Learning to predict where humans look

Learning to Predict Where Humans Look

Tilke Judd tjudd@mit.edu
Krista Ehinger kehinger@mit.edu
Frédo Durand fredo@csail.mit.edu
Antonio Torralba torralba@csail.mit.edu

MIT Computer Science Artificial Intelligence Laboratory and MIT Brain and Cognitive Sciences

Abstract

For many applications in graphics, design, and human computer interaction, it is essential to understand where humans look in a scene. Where eye tracking devices are not a viable option, models of saliency can be used to predict fixation locations. Most saliency approaches are based on bottom-up computation that does not consider top-down image semantics and often does not match actual eye movements. To address this problem, we collected eye tracking data of 15 viewers on 1003 images and use this database as training and testing examples to learn a model of saliency based on low, middle and high-level image features. This large database of eye tracking data is publicly available with this paper.

1. Introduction

For many applications in graphics, design, and human computer interaction, it is essential to understand where humans look in a scene. For example, an understanding of visual attention is useful for automatic image cropping [16], thumbnailing, or image search. It can be used to direct foveated image and video compression [22], [7] and levels of detail in non-photorealistic rendering [4]. It can also be used in advertising design, adaptive image display on small devices, or seam carving [14].

Some of these applications have been demonstrated by incorporating eye tracking into the process: a user sits in front of a computer with an eye tracker that records the user's fixations and feeds the data into the method. However, eye tracking is not always an option. Eye trackers are expensive and interactive techniques are a burden when processing lots of data. Therefore, it is necessary to have a way to predict where users will look without the eye tracking hardware. As an alternative, models of saliency have been used to measure the conspicuity of a location, or the likelihood of a location to attract the attention of human observers.

Most models of saliency [9] [13] [8] are biologically inspired and based on a bottom-up computational model. Typically, multiple low-level visual features such as intensity, color, orientation, texture and motion are extracted from the image at multiple scales. After a saliency map is computed for each of the features, they are normalized and combined in a linear or non-linear fashion into a master saliency map that represents the saliency of each pixel. Sometimes specific locations are identified through a combination of winner-take-all and inhibition-of-return operations.

Though the models do well qualitatively, the models have limited use because they frequently do not match actual human saccades from eye-tracking data, as in Fig 2, and finding a closer match depends on tuning many design parameters.

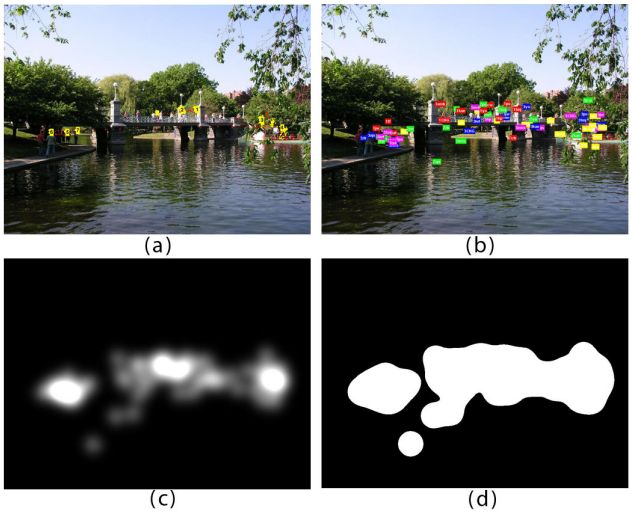

Figure 1. **Eye tracking data.** *We collected eye-tracking data on 1003 images from 15 viewers to use as ground truth data to train a model of saliency using machine learning. Gaze tracking paths and fixation locations are recorded for each viewer (b). A continuous saliency map (c) is found by convolving a gaussian over the fixation locations of all users. This saliency map can be thresholded to show the most salient 20 percent of the image (d).*

2009 IEEE 12th International Conference on Computer Vision (ICCV)
978-1-4244-4419-9/09/$25.00 ©2009 IEEE

(a) Original image (b) Itti and Koch Saliency Map (c) eye tracking locations

Figure 2. **Current saliency models do not accurately predict human fixations.** *In row one, the low-level model selects bright spots of light as salient while viewers look at the human. In row two, the low level model selects the building's strong edges and windows as salient while viewers fixate on the text.*

We make two contributions in this paper. The first is a large database of eye tracking experiments with labels and analysis, and the second is a supervised learning model of saliency which combines both bottom-up image-based saliency cues and top-down image semantic dependent cues. Our database consists of eye tracking data from 15 different users across 1003 images. To our knowledge, it is the first time such an extensive collection of eye tracking data is available for quantitative analysis. For a given image, the eye tracking data is used to create a "ground truth" saliency map which represents where viewers actually look (Fig 1). We propose a set of low, mid and high-level image features used to define salient locations and use a linear support vector machine to train a model of saliency. We compare the performance of saliency models created with different features and show how combining all features produces the highest performing model. As a demonstration that our model can be used for graphics applications, we show the DeCarlo and Santella [4] abstracted nonphotorealistic rendering technique adapted to use our saliency model instead of eye tracking input.

Other researchers have also made some headway on improving low level saliency models. Bruce and Tsotsos [2] present a model for visual saliency built on a first principles information theoretic formulation dubbed Attention based on Information Maximization (AIM) which performs marginally better than the Itti model. Avraham and Lindenbaum's work on Esaliency [1] uses a stochastic model to estimate the most probable targets mathematically. The main difference between these works and ours is that their models are derived mathematically and not trained directly from a large database of eye tracking data. Cerf et al. [3] improve upon the Itti model by adding face detection to the model. In addition to adding face detection, we add several other higher level features which provide us with an increased performance over both the Itti and Cerf models.

Our work is most closely related to the work of Kienzle et al. [10] who also learn a model of saliency directly from human eye movement data. Their model consists of a nonlinear mapping from a normalized image patch to a real value, trained to yield positive outputs on fixated patches, and negative outputs on randomly selected image patches. In contrast to our work, they only used low-level features. Furthermore, their training set comprises only 200 grayscale natural scene images.

In the specific situation of trying to predict where people look in a pedestrian search task Ehinger et al. [5] show that a model of search guidance combining three sources: low level saliency, target features, and scene context, outperforms models based on any of these single sources. Our work focuses on predicting saliency in a free viewing context and creates a model with a larger set of image features.

2. Database of eye tracking data

We collected a large database of eye tracking data to allow large-scale quantitative analysis of fixation points and gaze paths and to provide ground truth data for saliency model research. The images, eye tracking data, and accompanying code in Matlab are all available on the web to facilitate research in perception and saliency across the vision and graphics community.

2.1. Data gathering protocol

We collected 1003 random images from Flickr creative commons and LabelMe [15] (Fig 3) and recorded eye tracking data from fifteen users who free viewed these images. The longest dimension of each image was 1024 pixels and the other dimension ranged from 405 to 1024 with the majority at 768 pixels. There were 779 landscape images and 228 portrait images. The users were males and females between the ages of 18 and 35. Two of the viewers were researchers on the project and the others were naive viewers. All viewers sat at a distance of approximately two feet from a 19 inch computer screen of resolution 1280x1024 in a dark room and used a chin rest to stabilize their head. An eye tracker recorded their gaze path on a separate computer as they viewed each image at full resolution for 3 seconds separated by 1 second of viewing a gray screen. To ensure high-quality tracking results, we checked camera calibration every 50 images. We divided the viewing into two sessions of 500 randomly ordered images. Each session was done on average at one week apart. We provided a mem-

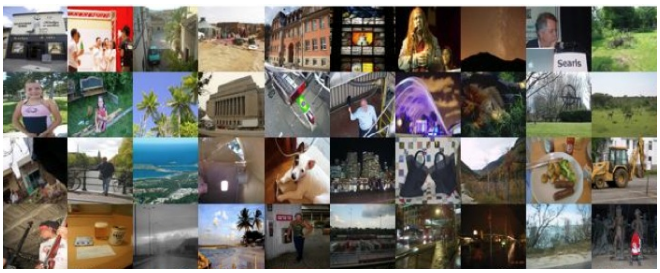

Figure 3. **Images.** *A sample of the 1003 images that we collected from Flickr and LabelMe. Though they were shown at original resolution and aspect ratio in the experiment, they have been resized for viewing here.*

ory test at the end of both viewings to motivate users to pay attention to the images: we showed them 100 images and asked them to indicate which ones they had seen before. We discarded the first fixation from each scanpath to avoid adding trivial information from the initial center fixation.

In order to obtain a continuous saliency map of an image from the eye tracking data of a user, we convolve a gaussian filter across the user's fixation locations, similar to the "landscape map" of [20]. We also generate a saliency map of the average locations fixated by all viewers. We can choose to threshold this continuous saliency map to get a binary map of the top n percent salient locations of the image (Fig 1d).

2.2. Analysis of dataset

For some images, all viewers fixate on the same locations, while in other images viewers' fixations are dispersed all over the image. We analyze this consistency of human fixations over an image by measuring the entropy of the average continuous saliency map across viewers. Though the original images were of varying aspect rations, we resized them to 200x200 pixel images before calculating entropy. Figure 4 shows a histogram of the entropies of the images in our database. It also shows a sample of 12 saliency maps with lowest and highest entropy and their corresponding images.

Our data indicates a strong bias for human fixations to be near the center of the image, as is consistent with previously analyzed eye tracking datasets [23] [19]. Figure 4 shows the average human saliency map from all 1003 images. 40% of fixations lie within the center 11% of the image; 70% of fixations lie within the center 25% of the image. This bias has often been attributed to the setup of the experiment where users are placed centrally in front of the screen, and to the fact that human photographers tend to place objects of interest in the center of photographs [23].

We use an ROC metric to evaluate the performance of human saliency maps to predict eye fixations. Using this method, the saliency map from the fixation locations of one

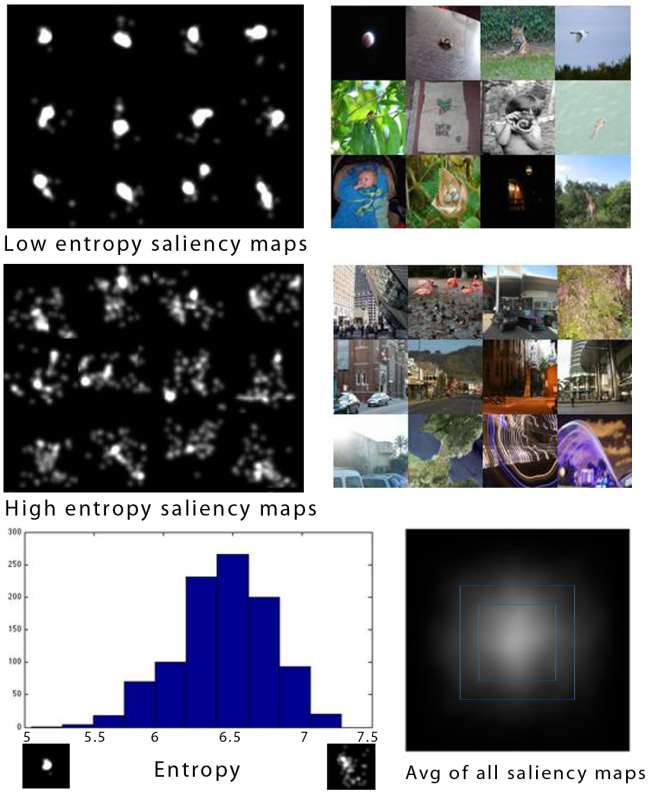

Figure 4. **Analysis of fixation locations.** *The first two rows show examples of saliency maps made from human fixations with low and high entropy and their corresponding images. Images with high consistency/low entropy tend to have one central object while images with low consistency/high entropy are often images with several different textures. Bottom left is a histogram of the saliency map entropies. Bottom right is a plot of all the saliency maps from human eye fixations indicating a strong bias to the center of the image. 40% and 70% of fixations lie within the indicated rectangles.*

user is treated as a binary classifier on every pixel in the image. Saliency maps are thresholded such that a given percent of the image pixels are classified as fixated and the rest are classified as not fixated. The human fixations from the other 14 humans are used as ground truth. By varying the threshold, the ROC curve is drawn and the area under the curve indicates how well the saliency map from one user can predict the ground truth fixations. Figure 5 shows the average ROC curve over all users and all images. Note that human performance is remarkably good: 60% of the ground truth human fixations are within the top 5% salient areas of a novel viewer's saliency map, and 90 percent are within the top 20 percent salient locations.

As stated before, the fixations in the database have a strong bias towards the center. Because of this, we find that simply using a Gaussian blob centered in the middle of the image as the saliency map produces excellent results,

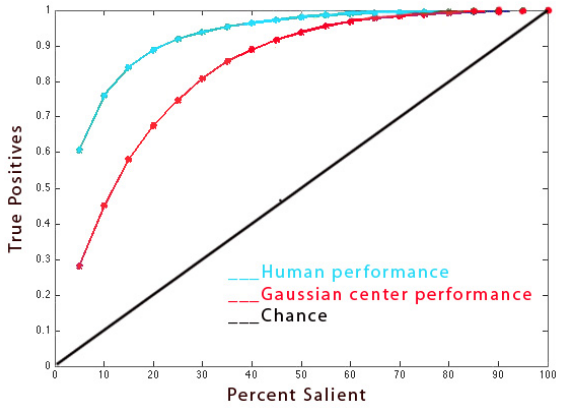

Figure 5. *In this ROC curve, human performance is very high demonstrating that the locations where a human looks are very indicative of where other humans have looked. The gaussian center model performs much better than chance because of the strong bias of the fixations in the database towards the center.*

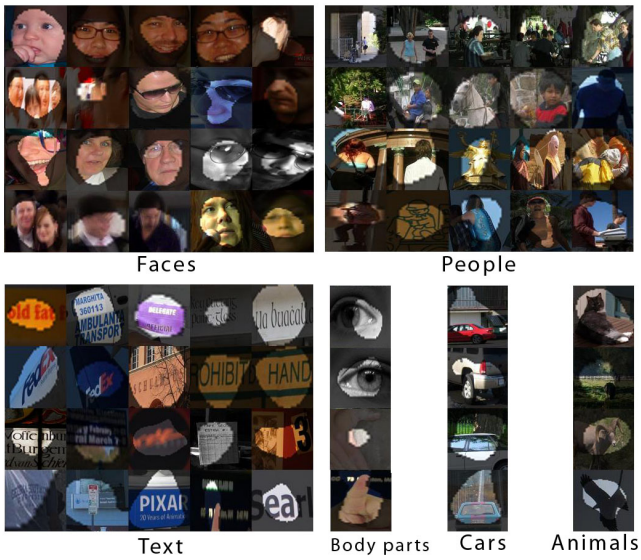

Figure 6. **Objects of interest.** *In our database, viewers frequently fixated on faces, people, and text. Other fixations were on body parts such as eyes and hands, cars and animals. We found the above image areas by selecting bounding boxes around connected areas of salient pixels on an image overlayed with its 3% salient mask.*

as noted for other datasets as well by [23] [11]. We plot the ROC curve for the center Gaussian on figure 5.

In order to analyze fixations on specific objects and image features we hand labeled our image dataset. For each image, we labeled bounding boxes around any faces and text, and indicated a line for the horizon if present. Using these labeled bounding boxes we calculated that 10% of fixations are on faces (Fig 6). Though we did not label all people, we noticed that many fixations landed on people (including representations of people like drawings or sculptures) even if their faces were not visible. In addition, 11% of fixations are on text. This may be because signs are innately designed to be salient (for example a stop sign or a store sign are created specifically to draw attention). We use these ground truth labels to study fixation prediction performance on faces and as a ground truth for face and horizon detection. We also qualitatively found that fixations from our database are often on animals, cars, and human body parts like eyes and hands. These objects reflect both a notion of what humans are attracted to and what objects are in our dataset.

By analyzing images with faces we noticed that viewers fixate on faces when they are within a certain size of the image but fixate of parts of the face (eyes, nose, lips) when presented with a close up of a face (Fig 7). This suggests that there is a certain size for a region of interest (ROI) that a person fixates on. To get a quick sense of the size of ROIs, we drew a rough bounding box around clustered fixations on 30 images. Figure 7 shows the histogram of the radii of the resulting 102 ROIs. Investigating this concept is an interesting area of future work.

Figure 7. **Size of regions of interest** *In many images, viewers fixate on human faces. However, when viewing the close up of a face, they look at specific parts of a face rather than the face as a whole, suggesting a constrained area of the region of interest. On the right is a histogram of the radii of the regions of interest in pixels.*

3. Learning a model of saliency

In contrast to previous computational models that combine a set of biologically plausible filters together to estimate visual saliency, we use a learning approach to train a classifier directly from human eye tracking data.

3.1. Features used for machine learning

The following are the low-, mid- and high-level features that we were motivated to work with after analyzing our dataset. For each image, we precomputed the features for every pixel of the image resized to 200x200 and used these to train our model.

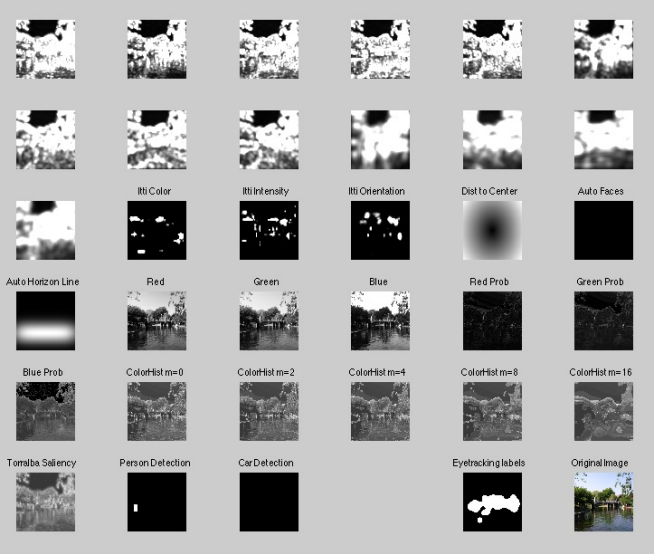

Figure 8. **Features.** *A sample image (bottom right) and 33 of the features that we use to train the model. These include subband features, Itti and Koch saliency channels, distance to the center, color features and automatic horizon, face, person and car detectors. The labels for our training on this image are based on a thresholded saliency map derived from human fixations (to the left of bottom right).*

Low-level features Because they are physiologically plausible and have been shown to correlate with visual attention, we use the local energy of the steerable pyramid filters [17] as features. We currently find the pyramid subbands in four orientations and three scales (see Fig 8, first 13 images). We also include features used in a simple saliency model described by Torralba [12] and Rosenholtz [13] based on subband pyramids (Fig 8, bottom left).

Intensity, orientation and color contrast have long been seen as important features for bottom-up saliency. We include the three channels corresponding to these image features as calculated by Itti and Koch's saliency method [9].

We include the values of the red, green and blue channels, as well as the probabilities of each of these channels as features (Fig 8, images 20 to 25) and the probability of each color as computed from 3D color histograms of the image filtered with a median filter at 6 different scales (Fig 8, images 26 to 31).

Mid-level features Because most objects rest on the surface of the earth, the horizon is a place humans naturally look for salient objects. We train a horizon line detector from mid-level gist features [12].

High-level features Because we found that humans fixated so consistently on people and faces we run the Viola Jones face detector [21] and the Felzenszwalb person detector [6] and include these as features to our model.

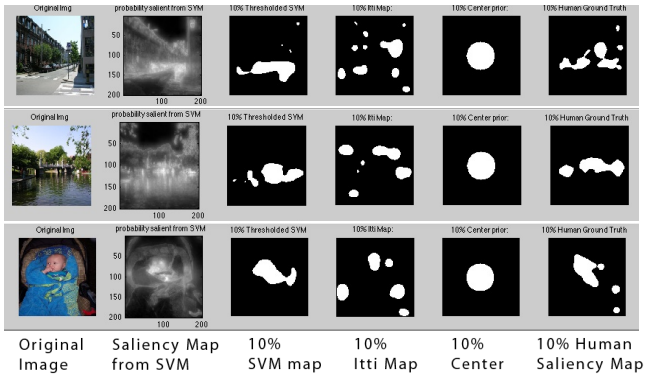

Figure 9. **Comparison of saliency maps.** *Each row of images compares the predictors of our SVM saliency model, the Itti saliency map, the center prior, and the human ground truth, all thresholded to show the top 10 percent salient locations.*

Center prior When humans take pictures, they naturally frame an object of interest near the center of the image. For this reason, we include a feature which indicates the distance to the center for each pixel.

3.2. Training

In order to train and test our model, we divided our set of images into 903 training images and 100 testing images. From each image we chose 10 positively labeled pixels randomly from the top 20% salient locations of the human ground truth saliency map and 10 negatively labeled pixels from the bottom 70% salient locations to yield a training set of 18060 samples and testing set of 2000 samples. We found that increasing the number of samples chosen per image above 10 did not increase performance. It is probable that after a certain number of samples per image, new samples only provide redundant information. We chose samples from the top 20% and bottom 70% in order to have samples that were strongly positive and strongly negative; we avoided samples on the boundary between the two. We did not choose any samples within 10 pixels of the boundary of the image.

Our tests on models trained using ratios of negative to positive samples ranging from 1 to 5 showed no change in the resulting ROC curve, so we chose to use a ratio of 1:1.

We normalized the features of our training set to have zero mean and unit variance and used the same normalization parameters to normalize our test data.

We used the liblinear support vector machine to train a model on the 9030 positive and 9030 negative training samples. We used models with linear kernels because we found from experimentation that they performed as well as models with radial basis function kernels and models found with

multiple kernel learning [18] for our specific task. Linear models are also faster to compute and the resulting weights of features are easier to understand. We set the misclassification cost c at 1. We found that performance was the same for c = 1 to c = 10,000 and decreased when smaller than 1.

3.3. Performance

We measure performance of saliency models in two ways. First, we measure performance of each model by its ROC curve. Second, we examine the performance of different models on specific subsets of samples: samples inside and outside a central area of the image and on faces.

Performance on testing images In Figure 10, we see a ROC curve describing the performance of different saliency models averaged over all testing images. For each image we predict the saliency per pixel using a specific trained model. Instead of using the predicted labels (indicated by the sign of $w^T x + b$ where w and b are learned parameters and x refers to the feature vector), we use the value of $w^T x + b$ as a continuous saliency map which indicates how salient each pixel is. Then we threshold this saliency map at n =1, 3, 5, 10, 15, 20, 25, and 30 percent of the image for binary saliency maps which are typically relevant for applications. For each binary map, we find the percentage of human fixations within the salient areas of the map as the measure of performance. Notice that as the percentage of the image considered salient goes to 100%, the predictability, or percentage of human fixations within the salient locations also goes to 100%.

We make the following observations from the ROC curves: (1) The model with all features combined outperforms models trained on single sets of features and models trained on competing saliency features from Torralba and Rozenholtz, Itti and Koch and Cerf et al. Note that we implement the Cerf et al. method by training an SVM on Itti features and face detection alone. We learn the best weights for the linear combination of features instead of using equal weights as they do. (2) The model with all features reaches 88% of the way to human performance. For example, when images are thresholded at 20% salient, our model performs at 75% while humans are at 85%. (3) The model with all features except the distance to the center performs as well as the model based on the distance to the center. This is quite good considering this model does not leverage any of the information about location and thus does not at all benefit from the huge bias of fixations toward the center. (4) The model trained on all features except the center performs much better than any of the models trained on single sets of features. For example, at the 20% salient location threshold, the Torralba based model performs at 50% while the all-in-without-center model performs at 60% for a 20% jump in performance. (5) Though object detectors may be

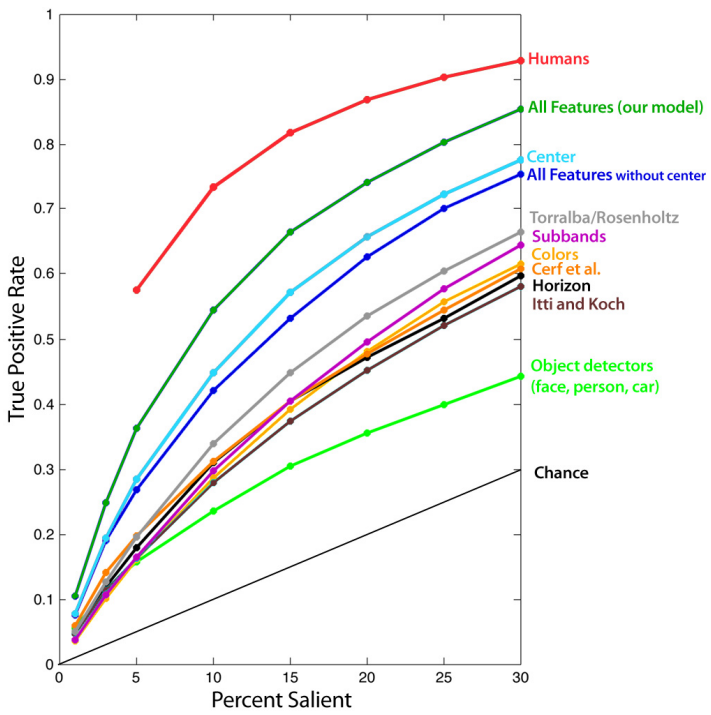

Figure 10. *The ROC curve of performances for SVMs trained on each set of features individually and combined together. We also plot human performance and chance for comparison.*

very good at locating salient objects when those objects are present in an image, it is not good at locating other salient locations when the objects are not present. Thus, the overall performance for the object detector model is low and these features should be used only in conjunction with other features. (6) All models perform significantly better than chance indicating that each of the features individually do have some power to predict salient locations.

We measure which features add most to the model by calculating the delta improvement between the center model and the center model with a given set of features. We observe that subband features and Torralba's features (which use subband features) add the greatest improvement. After that is color features, horizon detection, face and object detectors, and Itti channels.

Performance on testing samples To understand the impact of the bias towards the center of the dataset for some models, we divided each image into a circular central and a peripheral region. The central region was defined by the model based only on the feature which gave the distance of the example to the center. In this model, any sample farther than 0.42 units away from the center (where the distance from the center to the corner is 1) was labeled negative and anything closer was labeled positive. This is equivalent to the center 27.7% of the image. Given this threshold, we di-

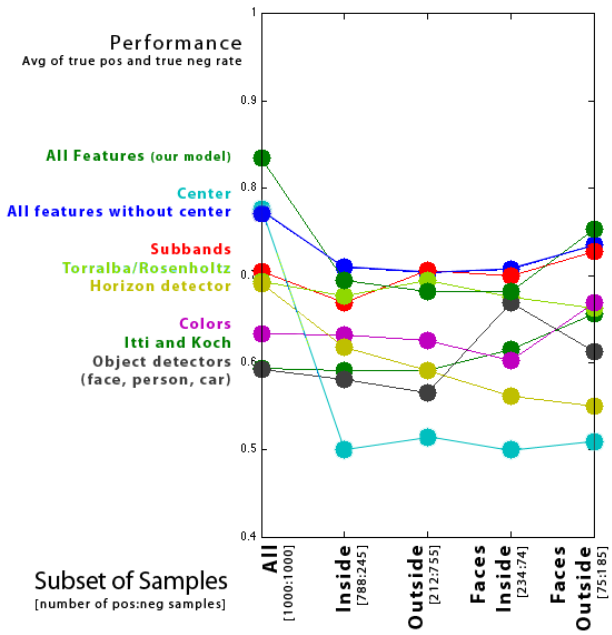

Figure 11. *Here we show the average rate of true positives and true negatives for SVMs trained with different feature sets on different* subsets *of samples. This value is equivalent to the performance of the model if there were an equal number of positive and negative samples in each subset.*

vided the samples to those *inside* and *outside* the center. In addition, we chose to look at samples that landed on faces since viewers were particularly attracted by them.

In Figure 11 we plot performance of the model for different subsets of samples. The performance here is defined as the average of the true positive and true negative rates. This is equivalent to the performance of the model if there were an equal number of positive and negative samples in each subset.

We make the following observations about the trained models from this measure of performance: (1) Even though center model performs well over all the samples (both samples inside and outside the center), it performs only as well as chance for the other subsets of samples. (2) While over all samples the performance of the center model and the all-features-without-center model perform the same, the later model performs more robustly over all subsets of samples. (3) Understandably, the model trained on features from object detectors for faces, people and cars performs better on the subsets with faces. (4) The SVMs using the center prior feature and the one using all features perform very well on 1000 positive and negative random testing points but are outperformed both in the inside and outside region. This paradox stems from the fact that 79% of the 1000 salient testing points are in the inside region, whereas 75% of the non-salient testing points are in the outside. One can show that this biased distribution provides a lift in performance for methods that would either have a high true negative rate outside or a high true positive rate inside, such as the center prior.

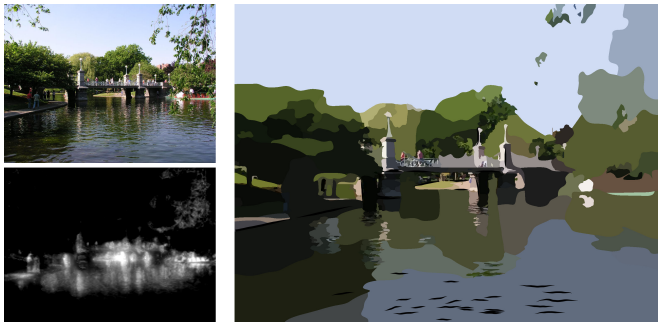

Figure 12. **Stylization and abstraction of photographs** *DeCarlo and Santella [4] use eye tracking data to decide how to render a photograph with differing levels of detail. We replicate this application without the need for eye tracking hardware.*

Discussion This eye tracking database allows us to quantify how consistent human fixations are across an image. In general, the fixation locations of several humans is strongly indicative of where a new viewer will look. So far, computer generated models have not matched humans' ability to predict fixation locations though we feel we have moved a step closer in that direction by using a model that combines low, mid and high level features.

Qualitatively, we learned that when free viewing images, humans consistently look at some common objects: They look at text, other people and specifically faces. If not people, they look at other living animals and specifically their faces. In the absence of specific objects or text, humans tend towards the center of the image or locations where low-level features are salient. As text, face, person and other object detectors get better, models of saliency which include object detectors will also get better. Though all these trends are not surprising, we are excited that this database will allow us to measure the trends quantitatively.

3.4. Applications

A good saliency model enables many applications that automatically take into account a notion of human perception: where humans look and what they are interested in. As an example, we use our model in conjunction with the technique of DeCarlo and Santella [4] to automatically create a non photorealistic rendering of a photograph with different levels of detail (Fig 12). They render more details at the locations users fixated on and less detail in the rest of the image. While they require information from an eye tracking device in order to tailor the level of detail, we use our saliency model to predict locations where people look.

4. Conclusion

In this work we make the following contributions: We develop a collection of eye tracking data from 15 people across 1003 images and have made it public for research use. This is the largest eye tracking database of natural images that we are aware of and permits large-scale quantitative analysis of fixations points and gaze paths. We use machine learning to train a bottom-up, top-down model of saliency based on low, mid and high-level image features. We demonstrate that our model outperforms several existing models and the center prior. Finally, we show an example of how our model can be used in practice for graphics applications.

For future work we are interested in understanding the impact of framing, cropping and scaling images on fixations. We believe that the same image cropped at different sizes will lead viewers to fixate on different objects in the image and should be more carefully examined.

Acknowledgments This work was supported by NSF CAREER awards 0447561 and IIS 0747120. Frédo Durand acknowledges a Microsoft Research New Faculty Fellowship and a Sloan Fellowship, in addition to Royal Dutch Shell, the Quanta T-Party, and the MIT-Singapore GAMBIT lab. Tilke Judd was supported by a Xerox graduate fellowship. We thank Aude Oliva for the use of her eye tracker and Barbara Hidalgo-Sotelo for help with eye tracking. We thank Nicolas Pinto and Yann LeTallec for insightful discussions, and the ICCV reviewers for their feedback on this work.

References

[1] T. Avraham and M. Lindenbaum. Esaliency: Meaningful attention using stochastic image modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 99(1), 2009.

[2] N. D. B. Bruce and J. K. Tsotsos. Saliency, attention, and visual search: An information theoretic approach. *Journal of Vision*, 9(3):1–24, 3 2009.

[3] M. Cerf, J. Harel, W. Einhauser, and C. Koch. Predicting human gaze using low-level saliency combined with face detection. In J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, editors, *NIPS*. MIT Press, 2007.

[4] D. DeCarlo and A. Santella. Stylization and abstraction of photographs. *ACM Transactions on Graphics*, 21(3):769–776, July 2002.

[5] K. Ehinger, B. Hidalgo-Sotelo, A. Torralba, and A. Oliva. Modeling search for people in 900 scenes: A combined source model of eye guidance. *Visual Cognition*, 2009.

[6] P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8, June 2008.

[7] W. S. Geisler and J. S. Perry. A real-time foveated multiresolution system for low-bandwidth video communication. In *in Proc. SPIE*, pages 294–305, 1998.

[8] X. Hou and L. Zhang. Saliency detection: A spectral residual approach. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 0:1–8, 2007.

[9] L. Itti and C. Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, 40:1489–1506, 2000.

[10] W. Kienzle, F. A. Wichmann, B. Schölkopf, and M. O. Franz. A nonparametric approach to bottom-up visual saliency. In B. Schölkopf, J. C. Platt, and T. Hoffman, editors, *NIPS*, pages 689–696. MIT Press, 2006.

[11] O. L. Meur, P. L. Callet, and D. Barba. Predicting visual fixations on video based on low-level visual features. *Vision Research*, 47(19):2483 – 2498, 2007.

[12] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42:145–175, 2001.

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

[14] M. Rubinstein, A. Shamir, and S. Avidan. Improved seam carving for video retargeting. *ACM Transactions on Graphics (SIGGRAPH)*, 2008.

[15] B. Russell, A. Torralba, K. Murphy, and W. Freeman. Labelme: a database and web-based tool for image annotation. MIT AI Lab Memo AIM-2005-025, MIT CSAIL, Sept. 2005.

[16] A. Santella, M. Agrawala, D. DeCarlo, D. Salesin, and M. Cohen. Gaze-based interaction for semi-automatic photo cropping. In *CHI '06: Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 771–780, New York, NY, USA, 2006. ACM.

[17] E. P. Simoncelli and W. T. Freeman. The steerable pyramid: A flexible architecture for multi-scale derivative computation. pages 444–447, 1995.

[18] S. Sonnenburg, G. Rätsch, C. Schäfer, and B. Schölkopf. Large scale multiple kernel learning. *J. Mach. Learn. Res.*, 7:1531–1565, 2006.

[19] 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.*, 7(14):1–17, 11 2007.

[20] B. M. Velichkovsky, M. Pomplun, J. Rieser, and H. J. Ritter. *Attention and Communication: Eye-Movement-Based Research Paradigms*. Visual Attention and Cognition. Elsevier Science B.V., Amsterdam, 1996.

[21] P. Viola and M. Jones. Robust real-time object detection. In *International Journal of Computer Vision*, 2001.

[22] Z. Wang, L. Lu, and A. C. Bovik. Foveation scalable video coding with automatic fixation selection. *IEEE Trans. Image Processing*, 12:243–254, 2003.

[23] 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.*, 8(7):1–20, 12 2008.