An Eye Fixation Database for Saliency Detection in Images

Subramanian Ramanathan[1], Harish Katti[2], Nicu Sebe[1],
Mohan Kankanhalli[2], and Tat-Seng Chua[2]

[1] Department of Information Engineering and Computer Science,
University of Trento, Italy
[2] School of Computing, National University of Singapore (NUS), Singapore
subramanian@disi.unitn.it

Abstract. To learn the preferential visual attention given by humans to specific image content, we present NUSEF- an eye fixation database compiled from a pool of 758 images and 75 subjects. Eye fixations are an excellent modality to learn semantics-driven human understanding of images, which is vastly different from feature-driven approaches employed by saliency computation algorithms. The database comprises fixation patterns acquired using an eye-tracker, as subjects free-viewed images corresponding to many semantic categories such as *faces* (human and mammal), *nudes* and *actions* (*look*, *read* and *shoot*). The consistent presence of fixation clusters around specific image regions confirms that visual attention is not subjective, but is directed towards *salient* objects and object-interactions.

We then show how the fixation clusters can be exploited for enhancing image understanding, by using our eye fixation database in an active image segmentation application. Apart from proposing a mechanism to automatically determine characteristic fixation seeds for segmentation, we show that the use of fixation seeds generated from multiple fixation clusters on the *salient* object can lead to a 10% improvement in segmentation performance over the state-of-the-art.

1 Introduction

The past decade has seen tremendous progress in the field of image understanding and retrieval. Breakthroughs have been achieved in robustly detecting and characterizing image objects [1,2], as well as in classifying scenes from image and video [3,4]. Nevertheless, computer vision's goal to 'enable computers to see what humans see' currently seems out of reach, and contemporary algorithms are focused on accurately interpreting and deriving a bag of keywords [5,6] for visual content.

Since human cognition is designed to process only limited information at any given time, our understanding of images is influenced by *what we attend to*, termed **visual attention**. Significant recent research has been devoted to understanding human visual attention. Through an urn model for object recall,

K. Daniilidis, P. Maragos, N. Paragios (Eds.): ECCV 2010, Part IV, LNCS 6314, pp. 30–43, 2010.

the authors in [7] demonstrate the inherent order of 'importance' assigned by human observers to scene objects. Most recently, the need for an eye-tracking database to train a model to predict where humans would look at in an image, is discussed in [8]. The database is motivated by the fact that human-observed 'regions of interest' are driven by top-down (task/semantics-based) as well as bottom-up (content/feature-based) processing, and generally don't match those predicted by image saliency computation methods [9,10,11,12,13,14].

In this paper, we present NUSEF- a database of eye-fixations compiled using an eye-tracker from a pool of 75 subjects and 758 images, spanning a large number of semantic categories. While [8] presents an eye-fixation database to learn what viewers attend to in everyday scenes, our database consists of a significant number of semantically *affective* (emotion-evoking) images. We believe that the analysis of visual attention for affective content can add a new dimension to eye-tracking research, and also offer interesting insights into how eye fixations are driven by image semantics- for *e.g.*, normal (*neutral*, *smiling*) faces are viewed differently from strongly expressive (*surprise*, *disgust*) faces and there are characteristic fixation patterns for images depicting actions (such as *look*, *read* and *shoot*) [15].

Our experimental results indicate that eye fixations are heavily influenced by image semantics and are consistently specific to *salient* (most important/meaningful) scene objects and object-interactions, which we call **attentional-bias**. Since the fixation data was acquired as subjects free-viewed images (*i.e*, in the absence of any pre-specified task), this observation is in contrast to the long-standing argument that top-down content processing by humans is subjective, and therefore, prone to extensive variability. Indeed, similar observations are also made in [16], where the authors argue that visual attention is essentially guided by recognized objects, with low-level saliency contributing only indirectly. We hope that this fixation database will particularly benefit members of the vision, multimedia, cognitive science and HCI communities.

Also, viewers exhibit exploratory behavior and attend to multiple regions-of-interest, as they observe salient objects. For *e.g.*, in *face* images, fixations are not concentrated around the center of the face but spread around the eyes, nose and mouth. We demonstrate how this phenomenon can be exploited for enhancing image understanding, using active image segmentation as an example. An algorithm for automatically segmenting the image region containing a fixation point is described in [17]. Employing the fixation point as a representative seed for the foreground object, the set of boundary edges around the fixated region are computed through energy minimization in polar space to produce promising results. While the authors claim that the fixation can be any random point in the object's interior, no methodology is provided to automatically select fixation points. On the contrary, a manually annotated point is taken as the fixation seed. Using acquired fixation patterns, we (i) propose a mechanism to automatically select the fixation seed and (ii) show how viewer's exploratory behavior can be exploited to generate multiple fixation seeds for segmentation, thereby contributing to a tremendous improvement in segmentation performance.

To summarize, the main contributions of this paper are the following:

1. A rich database of eye fixations for an image set spanning a comprehensive list of semantic categories, including a significant number of *affective* images. We believe that our eye fixation database, along with [8], will offer an excellent repository of ground truth data for visual attention and image understanding research.
2. Exploiting the attentional bias, or the clustering of fixations around the *salient* object, to automatically generate the fixation seed for active image segmentation.
3. Improving on the active segmentation performance achieved in [17] by 10%, upon generating multiple fixation seeds for segmentation within the *salient* object.

The paper outline is as follows. The next section describes acquisition, content, and other key characteristics of the eye fixation database. Section 3 discusses how attentional bias is exploited to improve the performance of active segmentation, along with the experimental results. We end with the main conclusions and directions for future work in Section 4.

2 Eye Fixation Database

The NUSEF (NUS Eye Fixation) database was acquired from undergraduate and graduate volunteers aged 18-35 years (μ=24.9, σ=3.4). The **ASL**^{TM} eye-tracker was used to non-invasively record eye fixations, as subjects free-viewed images. We chose a diverse set of 1024×728 resolution images, representative of various semantic concepts and capturing objects at varying scale, illumination and orientation, based on quality and aspect ratio constraints. Images comprised everyday scenes from *Flickr*, aesthetic content from *Photo.net*, *Google* images and emotion-evoking IAPS [18] pictures. The images [1] and Matlab code to visualize the image-wise and user-wise fixation characteristics have been made available at *http://mmas.comp.nus.edu.sg/NUSEF.html*.

2.1 Data Collection Protocol

From a collection of 1000 images, subjects were asked to view a random set of 400 images, over two passes, separated by a 10 minute interval. Each image was presented for 5 seconds and followed by a gray mask for 2 seconds, in order to destroy image persistence. The eye-tracker system consists of an infra-red sensing camera, placed alongside the computer monitor, at a distance of about 30 inches from the subject. Images were presented on a 17 inch LCD monitor, with a screen resolution of 96 dpi. Upon 9-point gaze calibration, the eye-tracker is accurate within the nearest 1^o visual angle at 3 feet viewing distance, which translates to an error radius of around 5 pixels on screen. The screen locations

[1] Except for copyrighted IAPS images, which may be obtained upon request from *http://csea.phhp.ufl.edu/media/*. IAPS-image IDs are provided, instead.

Table 1. Image distribution for NUSEF based on semantic category

Semantic Category	**Image Description**	**Image Count**
Face	Single or multiple human/mammal faces.	77
Portrait	Face and body of single human/mammal.	159
Nude		41
Action	Images with a pair of interacting objects (as in *look*, *read* and *shoot*).	60
Affect-variant group	Group of 2-3 images with varying affect.	46
Other concepts	Indoor, outdoor scenes, *world* images comprising living and non-living entities, *reptile*, *injury*.	375

that the subject observes (termed point-of-gaze), are sampled at 30 Hz, and processed to generate the coordinates and duration for every fixation. A fixation point represents the screen location where the point-of-gaze remains within 2^o visual angle for at least 100 milliseconds.

2.2 Image Content

The NUSEF database was compiled from images that were viewed by at least 13 subjects (containing a minimum of 50 fixations). Table 1 presents NUSEF's semantic category-based image distribution, while Table 2 compares our database to MIT's eye-tracking data [8]. Every image was viewed by an average of 25 subjects and over 57% of the images were viewed by more than 20 subjects. Therefore, the database provides statistically rich ground truth for image understanding applications.

Fig.1 shows the fixation patterns for various semantic image categories. Fixations are denoted by circles of varying sizes and gray-levels. The circle sizes

Table 2. Comparison between MIT database [8] and NUSEF in a nutshell

Database	**# images**	**Average # viewers per image**	**Semantics**	**Remarks**
MIT [8]	1003	15	Everyday scenes from *Flickr* and *LabelMe*	Fixations are found around faces, cars and text. Many fixations are biased towards the center.
NUSEF	758	25.3	Expressive face, nude, action, reptile and affect-variant group	Attentional-bias towards salient objects and object-interactions. Fixations are strongly influenced by scene semantics.

are indicative of the fixation duration at the point-of-gaze, while the gray-levels denote fixation starting time during the 5 second image presentation period. Evidently, a majority of the later fixations are around salient objects/regions even if early fixations may be influenced by other factors (image center, brightness, *etc.*). Low-level saliency drives visual attention in contextless indoor and outdoor scenes (Fig.1(a,b)). As also noted in [8], fixations are observed around specific regions like the eyes, nose and mouth for *faces* (Fig.1(c,d,e,f)). For *neutral* and *smiling* faces, attention is distributed almost equally between the upper (eyes) and lower (nose+mouth) halves of the face, while fixations are biased towards the lower half in highly expressive (*angry*, *surprise*, *disgust*) faces ((Fig.1(d)) (fixation statistics in [15]).

Semantic image categories unique to NUSEF include *nudes*, *actions* such as *look*, *read*, *shoot*, and *affect-variant groups*, which comprise a set of 2-3 images with similar content, but with each image inducing a different affect (*e.g.*, pleasant, neutral and unpleasant). Faces attract maximum attention in human and

Fig. 1. Exemplar images from various semantic categories (top) and corresponding gaze patterns (bottom) from NUSEF. Categories include Indoor (a) and Outdoor (b) scenes, *faces*- mammal (c) and human (d), *affect-variant group* (e,f), *action-look* (g) and *read* (h), *portrait*- human (i,j) and mammal (k), *nude* (l), *world* (m,n), *reptile* (o) and *injury* (p). Darker circles denote earlier fixations while whiter circles denote later fixations. Circle sizes denote fixation duration.

mammal *portraits* (Fig.1(i,j,k)), whereas most fixations occur on the body for *nudes* (Fig.1(l)). *Action* images (Fig.1(g,h)) are characterized by frequent fixation transitions between interacting objects, with more transitions occurring from the *action recipient* to the *action source* [15] (*e.g.* Man and book are *action source* and *recipient* respectively in Fig.1(h)). Affect-variant groups allow for a closer analysis of attentional bias, when objects are introduced/deleted in/from the image. The injured/missing eye in Fig.1(e) attracts the most attention, while the fixation distribution is more typical when the missing eye is replaced using image manipulation techniques in Fig.1(f). Fixations are observed around living beings in *world* images Fig.1(l,m), as well as unpleasant concepts such as *reptile* (Fig.1(o)) and *injury* (Fig.1(p)).

2.3 Analysis of Visual Attention Characteristics

Based on the fixation patterns observed for various semantic image categories, we summarize the following about human visual attention characteristics:

1. Human visual attention is undoubtedly influenced by image semantics. Except for contextless indoor and outdoor scenes, fixation clusters are clearly observed around *salient* objects/regions, and we term this phenomenon as attentional-bias. Concepts such as living beings, faces, *etc.* are *salient*, and generally attract considerable visual attention. Also, it appears that attentional-bias is independent of illumination, orientation as well as scale of the *salient* object/concept. This is evident from Fig.1(m,n), where over 90% of the total fixations are observed within 5% of the image area.
2. Scale of the object-of-focus and underlying semantics determine the *salient* image concept(s). Faces are *salient* in *portraits*, and within the face, the eyes, nose and mouth are *salient*. Unpleasant concepts such as reptiles, blood and injury, considerably influence visual attention whenever present. The fact that recognized concepts drive visual attention adds support to the theory that visual attention and object recognition are concurrent processes, and this is an interesting topic of research in the cognitive science community.
3. Visual attention patterns for *action* images are characterized by extensive fixation transitions between interacting objects. This inference is useful for characterizing actions, which otherwise cannot be detected using vision-based approaches. The observed fixation patterns are useful for developing a model to predict interesting regions in unknown images [8], or to localize the spatial locations of *salient* objects and actions [15].
4. Fixations around different 'regions of interest' confirm the exploratory behavior exhibited by the viewers, as they attend to *salient* content. This is particularly useful as human cognition can identify two content-wise dissimilar (due to differing color, texture *etc.*) image regions, as components of the same semantic entity. Overlap of the fixations corresponding to the two regions offer us vital cues, which can be exploited for enhancing automated image understanding. In the next section, we present one such example where the various fixation clusters observed on an object of interest are processed

to generate multiple fixation seeds for active image segmentation. Employing multiple fixation seeds instead of one for active segmentation is found to enhance segmentation performance tremendously.

3 Enhancing Active Image Segmentation with Multiple Fixations

Even as visual attention is specific to *salient* objects, all the fixations on the salient object are generally not restricted to a specific region. Instead, fixations tend to cluster around regions-of-interest within the salient object. If multiple, spatially overlapping, fixation clusters can be discovered from fixation patterns, information from the various clusters can be integrated to infer properties of the entire object. As an exemplar application, we demonstrate how statistically rich fixation data from NUSEF can be utilized to enhance fixation-based active segmentation.

A fixation-based image segmentation scheme, whose objective is to compute the enclosing contour containing the fixation point, has been recently proposed in [17]. Based on the premise that the human eye invariably fixates within the interior of an object, the algorithm attempts to find the set of boundary contours surrounding the fixation. Upon computing the probabilistic boundary edge map to determine the likelihood of an edge pixel being on an actual depth boundary through a combination of monocular, stereo and motion cues, the algorithm proceeds by transforming the edge-map onto polar space, with the fixation point at the pole. The polar space transformation is carried out in order to avoid the problem of graph-cut approaches preferring shorter contours over longer ones, so as to obtain the 'real' boundary contours.

Segmentation, then becomes the problem of finding the optimal cut through the polar edge map, so that edge pixels to the left of the cut are inside the fixation region, while those to the right are outside. An energy function is defined for assigning binary labels '0' and '1' to pixels inside and outside respectively, and the optimal segmentation is obtained as the graph-cut that minimizes the energy function.

3.1 Algorithm Analysis

While the segmentation procedure proposed in [17] is intuitive and the achieved segmentation performance is better than or comparable to other contemporary algorithms [19,20,21], the fixation-based active segmentation scheme suffers from the following shortcomings:

i. The active segmentation algorithm relies on a solitary fixation seed, which it considers to be representative of the foreground object (object-of-interest). This is not true of real fixation data as in general, humans tend to fixate at multiple locations on the object of interest (such as eyes, nose and mouth on a face). Intuitively, segmentation achieved from multiple fixations should

be more accurate and robust compared to the segmentation achieved using a solitary fixation.

ii. While the fixation point is assumed to be any random point within the interior of the object, there is no methodology provided to automatically select the fixation points. Instead, the algorithm requires the user to input the fixation point. Automatic selection of the fixation seed should be trivial with real fixation data, as most fixation points should lie within the *salient* object. At the least, the centroid of the fixation points can be safely assumed to lie within the foreground.

iii. In some cases, using multiple fixation seeds can enable a more accurate segmentation. The authors do not discuss how segments obtained from more than one fixation seed within the same object may be combined to generate the foreground segmentation.

To investigate the hypothesis that multiple fixations available from real eye-fixation data should enhance segmentation performance, and to exploit the

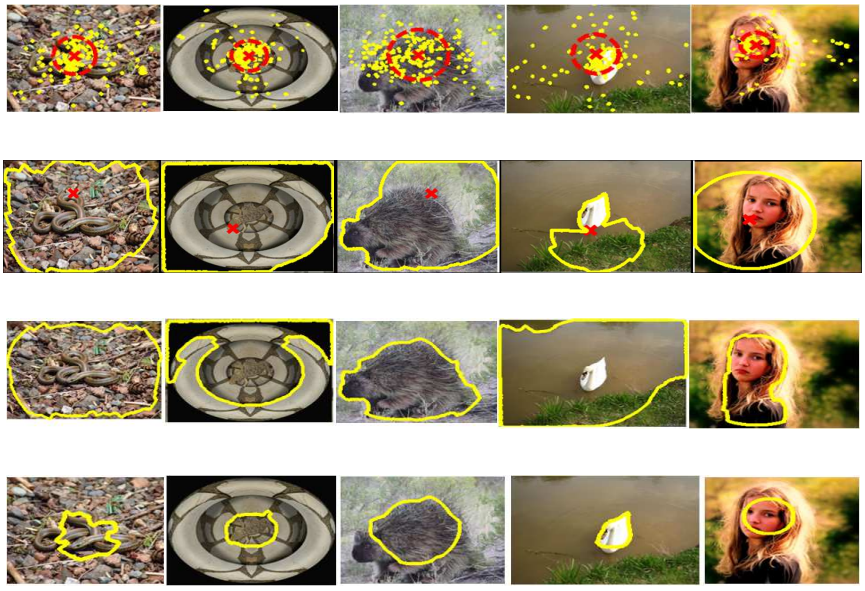
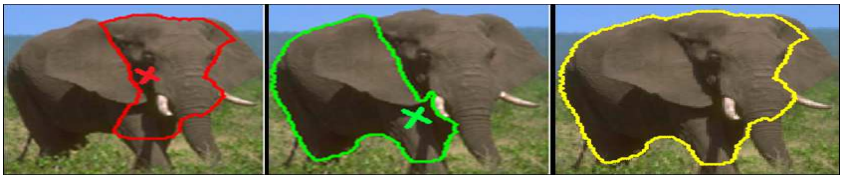


Fig. 2. Enhanced segmentation with multiple fixations. The first row shows the normalized fixation points (yellow). The red 'X' denotes centroid of the fixation cluster around the *salient* object, while the circle represents the mean radius of the cluster. Second row shows segmentation achieved with a random fixation seed inside the object of interest[17]. Third row contains segments obtained upon moving the segmentation seed to the fixation cluster centroid. Incorporating the fixation distribution around the centroid in the energy minimization process can lead to a 'tighter' segmentation of the foreground, as seen in the last row.



(a)

(b)

Fig. 3. More fixation seeds are better than one- Segments from multiple fixation clusters can be combined to achieve more precise segmentation as seen for the (a) *portrait* and (b) *face* images. The final segmentation map (yellow) is computed as the union of intersecting segments. Corresponding fixation patterns can be seen in Fig.1.

fixation clusters around salient objects owing to attentional bias, we performed the following experiments:

(a) To determine whether the segmentation performance of [17] is indeed stable and accurate irrespective of the fixation location, we obtained the output segments for 20 randomly selected fixation seeds from within the hand-drawn segmentation maps for 80 NUSEF images. The baseline segmentation performance is determined as the mean value of the F-measure for the 20 segments obtained from the random seeds. The F-measure, which is used as a measure of the segmentation performance accuracy, is defined as:

$$F = 2PR/(P + R) \tag{1}$$

where P and R denote precision and recall respectively. P denotes the fraction of the segmentation output overlapping with the ground truth, while R represents fraction of the ground-truth overlapping with the output segment.

(b) Considering the set of all fixation points for a given image, a characteristic fixation seed is generated as the centroid of the largest fixation cluster. This allows for the fixation seed to be computed automatically from real fixation data, and since the NUSEF contains statistically rich fixation data, the segmentation output for this characteristic seed, should be more stable than that obtained with a random fixation. Also, as seen from Figs.2 and 3, the centroid of the largest fixation cluster generally lies within the *salient* object, and therefore, the segmentation output with the centroidal fixation seed should be comparable to that obtained in (a). As seen from Fig.2 (rows 2 and 3), using the centroidal seed can sometimes produce a more desirable

segmentation. The largest fixation cluster is computed as follows. In order to account for the fixation duration at every fixated location, each fixation is weighted by the minimum fixation duration in order to generate a corresponding number of 'normalized fixation points' within a Gaussian kernel around the fixation location (this is the inverse of how a fixation is computed). Agglomerative hierarchical clustering is then employed to remove outliers and retain 90% of the original points based on Euclidian distance from the cluster center.

(c) As fewer fixations are observed as we travel radially away from the centroid, the fixation distribution around the centroid can be used as a reliable estimate of the foreground expanse. We recomputed the output segmentation by

Algorithm 1. Pseudo-code for (a), (b), (c), (d)

Steps in (a)
- Using [17], obtain segments for 20 random fixation seeds chosen from within the ground-truth segmentation.
- Compute F as the mean of the F-measures for the 20 segments (using Eq. 1).

Steps in (b)
- (i) for all fixation points fp, compute $weight_{fp} = (fixation\_duration\_at\_fp)/100$ (min fixation duration). Sample $weight_{fp}$ points within a Gaussian kernel around fp to generate normalized fixation points.
- (ii) Employ hierarchical clustering to compute the biggest fixation cluster based on Euclidian distance criterion.
- Use the centroid of this cluster as the fixation seed and invoke [17] to obtain the segmentation output.
- Compute F using Eq. 1.

Steps in (c)
- Perform step (i) to compute the normalized fixation point locations.
- Perform step (ii) to compute the biggest fixation cluster.
- (iii) Compute the centroid and assign r_{mean} as the mean distance of all points from the cluster centroid.
- (iv) Use the centroid of this cluster as the fixation seed for [17].
- (v) for all edge pixels p beyond $2 * r_{mean}$ distance from the fixation centroid, reset the labeling cost as $U_p(l_p = 0) = D$ and $U_p(l_p = 1) = 0$. This initialization discourages segmentation algorithm from labeling pixels outside $2 * r_{mean}$ distance as being 'inside' the fixation region.
- (vi)Perform the energy minimization to obtain the segmentation output.
- Compute F using Eq. 1.

Steps in (d)
- Perform steps (i),(ii) to compute the biggest fixation cluster.
- Compute sub-clusters within this cluster such that minimum cluster size $> D_{min}$ and distance between cluster centers $> D_{min}$, again employing agglomerative clustering.
- Repeat steps (ii), (iii), (iv), (v), and (vi) for all sub-clusters.
- Integrate the segments obtained from the various clusters in the final segmentation map by computing the union of segments having more than 10% overlap.
- Compute F using Eq. 1.

incorporating this information in the energy minimization process. In particular, we re-initialize the labeling cost U(.), so that all edge pixels at a distance greater than r_t from the centroid are deemed to be outside the foreground, *i.e.*, $U_p(l_p = 0) = D$ and $U_p(l_p = 1) = 0\ \forall p$ such that, $r_p \geq r_t$. Setting $r_t = 2r_{mean}$, where r_{mean} is the mean cluster radius from the centroid, works well for most images in practice. Incorporating fixation distribution information in the energy minimization process leads to a 'tighter' and more accurate foreground segmentation for difficult cases where the foreground-background similarity is high (Fig.2, fourth row).

(d) Penalizing the spread of the 'inside' region beyond r_t can at times, force the graph-cut algorithm to limit the foreground boundary at textural edges. In such cases, integrating the segmentation maps obtained from sub-clusters within the main cluster can lead to the optimal segmentation (Fig.3). From the main fixation cluster, we again employ agglomerative clustering to discover all sub-clusters that have a minimum membership (at least 5% of the total fixations) and whose centroids are separated by a minimum distance (100 pixels). The segmentation map for each cluster is computed as in (c), and we compute the final segmentation map as the union of segments that have at least 10% overlap.

The pseudo-code summarizing the steps involved in (a), (b), (c) and (d) is provided in Algorithm 1.

3.2 Results and Discussion

Performance evaluation to evaluate the effect of (a), (b), (c) and (d) was done on 80 NUSEF images, each comprising only one *salient* object. The data essentially corresponded to the following semantic categories- *Face*, *portrait*, *world* and *nude*, and included a number of challenging cases, where the foreground and background are visually similar.

As mentioned previously, the F-measure is used for evaluating segmentation accuracy. For the baseline method, the mean F-measure for the segmentation outputs produced from 20 random seeds was computed, while in all of (b), (c) and (d), a single segmentation output is produced for which the F-measure is computed. The F-measure scores for segmentation procedures (a), (b), (c) and (d) are tabulated in Table 3.

Table 3. Performance evaluation for segmentation outputs from (a), (b), (c) and (d)

Procedure	F-measure (mean ± variance)
(a)	0.6 ± 0.05
(b)	0.59 ± 0.06
(c)	0.60 ± 0.04
(d)	**0.66± 0.04**

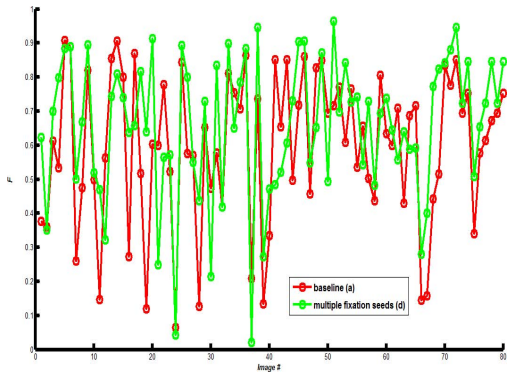


Fig. 4. F measure plot for 80 images. The legend is as follows - *red* baseline and *green* - Integration of segments obtained from multiple sub-clusters.

The F-measure scores for (a), (b) and (c) are found to be almost similar. While the fixation seeds for (a) were randomly picked from the hand-segmented ground truth, the seeds for (b) and (c) were automatically obtained from the fixation data. The fact that the segmentation performance obtained from all three procedures are comparable implies that our methodology for determining the fixation seed is valid. While incorporating the fixation distribution information in the segmentation framework can isolate the foreground more accurately for difficult cases (shown in Fig.2), it also causes the graph-cut algorithm to draw the boundaries along the edges closest to the fixation, sometimes leading to inefficient segmentation. Nevertheless, this deficiency can be overcome by considering overlapping segments obtained from multiple fixation clusters whose centers are sufficiently far away from one another, as in (d).

Fig.4 presents the F-measure plots for segmentation procedures (a) and (d). Clearly, the segmentation performance obtained using multiple fixation seeds is better than that obtained from a random fixation point for most images. This is because segments are conservatively computed in the multi-fixation seed case using the cluster spread as a cue, and then integrated to produce the final segmentation map. However, in some cases where spurious segments are picked up, the segmentation performance using multi-fixation seeds also falls. Overall, a significant 10% improvement in segmentation performance is obtained on using multiple seeds obtained from actual fixation data for segmentation as against a random fixation seed.

4 Conclusion and Future Work

This paper presents NUSEF- an eye fixation database acquired for images corresponding to many semantic categories, including *affective* content, in which visual attention is strongly driven by image semantics. The acquired fixation patterns confirm the hypothesis that eye fixations are influenced by *salient* image

content, and are largely independent of the viewer-specific preferences. We believe that this database would be particularly beneficial for visual attention and image understanding-related research. The fact that viewers show exploratory behavior while observing *salient* content, thereby generating clusters around interesting regions, is then exploited to enhance the segmentation performance achieved by the fixation-based active segmentation algorithm by as much as 10%.

Future work involves formalizing the segmentation procedure, which is currently based on certain heuristics. If fixation data can be efficiently used for object segmentation, it would benefit a number of vision and graphics applications such as content-based image retrieval and seam carving. Characterization of image data based on gaze patterns (*e.g.* action vs non-action images) is another direction for future work.

Acknowledgements

This research has been partially supported by the FP7 IP European project GLOCAL.

References

1. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vision 60(2), 91–110 (2004)
2. Shotton, J., Winn, J., Rother, C., Criminisi, A.: Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. Int. J. Comput. Vision 81(1), 2–23 (2009)
3. Oliva, A., Torralba, A.: Modeling the shape of the scene: A holistic representation of the spatial envelope. Int. J. Comput. Vision 42(3), 145–175 (2001)
4. Van Gemert, J.C., Geusebroek, J.M., Veenman, C.J., Snoek, C.G.M., Smeulders, A.W.M.: Robust scene categorization by learning image statistics in context. In: CVPR-SLAM Workshop (2006)
5. Zheng, Y.T., Neo, S.Y., Chua, T.S., Tian, Q.: Visual synset: a higher-level visual representation for object-based image retrieval. The Visual Computer 25(1), 13–23 (2009)
6. Uijlings, J.R.R., Smeulders, A.W.M., Scha, R.J.H.: Real-time bag of words, approximately. In: CIVR (2009)
7. Spain, M., Perona, P.: Some objects are more equal than others: Measuring and predicting importance. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part I. LNCS, vol. 5302, pp. 523–536. Springer, Heidelberg (2008)
8. Judd, T., Ehinger, K., Durand, F., Torralba, A.: Learning to predict where humans look. In: ICCV (2009)
9. Peters, R.J., Iyer, A., Itti, L., Koch, C.: Components of bottom-up gaze allocation in natural images. Vision Research 45(8), 2397–2416 (2005)
10. Valenti, R., Sebe, N., Gevers, T.: Image saliency by isocentric curvedness and color. In: ICCV (2009)
11. Liu, T., Sun, J., Zheng, N.N., Tang, X., Shum, H.Y.: Learning to detect a salient object. In: CVPR (2007)

12. Zhang, L., Tong, M.H., Marks, T.K., Shan, H., Cottrell, G.W.: SUN: A Bayesian framework for saliency using natural statistics. Journal of Vision 8(7), 1–20 (2008)
13. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. PAMI 20(11), 1254–1259 (1998)
14. Bruce, N., Tsotsos, J.: Saliency, attention, and visual search: An information theoretic approach. J. of Vision 9(3), 1–24 (2009)
15. Subramanian, R., Harish, K., Raymond, H., Chua, T.S., Kankanhalli, M.: Automated localization of affective objects and actions in images via caption text-cum-eye gaze analysis. In: ACM MM, pp. 729–732 (2009)
16. Einhuser, W., Spain, M., Perona, P.: Objects predict fixations better than early saliency. J. Vis. 8(14), 1–26 (2008)
17. Mishra, A., Aloimonos, Y., Fah, C.L.: Active segmentation with fixation. In: ICCV (2009)
18. Lang, P., Bradley, M., Cuthbert, B.: (iaps): Affective ratings of pictures and instruction manual. Technical report, University of Florida (2008)
19. Bagon, S., Boiman, O., Irani, M.: What is a good image segment? A unified approach to segment extraction. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part IV. LNCS, vol. 5305, pp. 30–44. Springer, Heidelberg (2008)
20. Alpert, S., Galun, M., Basri, R., Brandt, A.: Image segmentation by probabilistic bottom-up aggregation and cue integration. In: CVPR (2007)
21. Arbelaez, P., Cohen, L.: Constraine image segmentation from hierarchical boundaries. In: CVPR (2008)