusion techniques show good performance over a broad range of images, and enriched the applicability range by fusing different individual saliency models. Note that our proposed fusion techniques are general and as more discovered predictive models can be combined to construct a strong model.

We also discussed the performance of fusing top individual saliency models and the most comprehensive set of fusion techniques for saliency prediction. It is proven that our proposed fusion strategy excels the state-of-the-art work as it intelliegntly combines advantages of different individual saliency prediction models. Especially, classifiers which are designed for better generalization and less over-fitting purposes, such as modest adaboost, show their power over other learning models. For future work, we aim to explore the possibility of adaptive selection of individual models to achieve better performance. Furthermore, given one fusion strategy, it would be valuable to analyze its weakness and how its performance can be enhanced by adding new individual saliency models.

Acknowledgement

The authors affirm that the views expressed herein are solely their own, and do not represent the views of the United States government or any agency thereof.

References

[1] A. Borji and L. Itti, “State-of-the-art in visual attention modeling,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 185–207, Jan. 2013.

[2] A. Borji, D. N. Sihite, and L. Itti, “Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study,” *IEEE Trans. Image Process.*, vol. 22, no. 1, pp. 55–69, Jan. 2013.

[3] A. Borji, H. R. Tavakoli, D. N. Sihite, and L. Itti, “Analysis of scores, datasets, and models in visual saliency prediction,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2013, pp. 921–928.

[4] O. Le Meur and Z. Liu, “Saccadic model of eye movements for free-viewing condition,” *Vis. Res.*, vol. 116, pp. 152–164, Nov. 2015.

[5] Z. Liu, R. Shi, L. Shen, Y. Xue, K. N. Ngan, and Z. Zhang, “Unsupervised salient object segmentation based on kernel density estimation and two-phase graph cut,” *IEEE Trans. Multimedia*, vol. 14, no. 4, pp. 1275–1289, Aug. 2012.

[6] R. Shi, Z. Liu, H. Du, X. Zhang, and L. Shen, “Region diversity maximization for salient object detection,” *IEEE Signal Process. Lett.*, vol. 19, no. 4, pp. 215–218, Apr. 2012.

[7] C. Guo and L. Zhang, “A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression,” *IEEE Trans. Image Process.*, vol. 19, no. 1, pp. 185–198, Jan. 2009.

[8] L. Shen, Z. Liu, and Z. Zhang, “A novel H.264 rate control algorithm with consideration of visual attention,” *J. Multimedia Tools Appl.*, vol. 63, no. 3, pp. 709–727, 2013.

[9] D. Ćulibrk, M. Mirkovic, V. Zlokolica, M. Pokric, V. Crnojevic, and D. Kukolj, “Salient motion features for video quality assessment,” *IEEE Trans. Image Process.*, vol. 20, no. 4, pp. 948–958, Apr. 2010.

[10] H. Du, Z. Liu, J. Jiang, and L. Shen, “Stretchability-aware block scaling for image retargeting,” *J. Vis. Commun. Image Represent.*, vol. 24, no. 4, pp. 499–508, 2013.

[11] Z. Liu, X. Zhang, S. Luo, and O. Le Meur, “Superpixel-based spatiotemporal saliency detection,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 9, pp. 1522–1540, Sep. 2014.

[12] Y. Luo, J. Yuan, P. Xue, and Q. Tian, “Saliency density maximization for efficient visual objects discovery,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 12, pp. 1822–1834, Dec. 2011.

[13] A. Torralba and A. A. Efros, “Unbiased look at dataset bias,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2011, pp. 1521–1528.

[14] A. Borji, D. N. Sihite, and L. Itti, “Salient object detection: A benchmark,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2012, pp. 414–429.

[15] L. Itti, C. Koch, and E. Niebur, “A model of saliency-based visual attention for rapid scene analysis,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[16] C. Koch and S. Ullman, “Shifts in selective visual attention: Towards the underlying neural circuitry,” *Human Neurobiol.*, vol. 4, no. 4, pp. 219–227, 1985.

[17] A. M. Treisman and G. Gelade, “A feature-integration theory of attention,” *Cognit. Psychol.*, vol. 12, no. 1, pp. 97–136, 1980.

[18] J. Harel, C. Koch, and P. Perona, “Graph-based visual saliency,” in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2006, pp. 542–552.

[19] T. Liu *et al.*, “Learning to detect a salient object,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 2, pp. 353–367, Feb. 2011.

[20] N. D. B. Bruce and J. K. Tsotsos, “Saliency based on information maximization,” in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2005, pp. 155–162.

[21] X. Hou and L. Zhang, “Dynamic attention: Searching for coding length increments,” in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2008, pp. 681–688.

[22] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell, “SUN: A Bayesian framework for saliency using natural statistics,” *J. Vis.*, vol. 8, no. 7, pp. 1–20, 2008.

[23] L. Kong, L. Duan, W. Yang, and Y. Dou, “Salient region detection: An integration approach based on image pyramid and region property,” *IET Comput. Vis.*, vol. 9, no. 1, pp. 85–97, 2015.

[24] Y. Fang, W. Lin, Z. Chen, C.-M. Tsai, and C.-W. Lin, “A video saliency detection model in compressed domain,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 1, pp. 27–38, Jan. 2014.

[25] H. Jiang, J. Wang, Z. Yuan, T. Liu, N. Zheng, and S. Li, “Automatic salient object segmentation based on context and shape prior,” in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2011, pp. 1–12.

[26] A. Borji and L. Itti, “Exploiting local and global patch rarities for saliency detection,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2012, pp. 478–485.

[27] M. Cerf, E. P. Frady, and C. Koch, "Faces and text attract gaze independent of the task: Experimental data and computer model," *J. Vis.*, vol. 9, no. 12, pp. 1–15, 2009.

[28] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *Proc. IEEE 12th Int. Conf. Comput. Vis. (ICCV)*, Sep./Oct. 2009, pp. 795–825.

[29] A. Nuthmann and J. M. Henderson, "Object-based attentional selection in scene viewing," *J. Vis.*, vol. 10, no. 8, pp. 2237–2242, 2010.

[30] K.-Y. Chang, T.-L. Liu, H.-T. Chen, and S.-H. Lai, "Fusing generic objectness and visual saliency for salient object detection," in *Proc. Int. Conf. Comput. Vis. (ICCV)*, Nov. 2011, pp. 914–921.

[31] B. Alexe, T. Deselaers, and V. Ferrari, "What is an object?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2010, pp. 73–80.

[32] G. Kootstra, A. Nederveen, and B. de Boer, "Paying attention to symmetry," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2008, pp. 1115–1125.

[33] M. Pajak and A. Nuthmann, "Object-based saccadic selection during scene perception: Evidence from viewing position effects," *J. Vis.*, vol. 13, no. 5, pp. 1–21, 2013.

[34] B. W. Tatler, "The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions," *J. Vis.*, vol. 7, no. 14, pp. 1–17, 2007.

[35] K. A. Ehinger, B. Hidalgo-Sotelo, A. Torralba, and A. Oliva, "Modelling search for people in 900 scenes: A combined source model of eye guidance," *Vis. Cognit.*, vol. 17, nos. 6–7, pp. 945–978, 2009.

[36] Z. Li, "A saliency map in primary visual cortex," *Trends Cognit. Sci.*, vol. 6, no. 1, pp. 9–16, Jan. 2002.

[37] L. Itti, and C. Koch, "Feature combination strategies for saliency-based visual attention systems," *J. Electron. Imag.*, vol. 10, no. 1, pp. 161–169, 2001.

[38] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, 2004.

[39] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.

[40] Q. Zhao and C. Koch, "Learning a saliency map using fixated locations in natural scenes," *J. Vis.*, vol. 11, no. 3, pp. 1–15, 2011.

[41] A. Borji, "Boosting bottom-up and top-down visual features for saliency estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2012, pp. 438–445.

[42] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *J. Comput. Syst. Sci.*, vol. 55, no. 1, pp. 119–139, Aug. 1997.

[43] P. Reinagel and A. M. Zador, "Natural scene statistics at the centre of gaze," *Netw., Comput. Neural Syst.*, vol. 10, no. 4, pp. 341–350, 1999.

[44] W. Kienzle, A. F. Wichmann, B. Schölkopf, and M. O. Franz, "A nonparametric approach to bottom-up visual saliency," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2006, pp. 689–696.

[45] L. Zhang, Y. Xia, R. Ji, and X. Li, "Spatial-aware object-level saliency prediction by learning graphlet hierarchies," *IEEE Trans. Ind. Electron.*, vol. 62, no. 2, pp. 1301–1308, Feb. 2015.

[46] L. Mai, Y. Niu, and F. Liu, "Saliency aggregation: A data-driven approach," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2013, pp. 1131–1138.

[47] O. Le Meur and Z. Liu, "Saliency aggregation: Does unity make strength?" in *Proc. 12th Asia Conf. Comput. Vis. (ACCV)*, 2014, pp. 18–32.

[48] Y. Xie, H. Lu, and M.-H. Yang, "Bayesian saliency via low and mid level cues," *IEEE Trans. Image Process.*, vol. 22, no. 5, pp. 1689–1698, May 2013.

[49] X. Li, H. Lu, L. Zhang, X. Ruan, and M.-H. Yang, "Saliency detection via dense and sparse reconstruction," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2013, pp. 2976–2983.

[50] E. Rahtu, J. Kannala, M. Salo, and J. Heikkilä, "Segmenting salient objects from images and videos," in *Proc. 11th Eur. Conf. Comput. Vis. (ECCV)*, 2010, pp. 366–379.

[51] V. Navalpakkam and L. Itti, "An integrated model of top-down and bottom-up attention for optimizing detection speed," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2006, pp. 2049–2056.

[52] J. M. Wolfe, E. F. Risko, and A. Kingstone, *Visual Search*. Hove, U.K: Psychology Press, 1998, pp. 13–71.

[53] A. Borji, M. N. Ahmadabadi, and B. N. Araabi, "Cost-sensitive learning of top-down modulation for attentional control," *Mach. Vis. Appl.*, vol. 22, no. 1, pp. 61–76, 2011.

[54] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2007, pp. 1–8.

[55] A. Jain, K. Nandakumar, and A. Ross, "Score normalization in multimodal biometric systems," *Pattern Recognit.*, vol. 38, no. 12, pp. 2270–2285, Dec. 2005.

[56] R. Snelick, U. Uludag, A. Mink, M. Indovina, and A. K. Jain, "Large-scale evaluation of multimodal biometric authentication using state-of-the-art systems," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 3, pp. 450–455, Mar. 2005.

[57] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, "On combining classifiers," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 3, pp. 226–239, Mar. 1998.

[58] A. Hagen, "Robust speech recognition based on multi-stream processing," Ph.D. dissertation, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, 2001.

[59] J. Ming and F. J. Smith, "Speech recognition with unknown partial feature corruption—A review of the union model," *Comput. Speech Lang.*, vol. 17, nos. 2–3, pp. 287–305, 2003.

[60] T. Chen and R. R. Rao, "Audio-visual integration in multimodal communication," *Proc. IEEE*, vol. 86, no. 5, pp. 837–852, May 1998.

[61] N. Poh and S. Bengio, "EER of fixed and trainable fusion classifiers: A theoretical study with application to biometric authentication tasks," *Multiple Classier Syst.*, vol. 17, no. 5, pp. 74–85, 2005.

[62] K. Nandakumar, Y. Chen, S. C. Dass, and A. K. Jain, "Likelihood ratio-based biometric score fusion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 2, pp. 342–347, Feb. 2008.

[63] A. GarciaDiaz, X. R. FdezVidal, X. M. Pardo, and R. Dosil, "Decorrelation and distinctiveness provide with humanlike saliency," in *Advanced Concepts for Intelligent Vision Systems (ACIVS)*, vol. 5807. Berlin, Germany: Springer, 2009, pp. 343–354.

[64] W. Zhu, S. Liang, Y. Wei, and J. Sun, "Saliency optimization from robust background detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 2814–2821.

[65] Z. Liu, W. Zou, and O. Le Meur, "Saliency tree: A novel saliency detection framework," *IEEE Trans. Image Process.*, vol. 23, no. 5, pp. 1937–1952, May 2014.

[66] S. Prabhakar and A. K. Jain, "Decision-level fusion in fingerprint verification," *Pattern Recognit.*, vol. 35, no. 4, pp. 861–874, 2002.

[67] R. P. W. Duin and D. M. J. Tax, "Experiments with classifier combining rules," in *Multiple Classifier Systems* (Lecture Notes in Computer Science), vol. 1857. 2000, pp. 16–29.

[68] J. Li, M. D. Levine, X. An, and H. He, "Saliency detection based on frequency and spatial domain analyses," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2011, pp. 1–11.

[69] F. Wang and J. Han, "Multimodal biometric authentication based on score level fusion using support vector machine," *Opto-Electron. Rev.*, vol. 17, no. 1, pp. 59–64, 2009.

[70] P. Mehrani and O. Veksler, "Saliency segmentation based on learning and graph cut refinement," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, vol. 101. 2010, pp. 1–12.

[71] M.-M. Cheng, G.-X. Zhang, N. J. Mitra, X. Huang, and S.-M. Hu, "Global contrast based salient region detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2011, pp. 409–416.

[72] H. J. Seo and P. Milanfar, "Static and space-time visual saliency detection by self-resemblance," *J. Vis.*, vol. 9, no. 12, pp. 1–27, 2009.

[73] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2010, pp. 2376–2383.

[74] Y. Zhai and M. Shah, "Visual attention detection in video sequences using spatiotemporal cues," in *Proc. ACM Multimedia*, 2006, pp. 815–824.

[75] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2009, pp. 1597–1604.

[76] D. Gao and N. Vasconcelos, "Discriminant saliency for visual recognition from cluttered scenes," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2005, pp. 481–488.

[77] Y.-F. Ma and H.-J. Zhang, "Contrast-based image attention analysis by using fuzzy growing," in *Proc. ACM Multimedia*, 2003, pp. 374–381.

[78] V. Gopalakrishnan, Y. Hu, and D. Rajan, "Random walks on graphs to model saliency in images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2009, pp. 1698–1705.

[79] V. Movahedi and J. H. Elder, "Design and perceptual validation of performance measures for salient object segmentation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2010, pp. 49–56.

[80] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "LabelMe: A database and web-based tool for image annotation," *Int. J. Comput. Vis.*, vol. 77, nos. 1–3, pp. 157–173, 2008.

[81] L. Itti, "Automatic foveation for video compression using a neurobiological model of visual attention," *IEEE Trans. Image Process.*, vol. 20, no. 11, pp. 1304–1318, Oct. 2004.

[82] W. Einhauser, M. Spain, and P. Perona, "Objects predict fixations better than early saliency," *J. Vis.*, vol. 8, no. 14, pp. 1–26, 2008.

[83] C. Kanan and G. Cottrell, "Robust classification of objects, faces, and flowers using natural image statistics," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2010, pp. 2472–2479.

[84] Y.-F. Ma, X.-S. Hua, L. Lu, and H.-J. Zhang, "A generic framework of user attention model and its application in video summarization," *IEEE Trans. Multimedia*, vol. 7, no. 5, pp. 907–919, Oct. 2005.

[85] L. Elazary and L. Itti, "Interesting objects are visually salient," *J. Vis.*, vol. 8, no. 3, pp. 1–15, 2008.

[86] S. Ramanathan, H. Katti, N. Sebe, M. Kankanhalli, and T.-S. Chua, "An eye fixation database for saliency detection in images," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2010, pp. 30–43.

[87] H. F. Chua, J. E. Boland, and R. E. Nisbett, "Cultural variation in eye movements during scene perception," *Proc. Nat. Acad. Sci. USA*, vol. 102, no. 35, pp. 12629–12633, 2005.

[88] R. Carmi and L. Itti, "Visual causes versus correlates of attentional selection in dynamic scenes," *Vis. Res.*, vol. 46, no. 26, pp. 4333–4345, 2006.

[89] J. Richiardi, "Probabilistic models for multi-classifier biometric authentication using quality measures," Ph.D. dissertation, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, 2007.

[90] J. P. Hube, "Methods for estimating biometric score level fusion," in *Proc. IEEE 4th Int. Conf. Biometrics, Theory, Appl., Syst.*, Sep. 2010, pp. 1–6.

[91] S. Frintrop, "VOCUS: A visual attention system for object detection and goal-directed search," Ph.D. dissertation, Univ. Bonn, Bonn, Germany, 2005.

[92] Q. Zhao and C. Koch, "Learning visual saliency by combining feature maps in a nonlinear manner using AdaBoost," *J. Vis.*, vol. 12, no. 6, pp. 1–15, 2012.

[93] Y. Li, K. Fu, Z. Liu, and J. Yang, "Efficient saliency-model-guided visual co-saliency detection," *IEEE Signal Process. Lett.*, vol. 22, no. 5, pp. 588–592, May 2014.

Jingwei Wang was born in China in 1984. She received the B.S. and M.S. degrees from Xidian University, Xi'an, China, in 2005 and 2008, respectively, and the Ph.D. degree from the University of Southern California, Los Angeles, CA, in 2013, all in electrical engineering. Her main areas of research interest lies in salient object detection, segmentation, and recognition for 2D/3D automatically conversion purpose.

Ali Borji (M'11) received the B.S. degree in computer engineering from the Petroleum University of Technology, Tehran, Iran, in 2001, the M.S. degree in computer engineering from Shiraz University, Shiraz, Iran, in 2004, and the Ph.D. degree in cognitive neurosciences from the Institute for Studies in Fundamental Sciences, Tehran, in 2009. He then spent a year with the University of Bonn, Germany, as a Research Assistant. He is currently a Post-Doctoral Scholar with iLab, University of Southern California, Los Angeles, CA. His research interests include bottom-up and top-down visual attention, visual search and active learning, object and scene recognition, machine learning, cognitive robotics, cognitive and computational neurosciences, and biologically plausible vision models.

C.-C. Jay Kuo received the B.S. degree from National Taiwan University, Taipei, in 1980, and the M.S. and Ph.D. degrees from the Massachusetts Institute of Technology, Cambridge, in 1985 and 1987, respectively, all in electrical engineering. From 1987 to 1988, he was a Computational and Applied Mathematics Research Assistant Professor with the Department of Mathematics, University of California at Los Angeles. Since 1989, he has been with the University of Southern California (USC). He is currently the Director of the Multimedia Communication Laboratory and a Professor of Electrical Engineering and Computer Science with USC.

Laurent Itti received the M.S. degree in image processing from the Ecole Nationale Superieure des Telecommunications, Paris, France, in 1994, and the Ph.D. degree in computation and neural systems from Caltech, Pasadena, CA, in 2000. He has been an Assistant, Associate, and a Full Professor of Computer Science, Psychology, and Neuroscience with the University of Southern California. His research interests are in biologically inspired computational vision, in particular in the domains of visual attention, scene understanding, control of eye movements, and surprise. This basic research has technological applications to, among others, video compression, target detection, and robotics. He has co-authored over 150 publications in peer-reviewed journals, books and conferences, three patents, and several open source neuromorphic vision software toolkits.