Automated localization of affective objects and actions in images *via* text caption-cum-eye-gaze analysis

Subramanian Ramanathan†, Harish Katti†, Huang Z Raymond‡, Chua Tat-Seng†, Mohan Kankanhalli†
School of Computing†, Department of Psychology‡
National University of Singapore, Singapore
raman,harishk,chuats,mohan@comp.nus.edu.sg,raymondhuang@nus.edu.sg

ABSTRACT

We propose a novel framework to **localize** and **label** affective objects and actions in images through a combination of text, visual and gaze-based analysis. Human gaze provides useful cues to infer locations and interactions of affective objects. While concepts (labels) associated with an image can be determined from its caption, we demonstrate localization of these concepts upon learning from a statistical **affect model for world concepts**. The affect model is derived from non-invasively acquired fixation patterns on labeled images, and guides localization of affective objects (*faces*, *reptiles*) and actions (*look*, *read*) from fixations in unlabeled images. Experimental results obtained on a database of 500 images confirm the effectiveness and promise of the proposed approach.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Multimedia application

Keywords

Automated localization and labeling, text caption-cum-gaze analysis, affect model for world concepts, statistical model.

1. INTRODUCTION

Image understanding remains an unsolved problem, despite the many advances in computer vision. Description of natural images involves automated segmentation and recognition of the various scene objects appearing at multiple scales and orientations, which has inspired *LabelMe* [11]. Difficulty in determining image objects (concepts) from visual content has necessitated image retrieval algorithms [2] to rely on associated keywords and captions for image search.

Noise associated with text-based image retrieval led to the development of Supervised Multiclass labeling (SML) [1], which segments and labels unknown images by applying gained knowledge on the extracted 'bag of features'. However, the algorithm requires extensive training and fails to address the semantic gap. An urn model for object recall is used in [12] to establish the *importance* of some scene objects, even in simple scenes. Also, observations made from eye-gaze statistics in [5] suggest that humans are attentive to *interesting* objects in semantically rich photographs.

Eye gaze measurements have been employed for modeling user attention in a number of applications including visual search for Human-Computer Interaction (HCI) [7] and open signed video analysis [3]. [9] employs low-level image features (contrast, intensity *etc.*) for computing a saliency map to predict human gaze. However, as noted in [5], objects drive attention for semantically rich images, while low-level saliency contributes only indirectly.

The work done in this paper is perhaps most similar to [10], where caption text and image segments are combined to localize the subject of a natural image. On the other hand, we focus on localizing *affective* (attention grabbing, emotion evoking) concepts in images. Contrary to the notion that human subjectivity influences the choice of interesting scene objects, we observe that affective concepts are attention grabbing and consistently fixated upon by a majority of subjects. These concepts may correspond to individual objects or interactions between two objects (actions). An **affect model for world concepts** is derived from fixation patterns for labeled images, relying on the observation that visual attention is drawn towards affective concepts and actions. The affect model encodes world ontology as a tree, whose vertex weights denote concept affectiveness, and helps localize the most affective concepts corresponding to the caption of an unlabeled image.

Fig.1 demonstrates automatic labeling of generic faces using the proposed approach. Labeled images (Figs.1(a),(b)) are used for learning affective image concepts. Subject fixation patterns for these images, where a fixation is defined as attention around a point for a minimum time period (100 msec for our experiments), are shown in Figs. 1(d),(e)). Distinct colors represent fixation patterns for different subjects, numbers denote the sequence of fixations while circle sizes denote the fixation duration around a point. While the training images include labels like *body*, *grass etc*, we observe a majority of fixations on the face, implying that faces are affective. Also, most fixations within the face are observed around the *eyes*, *nose* and *mouth*. Fig. 1(c) is an unlabeled image with known fixation patterns (Fig. 1(f)), and whose caption reads 'A cute cat face'.

The hierarchy of affective concepts for Fig. 1(c) is determined through the affect model as *face* $\rightarrow$ {*nose*+*mouth*, *eyes*}. Using JSEG segmentation [4] as a guide, recursive

ACM MM '09 Beijing, China

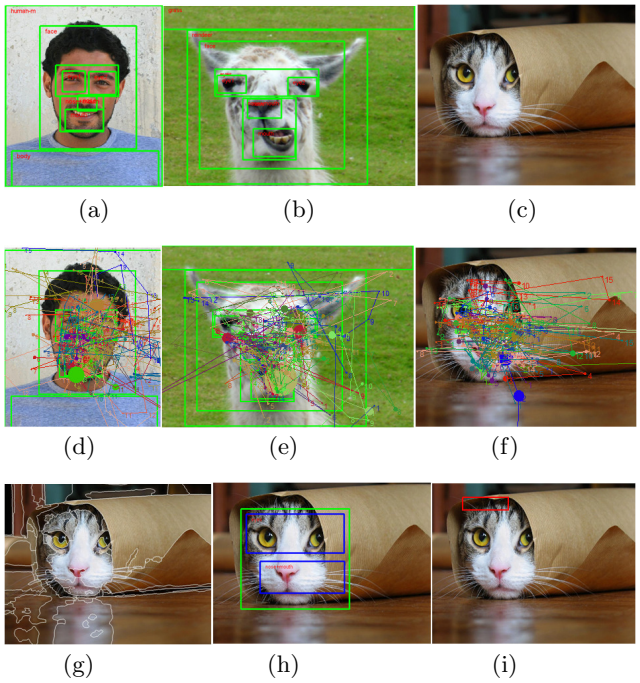


Figure 1: (a) and (b) are labeled images with fixation patterns as shown in (d),(e). (c) Unlabeled image with known fixation patterns (f). (g) JSEG segments and fixation clusters guide automatic localization of *face* (green), *eyes* and *nose+mouth* (blue) regions (h). (i) Viola-Jones face detector result (red).

fixation clustering is employed for affective concept localization (Fig.1(h)), which is not achievable using syntactic approaches (Fig.1(i)). We now describe affect model synthesis and affective concept localization in unlabeled images.

2. AFFECT MODEL SYNTHESIS

2.1 Experimental set-up and protocol

We use the *ASL* eye-tracker for recording subject fixation patterns. The eye-tracker operates at 60 Hz and is accurate within the nearest 0.5^{O} visual angle (0.5 cm error at 50 cm distance from display). Images corresponding to affective themes (*normal face*, *expressive face*, *reptile*, *blood*, *nude etc.*) and actions (*look*, *read*, *shoot*) are chosen from IAPS [8] and *Photo.net* (Fig.2). Also, image manipulation techniques are used to insert/delete affective objects and produce affect-variant image pairs (*e.g. unpleasant*/*neutral*) as shown Fig.2(i).

In two passes, subjects are shown a total of 300 1024x768 resolution images for 5 seconds each with a 2 second gray-mask image in-between. Each pass comprises 70 randomly selected affective (*pleasant*/*unpleasant*) stimuli interspersed with a random number of *neutral* stimuli. Subjects comprised 50 undergraduate and graduate student volunteers, all of whom were allowed a 10 minute break between the two passes to avoid fatigue.

2.2 Affect model synthesis from fixation data

Note from Figs.1(a),(b) that concept labels are assigned to rectangular image regions termed areas of interest (AOIs). Let n AOIs $\{a_1, .., a_n\}$ constitute image I, such that $\bigcup a_i \subseteq I$. AOIs can overlap, and the $\subseteq$ symbol denotes that some image regions may be unlabeled. If m subject fixation patterns are available for I, and $FD_{i,j}$ denotes the duration for which subject j has fixated on a_i, the representative fixation duration for concept $a_i \in I$, is given by

$$\overline{FD_{i\,I}} = \frac{1}{m}\sum_{j=1}^{m} FD_{i,j} \tag{1}$$

Given a concept pair $(a_p, a_q) \in I$, let $TC_{p,q,j}$, $NF_{p,j}$ respectively denote the fixation transition count from a_p to a_q and the number of fixations in a_p for subject j. The representative conditional probability $\overline{CP_{p,q\,I}}$, which models the likelihood of a fixation transition from a_p to a_q following a fixation in a_p is defined as

$$\overline{CP_{p,q\,I}} = \frac{\sum_{j=1}^{m}(TC_{p,q,j})}{\sum_{j=1}^{m}(NF_{p,j})} \tag{2}$$

Empirical observation shows high $\overline{FD_{i\,I}}$ and $\overline{CP_{p,q\,I}}$ values correspond to affective objects and actions respectively.From labeled image AOIs, we construct the affect model as an ontology tree incorporating hierarchical relationships between world concepts (Fig.3). Each concept is associated with an *affect weight*, which measures its affectiveness against other concepts at the same hierarchy level. If P_i is the parent concept for a_i as given by the ontology, then the AOI corresponding to P_i contains a_i in image I. Let S_i denotes the set of N_i images containing a_i, the representative affect weight $\overline{w_i}$ for concept a_i is

$$\overline{w_i} = \frac{1}{N_i}\sum_{\forall I \in S_i} \frac{\overline{FD_{i\,I}}}{\overline{FD_{P_i\,I}}} \tag{3}$$

Strongly affective objects are blue-shaded in Fig.3.

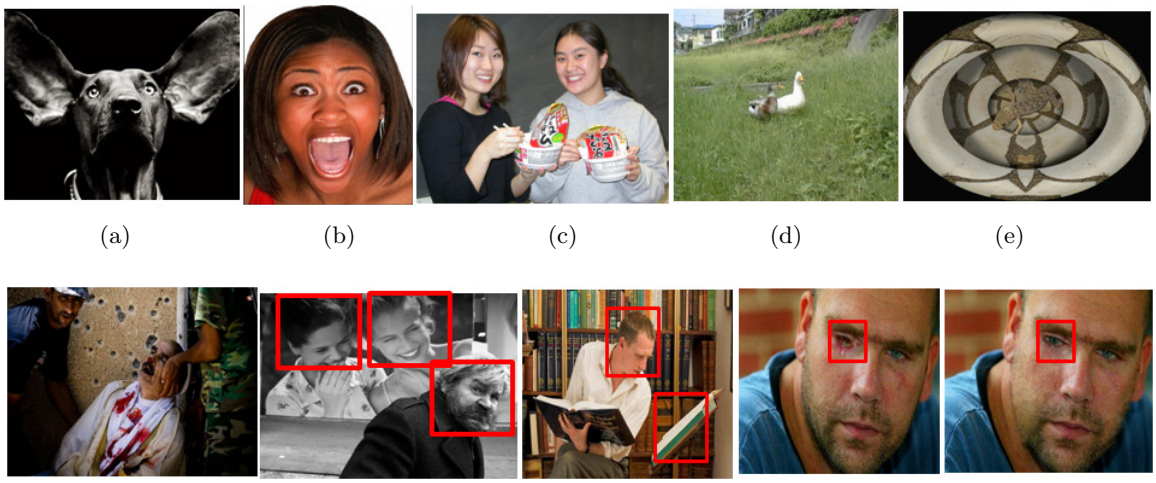
Affective concept learning from statistics is presented in Table 1. We learn the affectiveness of a particular concept from images where it is significant, and also co-occurs with other concepts in the world ontology. *World* images, which represent a collection of *living* and *inanimate* objects, are used to infer that *living beings* are highly affective. *Face* grabs attention in normal *humans*/*mammals*, while the *body* is substantially more affective in *nude* images. Within the *face*, *nose* and *mouth* correspond to a higher w_i, especially for *expressive* faces. Affective actions are characterized by extensive fixation transitions between interacting objects, as represented by dotted arrows in Fig.3. Also, in cases where the action source is clearly identifiable (as in *read*, *shoot*), we observe that the likelihood of transitions from the less affective *action recipient* to the more affective *action source* is higher, which is useful for inferring the direction of action.

Image theme	**#Images**	**Concept- $\overline{w_i}$ (or) $\overline{CP_{p,q\,I}}$**
World	30	*living*- 0.4, *inanimate*- 0.1
human/mammal	50	*face*- 0.75, *body*- 0.19
Nude	20	*face*- 0.22 *body*- 0.62
Normal faces	50	*eyes*- 0.37,*nose+mouth*- 0.4
Expressive faces	48	*eyes*- 0.35,*nose+mouth*- 0.5
Look,Read,Shoot	60	$mean(\overline{CP_{rec,src\,I}})$ – 0.4

Table 1: Statistics for affective concept learning.

3. LOCALIZING AFFECTIVE CONCEPTS IN UNLABELED IMAGES

The proposed framework for localizing and labeling affective objects/actions in unlabeled images consists of the following steps:



(a) (b) (c) (d) (e)

(f) (g) (h) (i)

Figure 2: Exemplar images corresponding to various affective themes. (a) *Normal* and (b) *expressive* face. (c) Multiple *human/mammal* (d) *world* consisting of multiple (living and inanimate) objects (e) *reptile* (f) *blood*. Interacting objects for (g) *look*, (h) *read* actions shown in red. (i) Synthesis of an affect-variant (*unpleasant/neutral*) image pair by restoration of the *damaged* eye (red) through image manipulation techniques.

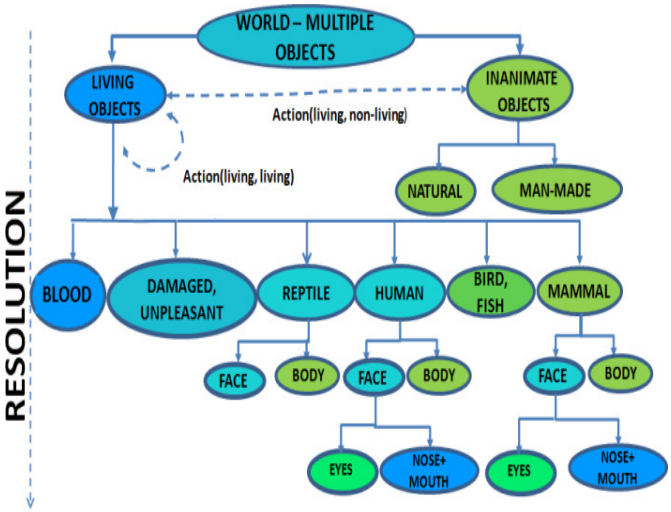
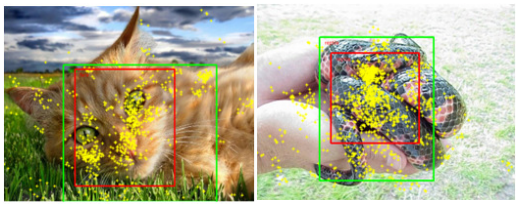


Figure 3: Affect model. A shift from blue to green-shaded ellipses denotes a transition from more affective to less affective concepts. Dotted arrows represent action-characteristic fixation transitions between objects.

- *Determining affective image concepts from caption analysis and affect model*- We assume noise-free and concise captions for unlabeled images, which list the key image objects and actions (Fig.5). The list of noun /verb /adjective image concepts are automatically determined from the caption using the *Lingua::Tagger* package, and mapped to the closest affect model concepts using *Wordnet* [6]. The caption concepts corresponding to the highest w_i values and their hierarchy are determined using the affect model.
- *Concept localization through recursive fixation clustering*- Fixations on the unlabeled image are used to localize AOIs corresponding to the affective caption concepts. In general, n affective concepts correspond to n distinct fixation clusters, which are determined via hierarchical clustering. Color-based JSEG segmentation [4] enables refinement of fixation clusters, which are noisy. Localization accuracy is increased by retaining only those cluster points that correspond to homogeneous color segments (Fig.4). For some concepts like *face*, AOI localization for sub-concepts in the hierarchy is achieved through recursive fixation clustering, where the largest cluster within the original cluster corresponds to the most affective sub-concept.
- *AOI-based post-processing for action localization*- Upon localization of AOIs corresponding to affective objects, actions can be inferred from extensive fixation transitions between interacting objects, as described in Section 2.2.



(a) (b)

Figure 4: Color-homogeneous cluster (red) obtained from original fixation cluster (green) on (a) *cat face* and (b) *reptile*. Fixation points are shown in yellow.

4. RESULTS

Localization of *italicized* objects and actions from textual image captions is demonstrated in Fig.5. Blue rectangles in (Fig.5(a),(b)) correspond to *face* sub-concepts local-

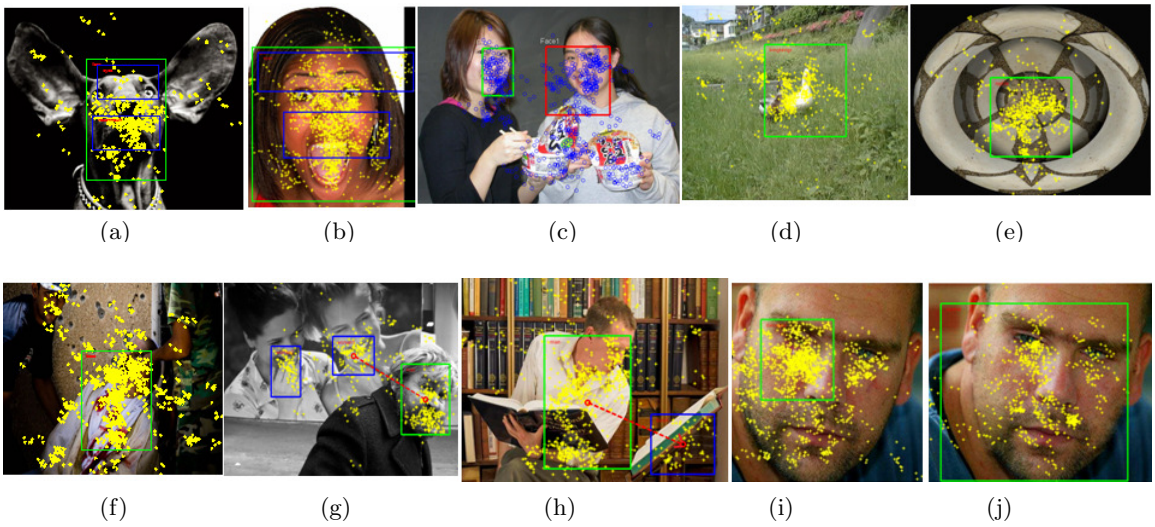


Figure 5: Affective object/action localization results for images with captions (a) A dog's *face* (b) Her *surprised face* said it all! (c) *Two girls* posing for a photo (d) *Birds* in the park. (e) *Lizard* on a plate (f) *Blood-stained* war victim rescued by soldiers (g) *Two ladies looking* and laughing at an *old man* (h) *Man reading* a *book* (i) Man with a *damaged* eye. (h) Fixation patterns and face localization when the damaged eye is restored.

ized through recursive fixation clustering. For action images (Figs.5(g),(h)), the action direction (dotted red arrow) and object labels therefrom, are inferred from the assumption that maximum fixation transitions occur from the *least affective* to the *most affective* object. For the AOIs localized in Fig.5(h), $\overline{CP_{2,1}}_I = 0.351$ and $\overline{CP_{1,2}}_I = 0.071$, which enables assignment of labels to AOI_1, AOI_2 as *man* and *book* respectively. The *look* direction in Fig.5(g) is inferred similarly ($\overline{CP_{p,q}}_I = 0.361$). For a representative set of 50 unlabeled images, correct labeling of affective concepts from image caption text is achieved with 80% accuracy. Localization to a wrong AOI is considered as a failure, the method works best for *face* images. Accuracy of gaze-based label assignment for multiple object and action images can improve tremendously when used along with object recognition algorithms.

5. CONCLUSION AND FUTURE WORK

Localization and labeling of affective caption *objects* and *actions* is successfully achieved using the affect model-based framework. Fixation clusters characterize affective objects while extensive inter-object fixation transitions indicate actions. Affect model-based labeling works best for face images, while all affective concepts in multiple object and action images may not be correctly localized/labeled. Future work involves combining the framework with object recognition algorithms for robust image labeling.

6. ACKNOWLEDGEMENT

We thank Dr. Why Yong Peng for his guidance in designing and conducting behavioural experiments.

7. REFERENCES

[1] G. Carneiro, A. B. Chan, and P. J. Moreno. Supervised learning of semantic classes for image annotation and retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(3):394–410, 2007.

[2] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Comput. Surv.*, 40(2), 2008.

[3] S. J. C. Davies, D. Agrafiotis, C. N. Canagarajah, and D. R. Bull. A multicue bayesian state estimator for gaze prediction in open signed video. *IEEE Trans. on Multimedia*, 11:39–48, 2009.

[4] Y. Deng, B. Manjunath, and H. Shin. Color image segmentation. *CVPR99*, 1999.

[5] W. Einhäuser, M. Spain, and P. Perona. Objects predict fixations better than early saliency. *J. Vis.*, 8(14):1–26, 11 2008.

[6] C. Fellbaum. *WordNet: An Electronical Lexical Database*. MIT Press, 1998.

[7] T. Halverson and A. J. Hornof. The effects of semantic grouping on visual search. In *CHI '08*, pages 3471–3476, 2008.

[8] P. Lang, M. Bradley, and B. Cuthbert. (iaps): Affective ratings of pictures and instruction manual. technical report a-8. Technical report, University of Florida, 2008.

[9] R. J. Peters, A. Iyer, L. Itti, and C. Koch. Components of bottom-up gaze allocation in natural images. *Vision Research*, 45(8):2397–2416, Aug 2005.

[10] N. C. Rowe. Finding and labeling the subject of a captioned depictive natural photograph. *IEEE Trans. on Knowl. and Data Eng.*, 14(1):202–207, 2002.

[11] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. Labelme: A database and web-based tool for image annotation. Technical report, Tech. Rep. MIT-CSAIL-TR-2005-056, 2005.

[12] M. Spain and P. Perona. Some objects are more equal than others: Measuring and predicting importance. In *ECCV '08*, pages 523–536, 2008.