Automated Localization of Affective Objects and Actions in Images Via Caption Text-cum-Eye Gaze Analysis

Subramanian Ramanathan†, Harish Katti†, Raymond Huang Z.W.‡, Tat-Seng Chua†,
Mohan Kankanhalli‡
School of Computing†, Department of Psychology‡
National University of 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, caption text-cum-eye 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.

This paper is perhaps most similar to [10], where caption text and image segments are combined to localize the subject of a natural image. Instead, 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 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. 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. Since eye-gaze is a strong indicator of visual attention, the proposed affect model can be easily extended to include interesting objects in semantically rich images.

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-i-mouth, eyes}. Using JSEG segmentation [4] as a guide, recursive fixation clustering is employed for affective concept localiza-
2. AFFECT MODEL SYNTHESIS

2.1 Experimental model 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 0.5º 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 may 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

\[
FD_{i} = \frac{1}{m} \sum_{j=1}^{m} FD_{i,j}
\]

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 \( CP_{p,q} \), which models the likelihood of a fixation transition from \( a_p \) to \( a_q \) following a fixation in \( a_p \) is defined as

\[
CP_{p,q} = \frac{\sum_{j=1}^{m} TC_{p,q,j}}{\sum_{j=1}^{m} NF_{p,j}}
\]

Empirical observations show that high \( FD_{i} \) and \( CP_{p,q} \) 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, \( T_{O} \) the AOI for \( P_i \), contains \( a_i \) in \( I \). Let \( S_i \) denote the set of \( N_i \) images containing \( a_i \), the representative affect weight \( \bar{w_i} \) for concept \( a_i \) is

\[
\bar{w_i} = \frac{1}{N_i} \sum_{I \in S_i} \frac{FD_{i}}{FD_{i,j}}
\]

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 (ar) to the more affective action source (as) is higher, which is useful for inferring the direction of action.

<table>
<thead>
<tr>
<th>Image theme</th>
<th>#Images</th>
<th>Concept- ( \bar{w_i} ) (or) ( CP_{p,q} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>World</td>
<td>30</td>
<td>living-0.4, mammal-0.1</td>
</tr>
<tr>
<td>Human/mammal</td>
<td>50</td>
<td>face-0.75, body-0.19</td>
</tr>
<tr>
<td>Nude</td>
<td>20</td>
<td>face-0.22 body-0.62</td>
</tr>
<tr>
<td>Normal faces</td>
<td>50</td>
<td>eyes-0.37, nose+mouth-0.4</td>
</tr>
<tr>
<td>Expressive faces</td>
<td>48</td>
<td>eyes-0.35, nose+mouth-0.5</td>
</tr>
<tr>
<td>Look, Read, Shoot</td>
<td>60</td>
<td>mean(( CP_{P_{ar,as}} )) - 0.4</td>
</tr>
</tbody>
</table>

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:

- 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 AOs 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 AOs corresponding to affective objects, actions can be inferred from extensive fixation transitions between interacting objects, as described in Section 2.2.

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 localized 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 AOs localized in Fig.5 (h), $CP_{21j} = 0.351$ and $CP_{12j} = 0.071$, 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 ($CP_{p,q} = 0.492$). For a representative set of 50 unla-
beled images, correct labeling of affective concepts from image caption text is achieved with 80% accuracy and works best for face images. Accuracy is computed based on the number of correctly localised and labeled concepts for the test dataset. While incorrect localization and labeling is observed for images with multiple affective concepts, this can be greatly improved by incorporating object recognition algorithms into the proposed framework.

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 valuable comments and suggestions.

7. REFERENCES


