Learning Attention Map from Images

Yao Lu, Wei Zhang, Cheng Jin and Xiangyang Xue
School of Computer Science, Fudan University, Shanghai, China
{yaolu, weizh, jc, xyxue}@fudan.edu.cn

Abstract

While bottom-up and top-down processes have shown effectiveness during predicting attention and eye fixation maps on images, in this paper, inspired by the perceptual organization mechanism before attention selection, we propose to utilize figure-ground maps for the purpose. So as to take both pixel-wise and region-wise interactions into consideration when predicting label probabilities for each pixel, we develop a context-aware model based on multiple segmentation to obtain final results. The MIT attention dataset [14] is applied finally to evaluate both new features and model. Quantitative experiments demonstrate that figure-ground cues are valid in predicting attention selection, and our proposed model produces improvements over baseline method.

1. Introduction

Due to the massive visual information received from the nature world, humankind has evolved the ability of attending to what they are interested in using eye saccades so as to reduce the complexity of visual processing [20]. Aiming at the same target in the computer vision society, predicting human visual attention and eye fixations on images has long been an open question considering its broad applications in various tasks including object detection, scene understanding, image/video retrieval, advertisement and UI design, etc.

Previous studies have shown that visual attention follows two main procedures: bottom-up and top-down. Bottom-up attention indicates that selection of visual attention depends on low-level features (i.e. color, texture, brightness etc.) of images, and previous works on saliency detection have done well on this direction [13, 1, 30]. Meanwhile, top-down attention claims that object information dominates over the attention selection: human attend to familiar object entities rather than regions with salient low-level features [5, 3, 14].

Some researchers have previously combined these two mechanisms together and construct models to predict visual attention. Cerf et al [4] build a model to mix Itti's saliency and a face detector, which performs much better than low-level saliency alone. Judd et al [14] prove that low-level saliency does not account for attention and eye fixations using eye tracking devices on human subjects. To solve this problem, they further propose to detect cars and pedestrians as well as horizons in their scheme, which they think are of great interest by human observers. All these pioneer works show the feasibility of both bottom-up and top-down cues to predict attention maps on images.

However, recent advances in neuroscience and psychology have revealed some other important mechanisms before attention selection. Kimchi et al [15] show by cognition experiments that figure-ground organization occurs before attention selection. Qiu et al [21] also prove in brain experiments that figure-ground guides the attention selection. Hence, in this paper, we have every reason to assume that, *figural objects capture attention and eye fixations*, which is another important cue often ignored by many state-of-the-art algorithms.

Figure-ground and perceptual organization can be traced back to the 1920s. Many rules have been found by Gestalt psychologists including convexity, parallelism, symmetry, orientation, surroundedness and object familiarity etc [20]. By utilizing the bias towards figural objects of these local cues, several previous works attempt to assign figure/ground labels for regions and contours [8, 24, 17, 16]. Hence in this paper, we focus on the applying of Gestalt rules and exploit several local figure-ground cues, i.e. convexity, symmetry and surroundedness, to predict visual attention maps. Furthermore, basic bottom-up and top-down features are also taken into consideration so as to evaluate effectiveness of the proposed figure-ground features. When finally predicting probabilities for each pixel in the image, a context-aware model upon multiple segmentation is applied to further improve detection result of the baseline model, which uses an SVM to learn weights between different bottom-up and top-down features.

The rest part of the paper is organized as follows. In section 2 we introduce basic feature as well as three perceptual cues used in our approach. Section 3 describes our model in detail. Experiments and evaluations are introduced in Section 4, while Section 5 ends the paper with a conclusion and a discussion on future works.

2. Feature used to predict attention map

2.1. Basic features

Among the features used in our approach, the basic features are much the same as that used in the baseline method [14], so we briefly illustrate and review them here.

Low-level features. As an important cue to indicate bottom-up attention, low-level feature saliency has been researched well in the literature. Below is a list of the low-level features used in our model.

F1. Local energy of steerable pyramid filters [29].

F2. Itti and Koch's saliency map [13].

F3. Color (RGB values, probabilities and histograms).

F4. Torralba and Rosenholtz's saliency [25, 19].

High-level features. Judd et al [14] consider car, pedestrian and face as object entities that people will draw their eyes on, and this satisfies the top-down process of human attention selection. Hence these high-level features are used:

F5. Felzenswalb's car and pedestrian detector [7].

F6. Viola Jones's face detector [32].

Spatial feature is another important cue to show where the target should be and the spatial relationship between targets in images. We use:

F7. Distance to the center for each pixel.

F8. Horizon position of the image [19].

2.2. Figure-ground maps

In this paper, we put emphasis on obtaining pixel-level probability maps of foreground objects, i.e. the figure-ground maps. We choose three important Gestalt cues and attempt to draw the map for each of them respectively.

2.2.1 F9. Convexity map

Convexity has long been proved as a cue to separate figural and background objects. It is shown by psychologists in lots of experiments that, if there is a boarder line between two neighboring regions, then the region on the convex side tends to be foreground. Imagine the famous face-vase picture, which demonstrates a black vase in the middle and two white faces on its two sides. In fact the borderline between the black vase and white face has the same degree of convexity when seen from two sides, which causes the confusion of which part is foreground in this picture.

Several previous works have utilized such useful cue for different tasks. Fowlkes et al [8] perform ecological statistics on nature images and find convexity indeed has the ability to discriminate which part is foreground. Ren et al [24] further design a model to classify contours using convexity.

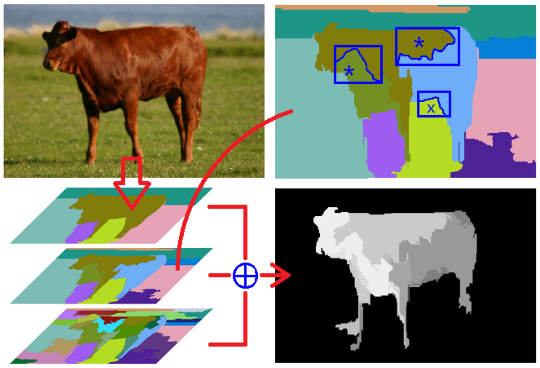

Figure 1. For a given image, multiple segmentation is calculated and convex regions are detected on each segmentation hypothesis. After merging all convex regions we obtain final convexity map (right-bottom). For example in the top-right image, curves in blue are detected concave arcs, and hence superpixels with "*" are convex regions detected. Notice superpixels with "x" do not meet the threshold th_a for the percentage of covered area by bounding-box of the concave arc.

Algorithm 1 Concave arc detection.

```
Input:
superpixel contour $s$ (clock-wise), concave threshold $th_c$
Steps:
Smooth $s$ using B-spline curve algorithm.
Draw bounding-box and split $s$ into four sections.
$ret = \phi$
for each section do
    Determine three main directions in sequence.
    Find starting points of concave arc when pixel movements violate main direction for more than $th_c$ steps.
    Find ending points of concave arc when pixel movements accord with main direction for more than $th_c$ steps after string points detected.
    $arc =$ combination of all the pieces in this section.
    $ret = ret \bigcup arc$.
end for
return $ret$
```

More recently, Lu et al [17] detect salient objects by constructing a multi-scale model and the cue of convexity is used to determine weights between superpixels. In this paper, in order to obtain pixel-level convexity maps for given images, we modify their method to meet our needs.

Generally, the main idea to construct convexity map describes as follows (see Figure 1 for the flow chart):

(1) Compute multiple segmentation for a given image. (2) For each segmentation hypothesis, find all concave arcs using Algorithm 1 and determine convex regions. (3) Map convex regions to pixels and add up results on different scales to obtain final convexity map.

The key problem in this scheme is to detect concave arcs,

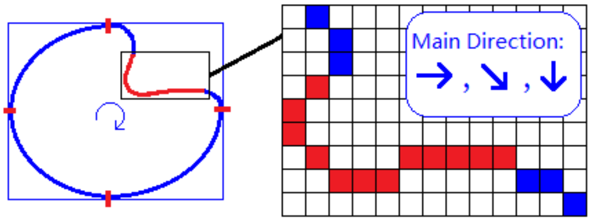

Figure 2. Example to detect concave arc. For the section of curve starting from the top to the right, main directions should be right, right-bottom and then bottom in sequence. Pixel movements of the blue part accord with the correct order of main directions, while pixel movements of the red part violate correct order so they are recognized as concave arc

which is to say, given the contour of a superpixel, how to find concave parts of the contour, meanwhile the part should be ”concave” enough [17]. Here we illustrates the algorithm in Algorithm 1 and Figure 2.

The usage of multiple segmentation summarizes convexity on different scales and benefits the final result, which is also applied in [12, 26]. Moreover, in our classification model introduced in Section 3, we will once again use the multiple segmentation result, hence no additional calculations will be made.

Once we get all the concave arcs in the multiple segmentation, we draw bounding-boxes for them, within which superpixels on the convex side are recognized as convex regions (percentage of covered area by the bounding-box should be greater than a threshold th_a, see Figure 1 for example). Finally, after simply adding all the convex regions detected in the multiple segmentation to the pixel level, we obtain the convexity map.

2.2.2 F10. Surroundedness map

Surroundedness is another Gestalt rule indicating that region surrounded by other region is likely to be foreground [20]. It can be seen as a special case of convexity – the whole object is convex rather than partly. Several previous works have used this cue to detect closed contours and perform interactive segmentation [31, 10].

Following the same opinion as the convexity map to use multiple segmentation, it is easy to construct a surroundedness map:

(1) Compute multiple segmentation for a given image. (2) For each segmentation, find superpixels that are *fully* surrounded by other superpixels. (3) Map the found superpixels to pixels and add up results on different segmentation hypotheses to obtain final surroundedness map.

Figure 3 demonstrates flow chart to obtain surroundedness map.

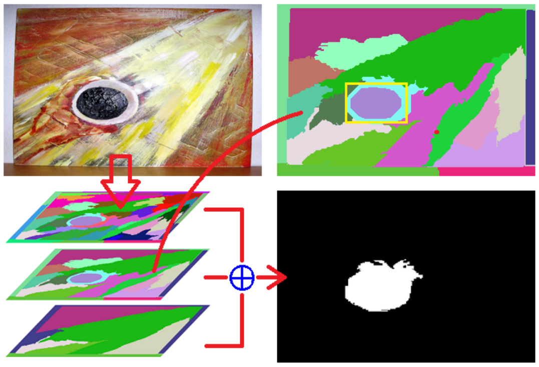

Figure 3. Flowchart to obtain surroundedness map, which is basically similar to the detection of convexity map. In the top-right image, rectangle in yellow is bounding-box of a fully surrounded superpixel.

2.2.3 F11. Symmetry map

Another important Gestalt cue to do figure-ground is symmetry. People pay more attention to high symmetric regions, and symmetry can also be used to predict center of objects [16]. Here we apply an isotropic symmetry operator described in [16, 23] to construct a symmetry map for a given image.

Given a point p, p_i and p_j are symmetric points according to p. The local symmetry of the pixel pair is defined as

$$c(i,j) = d(i,j,\sigma) \cdot p(i,j) \cdot m_i \cdot m_j \tag{1}$$

Where m_i is the magnitude of gradient at point p_i, $d(i,j,\sigma)$ represents the Gaussian weighting function on the distance between p_i and p_j with standard deviation σ, and

$$p(i,j) = (1 - cos(\gamma_i + \gamma_j)) \cdot (1 - cos(\gamma_i - \gamma_j)) \tag{2}$$

is the symmetry measurement where γ_i (γ_j) is the angle between the gradient direction of p_i (p_j) and the connection line between p_i and p_j (anti-clockwise direction, see Figure 4 for example). Hence the isotropic symmetry value of p is defined as

$$M^{iso}(x,y) = \sum_{(i,j)\in\Gamma(p)} c(i,j) \tag{3}$$

where $\Gamma(p)$ is all the pixel pairs within the radius r of p.

After the calculation of M^{iso} on each pixel in the image, we obtain the symmetry map on that scale, and practically, the symmetry detection is performed on several different scales and the merging technique is applied

$$S = \oplus_s N(M_s) \tag{4}$$

where $\oplus$ resizes different M to the same size, and 4 is chosen as the quantities of scales s, between pair of which there

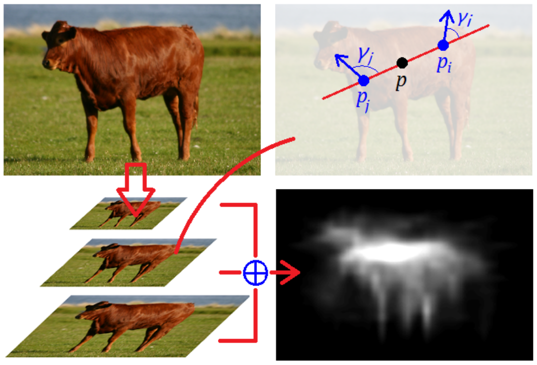

Figure 4. Flowchart to obtain symmetry maps. In the top-right image, p_i and p_j are symmetry points according to p.

is a down scaling by a factor of 2. N is the normalization operator used in [13], which normalizes the feature map to [0..1], and then multiply with $(1 - \bar{m})^2$. $\bar{m}$ is the average value of the feature map.

3. Learning a context-aware model

In the previous section, we have introduced different kinds of features used in our approach including bottom-up low-level feature cues, top-down object cues, as well as figure-ground perceptual cues. In this part, the classification model is introduced to predict final attention and eye fixation results.

In the baseline MIT attention model [14], 10 positive pixels and 10 negative pixels are chosen among the top 20% and bottom 70% salient locations for each image as training samples. Then a linear-kernel support vector machine is applied to predict label probabilities for testing images using the basic features (F1-F8), and the results show promising detection rates. However, their model do not consider the spatial relationship between pixels and regions.

Recently in the literature, the random walk segmentation algorithm [9] is influential and successfully model pixel-wise relations and perform interactive segmentation given starting points and labels. TextonBoost [28] also take relation between local pixels into consideration to perform semantic segmentation. However, these models are too local and do not consider regional interactions. Some other graph-based models such as mCRF [11] take care of regional impact, but it is hard to implement every pixel into the model due to the size limitation of the graph. Moreover, their approaches are relatively complicated and the computational cost is high.

Here we propose a simple approach to take both pixel-level and region-level contexts into consideration. To represent regional context, a multiple segmentation model is developed: for a given local pixel, we use the superpixel containing that pixel as context, and comparison of context is thus formulated as comparison of supuerpixels. Next, a random-walk scheme is applied to infer label probabilities for both superpixels and local pixels. The approach is illustrated in Figure 5 and detailed as follows:

Region-wise random walk. Suppose we have multiple segmentation hypothesis $S = \{S_1, S_2, \ldots, S_n\}$ for a given image. M^t is the superpixel set of S_t, and Δf_{mn} is feature distance between superpixel M^t_m and M^t_n. Given the initial label probability $P(M^t_m|f^t_m)$ for superpixel m in the hypothesis $S_t \in S$, we do: (1) move to a random superpixel n neighboring to m, calculate $P(M^t_n|M^t_m, f^t_n, \Delta f_{mn})$. (2) Continues to other superpixels using step 1 until all the superpixels are processed.

Pixel-wise random walk. After obtaining superpixel label probabilities for all the segmentation hypotheses, we have a label vector $\vec{X}_p = \{l_1, l_2, \ldots, l_n\}$ for each pixel p in the image, which implies label probabilities of different superpixels that contain this pixel in different segmentation hypotheses. Thus, another similar algorithm to obtain pixel label probabilities is used: given the initial label probability $P(N_p|f_p)$ for pixel p, we do (1) move to a random pixel q neighboring to p, calculate $P(N_q|N_p, f_q, \Delta f_{pq}, \Delta X_{pq})$. (2) Continues to other pixels using step 1, until all the pixels are processed. ΔX_{pq} depicts the contextual distance between pixel p and q

$$\begin{aligned} \Delta X_{pq} &= \vec{\alpha} \cdot (\vec{X}_p \ominus \vec{X}_q)^T \\ &= (\alpha_1, .., \alpha_n) \cdot (|x_{p1} - x_{q1}|, .., |x_{pn} - x_{qn}|)^T \end{aligned} \tag{5}$$

where x_{pn} and x_{qn} depict the nth component of $\vec{X}_p$ and $\vec{X}_q$. The addition of ΔX enables us to compare the context between pixels, which is hard to valuate in previous works. And it is especially useful at the intersection areas of objects. We import vector $\vec{\alpha} = \{\alpha_1, \ldots, \alpha_n\}$ here to represent weights of different context when comparing pixels. Intuitively the weights are correlated with various factors. For example, a pixel is highly related with its context in small area comparing with a large area of context, and under such circumstance the weigh for a small superpixel containing that pixel is high.

4. Experiments and evaluations

4.1. Dataset

With the development and help of eye tracking devices, researchers have gained access to real attention and eye fixation data these years. Several high quality eye tracking databases have been published in the literature recently, including the MIT dataset [14], the NUS dataset [22], the CIT FIFA dataset [3], as well as the Tsotsos dataset [2]. Among them the MIT dataset has a better variety of images classes and it is the closest to our natural world. We choose to evaluate our approach using this image database.

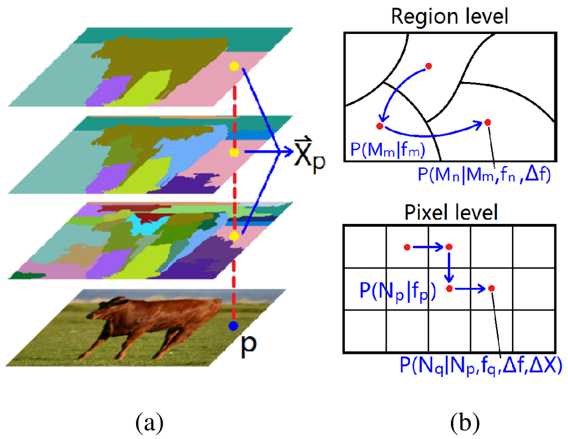

Figure 5. (a) We segment the image into multiple segmentation hypotheses, and label probabilities of superpixels in different hypotheses that contain the same pixel p form the contextual vector $\vec{X_p}$ for that pixel. (b) Given the label probability of a superpixel (resp. pixel), we take into account feature difference (resp. and contextual different) between two superpixels (resp. pixels) so as to predict label probability for a random neighboring superpixel (resp. pixel).

The MIT dataset contains 1003 images originally from Flickr and LabelMe [27] with a wide scope of different scenes and objects. The eye fixation data is then obtained from multiple users using eye tracking devices. Next a gaussian filter transforms discrete eye fixation points into continuous attention maps, and groundtruth can be attained by applying a threshold on the attention maps, which enables us to evaluate experimental results qualitatively and quantitatively.

4.2. Learning the model

Follow the baseline method in [14], we divide the dataset into 903 training images and 100 testing images. The images are resized to 200×200 pixels, and features F1 - F11 are calculated then. The training of the model can be separated into several parts:

Learning weights for distances There are two kinds of distance measurement applied in our approach: Δf for feature distance and ΔX for contextual distance.

For Δf, we use Joint Equal Contribution (JEC) proposed in [18] to scale each group of feature (F1-F11) to [0,1], and weight them the same because of the explicit physical meaning of each component.

For ΔX, the weight $\vec{\alpha}$ depicts importance of different contexts for a pixel. We randomly select 20 pairs of neighboring pixels that locate at intersection areas of superpixels for each training image (otherwise $\vec{X} = 0$ happens mostly regardless of the weight). Moreover, these pixels should be strongly positive or negative (top 20% or bottom 70% saliency). Thus the training set P of 9030 pairs of pixels is formed. We let $l = 0$ for negative pixels and $l = 1$ for positive pixels. The main target to learn the contextual weight is that, we maximize the contextual distance between pixels with different labels, while minimize that between pixels with same labels. The target function is below:

$$\operatorname*{argmin}_{\vec{\alpha}} \sum_{(a,b)\in P} (1-2|l_a-l_b|)\cdot\vec{\alpha}\cdot(\vec{X}_a \ominus \vec{X}_b)^T \qquad (6)$$

When pixel a and b have the same label, Eq.6 equals to minimizing $\vec{\alpha}\cdot(\vec{X}_a - \vec{X}_b)^T$, and when they have different labels, Eq.6 equals to maximizing the same item.

We employ a gradient decent algorithm to obtain the best $\vec{\alpha}$. The initial value of each component of the vector is set to be inversely proportional to the area of that superpixel. Because $\vec{\alpha}$ is low-dimensioned (chosen as 4-dim described later), the algorithm will converge quickly.

Learning initial guesses.

On the superpixel level, we first segment the images using Felzenswalb's segmentation algorithm [6] with different parameters, so that the images are separated into 10 - 30 pieces. n is chosen for 4, hence there are 4 hypotheses in the multiple segmentation model. Next, given the groundtruth attention map, we choose superpixels which contain more than 80% area of top 20% saliency pixels as positive samples (totally 8669 samples), while superpixels which contain more than 80% area of bottom 80% saliency pixels as negative samples (totally 16615 samples). The feature value for a superpixel is the mean value of all its containing pixels. Finally, we apply an liblinear support vector machine to train the model, where parameter c is selected as 1.

On the pixel level, 10 pixels of top 20% saliency are randomly chosen as positive samples for each training image (9030 totally), and 10 pixels of bottom 70% saliency as negative samples (9030 totally). No samples are chosen within 10 pixels near boundary of images as the baseline method does. A liblinear support vector machine is thus applied again to train the model, and parameter c is selected as 1.

Learning contextual interactions After initial guess of samples, we take contextual interactions into consideration using a Bayesian classifier, that is,$P(M_n|M_m, f_n, \Delta f_{mn})$ and $P(N_q|N_p, f_q, \Delta f_{pq}, \Delta X_{pq})$ respectively.

On regional level, given the feature difference between one superpixel and its preceding (neighboring) superpixel, we suppose the independency between variables and have

$$\begin{aligned} & P(M_n|M_m, f_n, \Delta f_{mn}) \\ \propto \quad & P(M_m, \Delta f_{mn}|M_n)P(M_n|f_n) \end{aligned} \qquad (7)$$

where the first term can be learned from training data by statistics, and the second term can be obtained from the initial guess stated above.

On pixel level, similarly we have

$$\begin{aligned} & P(N_q|N_p, f_q, \Delta f_{pq}, \Delta X_{pq}) \\ \propto \quad & P(N_p, \Delta f_{pq}, \Delta X_{pr}|N_q)P(N_q|f_q) \end{aligned} \tag{8}$$

Practically, we make statistics on the training sample set, which consists of all the inter-superpixel relationship in the multiple segmentation for Eq. 7 and 100 pairs of neighboring pixels per training image for Eq. 8.

4.3. Comparisons

To quantitatively evaluate both the new features and model, we (i) use the baseline model while adding new features; (ii) use our model while keeping original features; (iii) use both new model and features.

Figure 6 and 7 illustrate results of the three purpose. In Figure 6, we utilize the baseline SVM model while adding the figure-ground features we proposed in this paper. We threshold the groundtruth and detection result at $n = 5, 10, 15, 20, 25, 30$ percent saliency simultaneously to obtain binary attention map and comparison results. In Figure 7 we demonstrate the usage of all the features and new model at 30 percent saliency. Importances for different feature are also illustrated.

We make the following conclusions from our experiments and evaluations.

(1) We can see from Figure 7 that performances of symmetry and convexity are competitive to other bottom-up and top-down features, while combination of the three figure-ground cues performs almost the best among all the features. This show that the addition of figure-ground features is effective.

(2) Surroundedness performs not as good as other figure-ground features despite its explicit psychological meaning. It is still hard to obtain fully enclosed areas and objects simply by an unsupervised segmentation algorithm. Despite surroundedness performs better than random guess, this cue should be used when combining with other features.

(3) Our context-aware model upon multiple segmentation is also effective compared with the baseline SVM method. There is more or less performance enhancement for each single set and combination set of features.

(4) The true positive rate for the baseline approach is 0.87, and the addition of figure-ground features while using the baseline model gives 2% performance enhancement. The applying of our context-aware model while keeping basic features obtains 1.2% performance enhancement. And when we use both new model and features we gain 3.6% performance enhancement. Figure 8 demonstrates some good results of our proposed features and model.

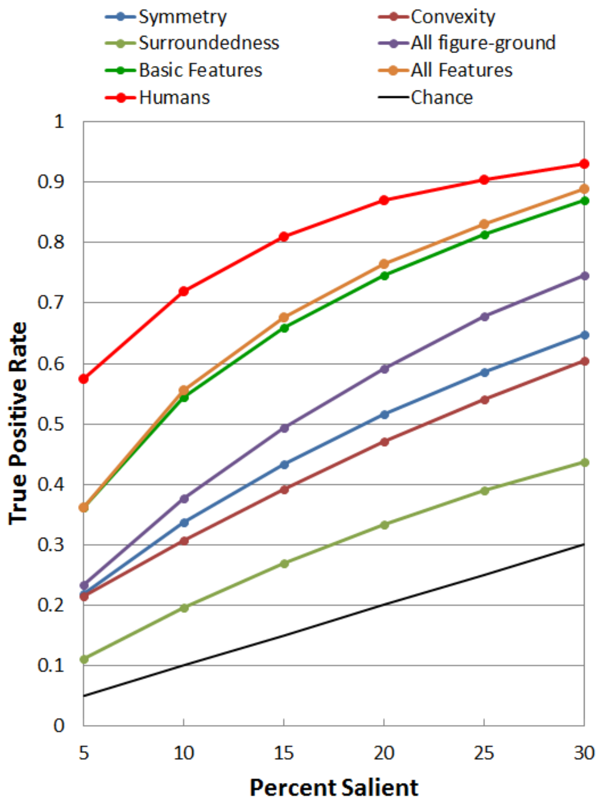

Figure 6. The ROC curve for the performance using baseline SVM model while adding new figure-ground cues. The addition of figure-ground cues provides 2% improvement.

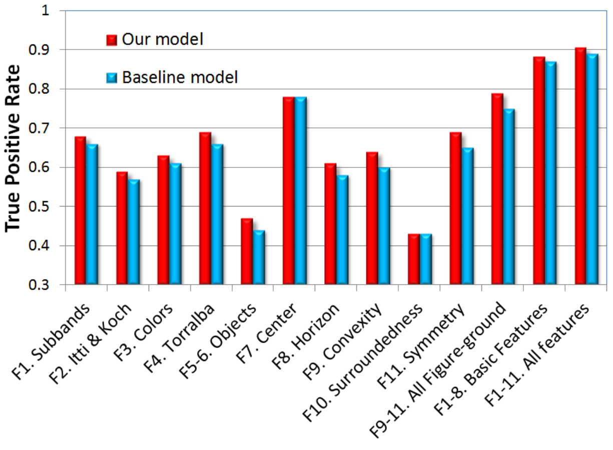

Figure 7. Comparison of true positive rates using both new model and new features under 30 percent salient, and importance of different features are also illustrated. Combination of the three figure-ground cues performs well compared with other top-down and bottom-up features. Moreover, when we us both new features and model, 3.6% performance enhancement is obtained compared with baseline approach.

5. Conclusion and discussion

In this work, by observation from experiments we conclude that figure-ground segmentation indeed contributes to the attention selection, which confirms recent conclusions in [21, 15] and shows a brand new way to predict attention and eye fixations using perceptual organization and figure-ground cues other than traditional bottom-up and top-down methods. Moreover, the proposed context-aware model further improves the baseline SVM model by considering both pixel-level and region-level interactions.

In our future work, we plan to study another kind of attention, i.e. controlled attention. It is really interesting and promising to combine different levels of perceptual cues into this task because recent studies have already shown that when people search for a certain object in their visual field, the mechanism of attention and eye saccades operates totally different, which is of the same target of object locating using sliding windows techniques in current computer vision society.

6. Acknowledgement

This work was partially funded by 973 Program (2010CB327906), Shanghai Leading Academic Discipline Project (B114), Doctoral Fund of Ministry of Education of China (20100071120033), and the NSF of China (No.60903077). We appreciate useful comments and suggestions from anonymous reviewers, as well as colleagues from institution of media computing (IMC) in Fudan University.

References

[1] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk. Frequency-tuned salient region detection. In *CVPR*, pages 1597–1604, 2009. 1

[2] N. Bruce and J. Tsotsos. Saliency, attention, and visual search: An information theoretic approach. *Journal of Vision*, 9(3), 2009. 4

[3] M. Cerf, E. Frady, and C. Koch. Faces and text attract gaze independent of the task: Experimental data and computer model. *Journal of vision*, 9(12), 2009. 1, 4

[4] M. Cerf, J. Harel, W. Einhäuser, and C. Koch. Predicting human gaze using low-level saliency combined with face detection. *NIPS*, 20:241–248, 2008. 1

[5] W. Einhäuser, M. Spain, and P. Perona. Objects predict fixations better than early saliency. *Journal of Vision*, 8(14), 2008. 1

[6] P. Felzenszwalb and D. Huttenlocher. Efficient graph-based image segmentation. *IJCV*, 59(2):167–181, 2004. 5

[7] P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *CVPR*, pages 1–8. 2

[8] C. Fowlkes, D. Martin, and J. Malik. Local figure–ground cues are valid for natural images. *Journal of Vision*, 7(8), 2007. 1, 2

[9] L. Grady. Random walks for image segmentation. *TPAMI*, pages 1768–1783, 2006. 4

[10] V. Gulshan, C. Rother, A. Criminisi, A. Blake, and A. Zisserman. Geodesic star convexity for interactive image segmentation. 2010. 3

[11] X. He, R. Zemel, and M. Carreira-Perpinán. Multiscale conditional random fields for image labeling. In *CVPR*, volume 2, pages II–695, 2004. 4

[12] D. Hoiem, A. Efros, and M. Hebert. Recovering surface layout from an image. *IJCV*, 75(1):151–172, 2007. 3

[13] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *TPAMI*, 20(11):1254–1259, 1998. 1, 2, 4

[14] T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. In *ICCV*, pages 2106–2113, 2009. 1, 2, 4, 5

[15] R. Kimchi and M. Peterson. Figure-ground segmentation can occur without attention. *Psychological Science*, 19(7):660, 2008. 1, 7

[16] G. Kootstra, A. Nederveen, and B. De Boer. Paying attention to symmetry. In *BMVC2008*, pages 1115–1125. Citeseer, 2008. 1, 3

[17] Y. Lu, W. Zhang, H. Lu, and X. Xue. Salient object detection using concavity context. *ICCV*, 2011. 1, 2, 3

[18] A. Makadia, V. Pavlovic, and S. Kumar. A new baseline for image annotation. In *ECCV*, volume 8, pages 316–329. Citeseer, 2008. 5

[19] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*, 42(3):145–175, 2001. 2

[20] S. Palmer. *Vision science: Photons to phenomenology*, volume 1. MIT press Cambridge, MA., 1999. 1, 3

[21] F. Qiu, T. Sugihara, and R. von der Heydt. Figure-ground mechanisms provide structure for selective attention. *Nature neuroscience*, 10(11):1492–1499, 2007. 1, 7

[22] S. Ramanathan, H. Katti, N. Sebe, M. Kankanhalli, and T. Chua. An eye fixation database for saliency detection in images. *ECCV*, pages 30–43, 2010. 4

[23] D. Reisfeld, H. Wolfson, and Y. Yeshurun. Context-free attentional operators: the generalized symmetry transform. *IJCV*, 14(2):119–130, 1995. 3

[24] X. Ren, C. Fowlkes, and J. Malik. Figure/ground assignment in natural images. *ECCV*, pages 614–627, 2006. 1, 2

[25] R. Rosenholtz. A simple saliency model predicts a number of motion popout phenomena. *Vision Research*, 39(19):3157–3163, 1999. 2

[26] B. Russell, W. Freeman, A. Efros, J. Sivic, and A. Zisserman. Using multiple segmentations to discover objects and their extent in image collections. In *CVPR*, volume 2, pages 1605–1614, 2006. 3

[27] B. Russell, A. Torralba, K. Murphy, and W. Freeman. Labelme: a database and web-based tool for image annotation. *IJCV*, 77(1):157–173, 2008. 5

[28] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. *IJCV*, 81(1):2–23, 2009. 4

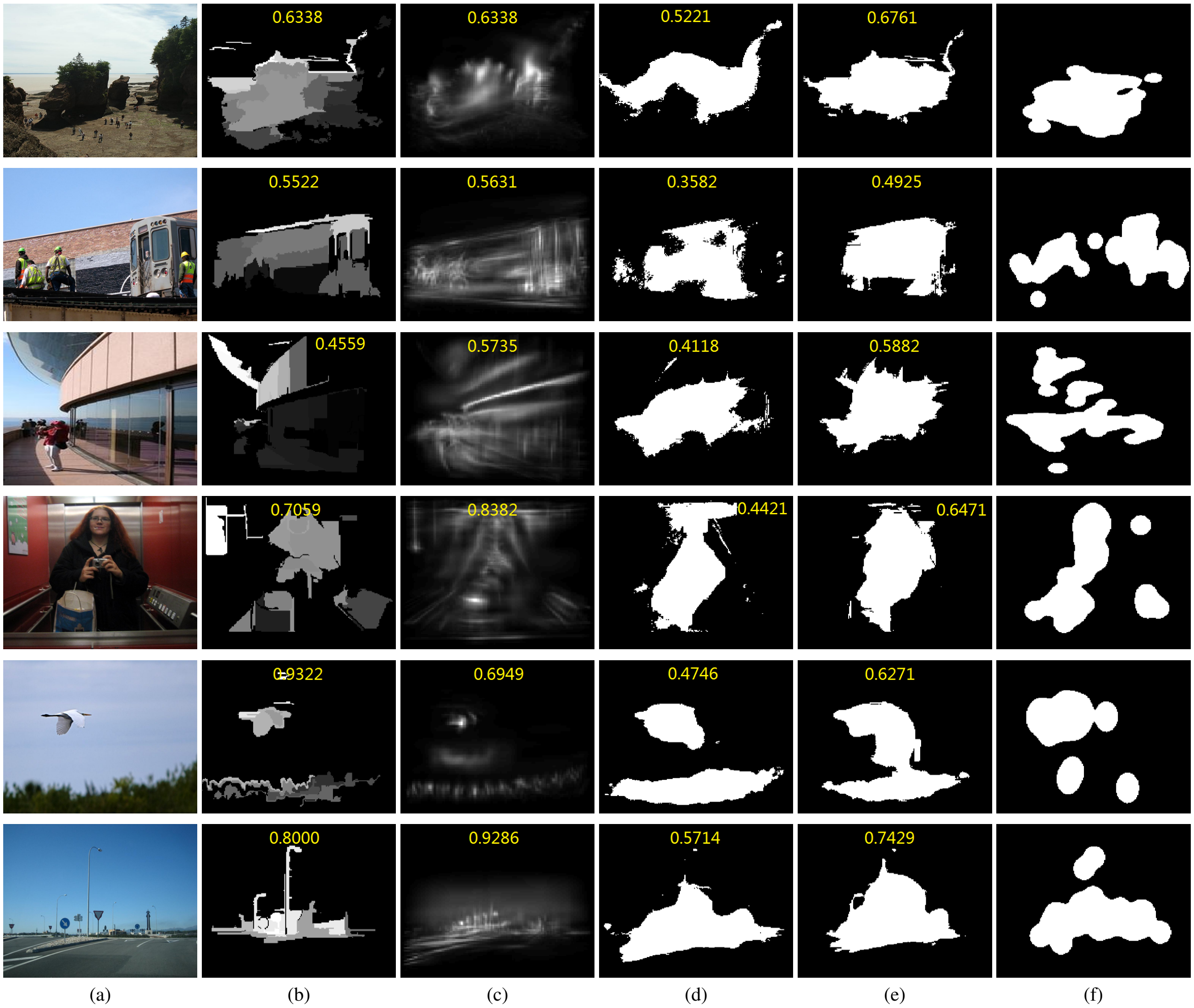

Figure 8. Some good results of our proposed features and model under 20 percent salient. (a) Input image. (b) Convexity map. (c) Symmetry map. (d) Baseline model + basic features (F1-8). (e) All feature (F1-11) + new model. (f) Groundtruth. True positive rate is shown in yellow digits for each map (binary maps are not shown for (a) and (b) due to space limitation.

[29] E. Simoncelli and W. Freeman. The steerable pyramid: A flexible architecture for multi-scale derivative computation. In *ICIP*, volume 3, pages 444–447, 1995. 2

[30] Y. Sugano, Y. Matsushita, and Y. Sato. Calibration-free gaze sensing using saliency maps. In *CVPR*, 2010. 1

[31] O. Veksler. Star shape prior for graph-cut image segmentation. *ECCV*, pages 454–467, 2008. 3

[32] P. Viola and M. Jones. Robust real-time object detection. *IJCV*, 57(2):137–154, 2002. 2