Can Computers Learn From Humans to *See Better*? Inferring Scene Semantics From Viewers' Eye Movements

Ramanathan Subramanian
Dept. of Information Engineering and Computer Science
University of Trento
subramanian@disi.unitn.it

Victoria Yanulevskaya
Dept. of Information Engineering and Computer Science
University of Trento
yanulevskaya@disi.unitn.it

Nicu Sebe
Dept. of Information Engineering and Computer Science
University of Trento
sebe@disi.unitn.it

ABSTRACT

This paper describes an attempt to bridge the **semantic gap** between computer vision and scene understanding employing eye movements. Even as computer vision algorithms can efficiently detect scene objects, discovering semantic relationships between these objects is as essential for scene understanding. Humans understand complex scenes by rapidly moving their eyes (saccades) to selectively focus on *salient* entities (fixations). For 110 social scenes, we compared verbal descriptions provided by observers against eye movements recorded during a free-viewing task. Data analysis confirms (i) a strong correlation between task-explicit linguistic descriptions and task-implicit eye movements, both of which are influenced by underlying scene semantics and (ii) the ability of eye movements in the form of *fixations* and *saccades* to indicate salient *entities* and *entity relationships* mentioned in scene descriptions.

We demonstrate how eye movements are useful for inferring the meaning of **social** (everyday scenes depicting human activities) and **affective** (emotion-evoking content like *expressive faces*, *nudes*) scenes. While saliency has always been studied through the prism of fixations, we show that saccades are particularly useful for (i) distinguishing mild and high-intensity facial expressions and (ii) discovering interactive actions between scene entities.

Categories and Subject Descriptors

H.1.2 [**User/Machine Systems**]: Human information processing; I.5.4 [**Pattern Recognition Applications**]: Computer vision

General Terms

Algorithms, Measurement, Human Factors

Keywords

eye movements, fixations and saccades, salient entities and interactions, scene semantics

1. INTRODUCTION

The flagship objective of computer vision is to enable computers 'to see what we see', *i.e.*, to endow them with human-like perception so as to extract meaningful descriptions from images [7]. While excellent progress has been achieved in recognizing individual objects [24, 5], analysis of state-of-the-art algorithms in activity and scene understanding [35, 13] indicates the extremely limited use of *context* in scene analysis. This is because the *viewer interprets context* upon discovering salient scene entities and relationships between them. These interpretation rules are based on our understanding of the world and are influenced by many factors which are difficult to model automatically. Therefore, there usually exists a difference between measured scene information and the actual meaning known as **semantic gap**.

Despite the many challenges involved, the need for associating context with content for image description and retrieval applications is strongly advocated in [11]. Also, psychological literature [27] has stressed the importance of object relationships in scene understanding and search. Context is mainly incorporated in multimedia retrieval through user-supplied *metadata*, as typified by ESP game and LabelMe [22], which embellish object-centric information regarding the scene. Tagging images with metadata concerning object relations can be quite tricky though. In everyday scenes comprising many objects, one can come up with multiple descriptions for the each object; the problem becomes intractable when relationships between objects also need to be considered. Nevertheless, studies have shown that human visual attention is limited to *salient* scene objects and their relationships [20, 25], which contribute maximally to scene semantics.

This paper investigates the utility of eye movements as metadata. Eye movements are a reflection of visual attention which is highly contextual- this enables discovery of content central to the scene meaning. Also, eye tracking technology has become inexpensive today, and reliable eye-gaze estimation is achievable using webcams [29]. With time, we believe it should be possible to non-invasively compile large-scale eye gaze data for images browsed on the web.

Humans see by making series of saccades and fixations; saccades are rapid eye movements that enable selective attention, while scene content is assimilated during fixations. Fixations and saccades are indicative of salient *entities* and certain forms of *entity relations*. These entities may be individual objects or object parts (*e.g.*, eyes, nose and mouth within a face). While many works on scene understanding

have exploited fixations for salient object detection, our experiments confirm that saccades are also highly informative.

This work makes two main contributions:

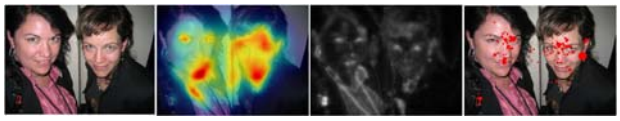
1. This is one of the first works to investigate how eye movements are indicative of human perception in social and affective scenes. Social scenes involve humans performing various actions (talking, walking, reading, *etc.*) in real-life, everyday settings while affective stimuli are capable of arousing emotions (emotive faces, erotica, *etc.*) in viewers. A significant proportion of the image content that we see in newspapers and on photo sharing websites constitute social scenes. Also, the multimedia community has acknowledged the need to analyze affective content for retrieval and understanding [17]. This paper demonstrates why eye movements can be regarded as useful metadata for inferring semantics of such scenes.
2. While previous scene understanding works employing eye movements have focused exclusively on fixations, we demonstrate that saccades are also highly informative. Humans instantly perceive interacting entities in social scenes, and such interactions are characterized by vacillating saccades. Moreover, saccades are found to be more effective than fixations for determining the emotional intensity of faces.

The paper is organized as follows. The next section discusses how eye movements have been utilized for scene understanding. In section 3, we describe in detail (a) the methodology we employed to acquire ground truth for social and affective scenes, (b) key observations we made upon analyzing the annotations, and (c) how eye movements are highly indicative of human scene understanding. We demonstrate how understanding scene semantics can benefit a number of contemporary applications in Section 4 and present our conclusions in Section 5.

2. EYE MOVEMENTS FOR SCENE UNDERSTANDING

Interest in studying and predicting human eye movements began decades ago. Most saliency prediction approaches [10, 28, 33] employ context-independent low-level features such as intensity, color and orientation to determine regions-of-interest in natural images. These 'bottom-up'/early saliency models imply that our visual attention is independent of scene content. However, many studies [34, 9] have confirmed that besides low-level features, our eye movements are driven by a number of 'top-down' factors such as the task on hand and the recognized scene objects. Although some studies [4, 18] argue that interesting image regions are visually salient, early saliency does not correlate well with fixations in meaningful scenes [9, 3, 2] as seen in Fig.1. Recent approaches [12, 36] have achieved higher saliency prediction accuracies upon learning from eye fixations. Even as these efforts have focused on exploiting fixations for understanding scene content, eye movements have rarely been used to infer the scene meaning.

Building on previous work [26, 19] that have examined the use of eye gaze data for scene understanding, we demonstrate how, when combined with scene knowledge (in the form of detected scene objects), fixations and saccades enable semantics inference in *social* and *affective* scenes. In complex



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

Figure 1: *'Bottom-up' saliency does not correlate well with eye fixations in meaningful scenes*- For the original image (a), saliency maps predicted by 'bottom-up' models [10],[28] are shown in (b),(c)- parts of the persons' clothing are discovered as salient, while fixations appear on faces (d).

social scenes, determining relationships between entities is an extremely challenging task, and has been attempted only recently [5]. Nevertheless, humans can instantly recognize the 'central activity' in the scene with little effort. We asked subjects to provide brief descriptions of social scenes that included important scene objects (nouns) and actions (verbs). Upon manually analyzing those descriptions, regularity in the use of verbs was used to determine whether a scene contained an interaction or not. Most subjects consistently identified and reported scene interactions in the initial part of the descriptions. Also, the remaining scene details were not reported as frequently for interactive scenes when compared to non-interactive scenes. In essence, the interacting entities and the interaction mainly constituted the scene gist.

Analysis of the eye-gaze recordings for scenes perceived as *interactive* in the above experiment confirms characteristic saccades vacillating between the interacting entities- these saccades are not observed for scenes without interaction. This is owing to two reasons:

- Social interactions are characterized by pose/gaze cues, and being able to determine (and explore details in) the direction of others' attention is an important ability in humans [32, 14]. Therefore, fixations on one interacting entity are almost always succeeded by a fixation on the other.
- Also, short-term memory influences viewers to re-fixate on semantically interesting/informative areas [9]. This is perhaps responsible for saccades vacillating between interacting entities. We generally observe symmetrically vacillating saccades,*i.e.*, the likelihood of a forward/backward saccadic transition between interacting entities is roughly the same. This allows for reliable and automatic detection of interactions in social scenes employing eye movements.

Furthermore, we investigated how eye data are useful for semantics inference in *affective* images. Of late, there has been considerable interest in analyzing multimedia data at the affective (emotional) level, besides the cognitive (content-specific) level [8, 17]. In affective images, viewers are found to attend to those entities which are *emotionally salient* [26]. We demonstrate how this phenomenon can be exploited to distinguish (a) *highly-intense* from *mild/moderately-intense* facial expressions and (b) portraits of *clothed persons* from *nudes*. We believe that being able to automatically recognize these semantic classes is important because of two reasons:

1. Two people discussing something behind a laptop.
2. Two persons talk with each other.
3. An old man is listening to a guy.
4. Two men discussing.
5. Two adult males having an informal conversation in front of a Macbook.

1. Guy reading a book next to a teddy bear.
2. A man reading a novel while sitting on a sofa and holding a toy bear.
3. A guy reading a book.
4. A man reading a book on the sofa, with a teddy bear in his arm.
5. A man reading a book with a teddy in his hand.

1. A starved woman with a Sleeping dog.
2. A beggar with her dog.
3. A poor woman sitting on the Street, with a dog next to her.
4. Poor woman is sitting near a small dog.
5. A beggar woman on the Street along with her dog.

1. Policeman posing before portrait.
2. Guard in front of the Forbidden City with portrait Of Mao.
3. Mao and a soldier standing below him.
4. A soldier standing in the square in front of Mao.
5. A soldier standing in front of the entrance.

Figure 2: Exemplar interactive (top row) and non-interactive (bottom row) scenes along with their descriptions. All nouns are marked in red, while verbs are marked in green.

- Even as numerous works have focused on emotion recognition, it is still hard to reliably detect facial expressions from images. While we do not know whether eye movements are characteristic of the facial expression, discovering highly/moderately emotional content from a database can facilitate applications such as content publishing (*e.g.*, advertisements).
- Likewise, while content-based (usually employing skin color) nudity/eroticism detection has generated significant attention over the years, state-of-the-art approaches [15] can only achieve moderate detection rates while maintaining a low number of false positives. However, concepts of nudity and eroticism can be fully comprehended only by humans and psychological literature has already reported the significance of eye movements for nudity detection [16].

We asked observers to rate faces and portraits in order to compile ground truth reflecting human perception. Observers were required to rate intensity of emotion portrayed in face images, whereas for portraits, they were required to rate the degree of nudity/eroticism on a Liekert 7-point scale. Based on the ratings, we noted that observers perceived facial expressions to be 'more intense' when significant facial deformations appeared on the lower half of the face, around the nose and mouth (Fig.7). This is mirrored in the eye movements as well- the eyes are usually the most attended regions in faces. However, visual attention shifts to the lower half of the face when emotions are strongly expressed. Our experiments confirm that saccades are more informative than fixations for distinguishing highly expressive from mildly expressive faces. Also, while faces strongly attract visual attention in normal (clothed) portraits, viewers mostly attend to body parts in nude/erotic stimuli. Fixations are found to be slightly more informative than saccades for clothed/nude portrait classification.

Even as our work is closely related to [25, 18], it is important to note a few significant differences. In [25], the authors ask viewers to name 10 objects they see in each image, while we ask for explicit descriptions comprising nouns and verbs- we therefore, seek to acquire a more natural ground-truth for evaluating scene understanding applications. In [18], significant correlation is observed between what viewers consciously perceive as interesting, eye fixations and low level saliency for 100 images spanning four specific semantic categories. We believe that such an analysis needs to be repeated for a wider class of images including social and affective scenes, as visual attention patterns differ significantly for semantically rich stimuli as compared to simple stimuli used in [25, 18]. The following section discusses our ground truth compilation procedure and how eye movements are highly reflective of scene semantics.

3. CORRELATING EYE MOVEMENTS WITH HUMAN PERCEPTION

We selected 110 social scenes and 160 affective scenes (60 faces, 100 portraits) from the MIT [12] and NUSEF [19] eye tracking data sets for compiling ground truth concerning human perception of semantics. These publicly available databases contain eye movement recordings for a substantial number of semantically rich stimuli compiled under *free-viewing* conditions, where viewers are required to examine images in the absence of any high-level tasks that might bias their viewing patterns. A free-viewing paradigm is most representative of task-agnostic human scene understanding, and facilitates objective measurement of the notion of *saliency*.

The **MIT database** [12] contains eye gaze data for 1003 images including indoor and outdoor scenes, scenes with single or multiple humans and animals captured at mid-resolution, and a few high-resolution face images. Eye-gaze patterns of 15 viewers were recorded as they examined each image for 3 seconds, over two sessions scheduled one week apart. The **NUSEF** database [19] consists of at least 13 gaze patterns per image for 758 gray-scale/color images spanning many semantic categories such as high-resolution human and mammal faces, mid-resolution *portraits* showing face and body of humans/mammals, including *nudes*, and interactive

scenes. Eye movements were recorded using a desktop-based eye tracker as viewers observed 350 images for 5 seconds each, over two sessions separated by a 10 minute interval.

3.1 Analyzing linguistic descriptions and eye gaze for social scenes

We asked observers to describe 110 social scenes 'in one or two sentences' for compilation of ground truth reflecting human scene understanding. Viewers were required to include objects (nouns) and object relationships (verbs) that they considered as important in the scene. To minimize noise, we asked viewers to only describe what they saw in the image and not to guess invisible details. All images were described by at least 12 observers; for some images, we collected 20 descriptions. Four exemplar images with five sample descriptions per image are shown in Fig.2.

	Man (1)	Book (1)	Teddy bear (0.75)	Sofa (0.2)
Man (1)		1 (reading)	0.2 (holding)	0.06 (sitting)
Book (1)				
Teddy Bear (0.75)				
Sofa (0.2)				

Figure 3: Interaction matrix for the image on the left containing normalized named frequencies for object and object interactions. Refer to Fig.2 for sample descriptions.

To analyze the descriptions, we manually listed all the objects named by observers, using the most obvious synonym denoting similar words wherever necessary. Then, we computed a normalized *named frequency* for each scene object, which denotes how frequently the object is named in the descriptions (a named frequency of 1 implies that an object was named in all the descriptions). Furthermore, we represented relationships between objects as an $N \times N$ interaction matrix, where N is the number of named objects for the image. In this work, we are only interested in verbs identifying *interactions* between objects *i.e.*, 'A man *talking* to a woman' or 'A man *carrying* a child' are considered interactive, while 'A man *walking*' does not imply any interaction. We focused on those verbs connecting two or more nouns in the descriptions, taking into account plural noun forms as in 'People are talking'. As for the nouns, we also computed the named frequencies of these interactive verbs. Fig.3 shows an image and the corresponding interaction matrix computed from 20 descriptions.

Fig.4 presents the results of ground-truth data analysis. We computed the distribution of the number of named objects per image across all images (Fig.4(a)). On the whole, 484 objects were named for 110 images with a mean of 4.4 ± 1.6 objects per image. This number is much lower than the median object count reported in [25]. This difference is mainly due to the nature of the scene description task involved- in [25], every viewer is asked for a list of 10 objects from which 'important objects' are derived. In an object-centric scene description task, it is highly likely that the viewer starts to look at objects peripheral to the scene meaning. In fact, the authors in [25] employ the 'forgetful urn' model to account for this phenomenon. On the contrary, we ask viewers to perform a more natural scene description task, which results in only the important scene objects being named in the descriptions. This is reflected in the high degree of consistency with which viewers name scene objects (Fig.4(b)). 48% of the objects appear in more than 90% of the descriptions, while only 12% of the objects are named by less than 10% of the viewers. Overall, about 70% of all named objects occur in at least 40% scene descriptions, indicating a high degree of agreement regarding what observers deemed as 'important'.

An equally interesting observation is the consistency with which viewers report object interactions. Fig.4(c) presents the distribution of images according to the degree of interaction between objects, as identified by the viewers. For 45 images, over 80% of the descriptions contained an interaction verb- we assumed these scenes to be interactive. For 37 images, fewer than 30% viewers reported any form of interactions- we assumed these scenes to be non-interactive. Also, when interactions were reported, (i) they mostly represented relationship between two objects; this is probably perhaps because the central activity in most scenes involved only dyadic object interactions. (ii) The interactive verbs appeared in at the beginning of sentences in a vast majority of cases. Table 1 lists the set of interactive verbs most commonly reported by viewers.

We then considered the named frequency of objects in interactive and non-interactive scenes. For scenes involving object interactions, the interacting objects constituted 87% of all named scene objects. For non-interactive scenes, the frequently reported scene objects constituted only 78% of all named objects. The result of a two-sample t-test (assuming equal variances) confirms a significant difference in the proportion of scene content reported by observers in their descriptions (null hypothesis assuming that the proportions are similar is rejected at critical $p = 0.046$). This implies that even though viewers generally succeed in identifying the most important objects, fewer scene details are reported in interactive scenes as compared to non-interactive scenes. In other words, *the interaction along with the interacting entities is perceived as central to the scene meaning and essentially constitute the scene gist.* This observation is reinforced from the analysis of eye movements as well.

Frequently reported verbs	Less frequently reported verbs
feeding	*leaning*
reading	*embracing*
fighting	*wearing*
holding	*waving*
working	*sitting*
pointing	*carrying*

Table 1: More frequently and less frequently reported interactive verbs in descriptions of social scenes.

For these 110 images, we manually marked out rectangles denoting scene regions containing objects with high named frequency. Overall, the bounding boxes around most frequently named objects covered 38.6% of the total image area. We then computed the proportion of total eye fixations occurring in these regions during the first 0.5, 1.5 and 3 seconds of scene viewing- the proportions were found to be 0.83, 0.82 and 0.8 respectively. This implies that objects frequently reported by viewers are also highly attended to during scene viewing. Also, semantically important objects

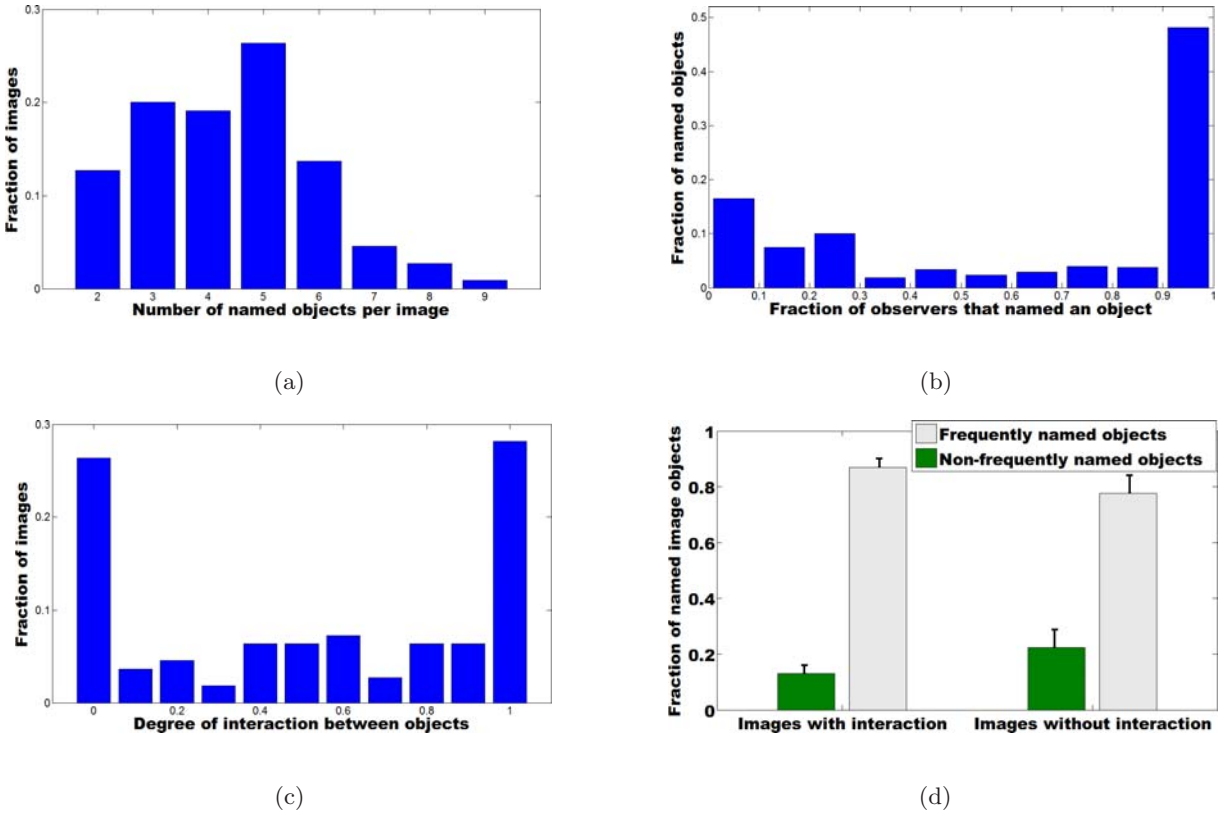


Figure 4: Summary of ground-truth analysis.

are fixated at very early by observers, and with time, observers start exploring the remaining scene details.

The *entropy* is used in [12] to measure how much eye fixations are dispersed over the image. We obtained an averaged continuous saliency map for all images by convolving a Gaussian filter with the eye fixation maps of all viewers, in order to measure the variance in user-fixated locations for the different time intervals (Fig.5). It is evident from the figure that viewers choose to explore more and more scene details with time- the entropy increases from 6.37 to 6.94 as we proceed from 500 ms to 3 sec of scene viewing time. However, across all these time intervals, the difference in entropy observed for interactive scenes and non-interactive scenes is significant at $p < 0.05$ ($p_{max} = 0.03$ for 0.5 sec viewing time). This implies that scene details visually processed by viewers during (task-explicit) scene description and (task-implicit) visual exploration are similar- semantically meaningful entities are **attended to** and **named** consistently.

We term this phenomenon of viewers preferentially directing their visual attention towards salient objects and object relationships as *attentional-bias*. In social scenes, we observe an attentional-bias towards interacting objects and the resulting interaction. If we compare saccades for interactive and non-interactive images, a crucial difference is that interactions (such as *read*, *fight*, *point*, *etc.*) are consistently characterized by vacillating saccades between interacting objects. Psychology literature provides support to this phenomenon- such interactions are characterized by pose cues and humans are instantly able to determine and follow the direction of others' attention [14]. Also, short-term or episodic scene memory causes viewers to semantically re-fixate on interesting objects. This influences the occurrence of vacillating saccades between interacting objects. Now, we briefly discuss how we can model this attentional-bias for automated interaction detection.

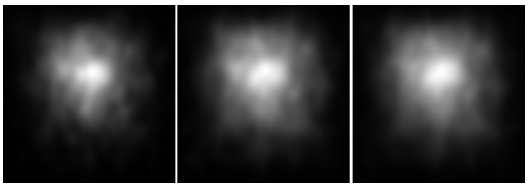
3.1.1 Automated detection of interactions

Let us represent different scene regions using rectangular regions-of-interest or ROIs. The representative conditional probability $P(m|l)$, which models the likelihood of a saccade to ROI m following a fixation in ROI l is defined as

$$P(m|l) = \frac{|S_{l,m}|}{|F_l|} \quad (1)$$

where $|S_{l,m}|$ denotes the number of saccades from ROI l to m, and $|F_l|$ denotes the number of fixations in region l.

In the absence of algorithms that can achieve 'semantic segmentation' (*i.e.*, reliably identify semantically relevant components of objects), and the error involved in gaze estimation, we employ the property of vacillating fixations to identify interacting clusters, and consequently interacting objects. The top row of Fig.6 presents an illustration of how we identify interaction between the man and the book. We employ the mean-shift based, multi-scale fixation clustering approach proposed in [23] to identify fixation clusters at 3 different scales. At any scale, when multiple clusters are detected, we compute the conditional probability of saccading to a different cluster. We choose that scale where the likelihood of to and from saccades are maximally likely,



(a) (b) (c)

Figure 5: Continuous Saliency map computed computed for all social scenes from eye fixations over the initial (a) 0.5 sec (Entropy=6.37) (b) 1.5 sec (Entropy=6.76) and (c) 3 sec of scene viewing (Entropy=6.94).

Figure 6: Illustrative example showing how multi-scale fixation clustering and saccadic analysis at scales S1, S2 and S3 enables interaction detection between man and the book.

characterizing vacillating saccades.

$$S = \arg\max_{s}(\eta_s) \tag{2}$$

where

$$\eta_s = min(\frac{P(m|l)_s}{P(l|m)_s}, \frac{P(l|m)_s}{P(m|l)_s}). \tag{3}$$

To minimize false detections, we consider only clusters with sufficient membership associated with high saccade probability and η values ($P(m|l) \geq 0.3, \eta \geq 0.5$). In the figure, the thickness of the arrows connecting clusters denote the likelihood of a saccade between the two clusters. The strongest and most symmetric interaction occurs between the man and the book at scale $S2$, where $\eta = 0.6$ with $P(book|man) = 0.36$ and $P(man|book) = 0.22$. It is inevitable that some saccades are observed between non-interacting objects as well as viewers explore scene details. However, in most cases, the saccade likelihoods are low (in Fig.6, $P(book|teddy) = 0.2$ and $P(teddy|book) = 0.18$ at scale $S1$) enabling us to discard such cluster-pairs from the analysis.

We attempted saccade-based detection of object interactions at multiple scales to automatically classify the interactive and non-interactive images identified from ground-truth analysis. To determine the effectiveness of saccadic features for discovering the presence or absence of object interactions in social scenes, we employed a leave-pair-out-cross-validation approach (LPOCV). Classification was attempted using linear SVMs and at any time, two samples (one positive and one negative) were used for testing, while the remaining data were used for training. The mean classification accuracy (Acc) and the mean square error (MSE) in accuracies obtained during the different trials are tabulated in Table 2. We obtain an overall accuracy of 77.8%, thereby demonstrating the effectiveness of saccade features for detecting object interactions.

It is also important to note that while vacillating saccades enable discovery of interactions, it is not possible to recognize 'what the interaction is' employing eye movements alone, *i.e.*, we cannot determine whether two persons (discovered employing a person detector) are talking or fighting, employing saccadic cues. However, as contemporary vision algorithms are unable to detect the presence/absence of interactions in images, we believe that detecting interactions, is in itself, a significant step. Further examples of how interactions may be detected in social scenes containing multiple people are shown in the next section.

3.2 Correlating eye movements with viewer ratings for affective scenes

Eye gaze is not only affected by interaction within a scene, but is also strongly influenced by *emotionally salient* content in affective images. We now describe how eye-gaze patterns are useful for distinguishing between (a) high-intensity vs low-intensity facial expressions and (b) portraits of clothed vs nude persons.

3.2.1 Distinguishing highly-expressive from mildly-expressive faces

We asked 12 viewers to rate the intensity of the portrayed facial expressions in 60 images, to understand how humans interpret emotive faces. All images contained an upright and frontal/slightly profile view of a face captured at high resolution. Observers were required to determine whether the portrayed emotion was *positive*, *neutral* or *negative* (termed valence in psychological literature). If they considered the emotional valence to be positive or negative, they were required to rate the intensity of the emotion on a Liekert scale of 1-7. In case an observer perceived a face to be *neutral*, the emotional intensity score was assigned to 0.

At least one observer assigned a non-zero score for every face, implying that viewers perceived some emotion in all the faces. We used the median of the observer scores, M_S, as a threshold to determine whether the face was perceived as mildly or highly expressive- faces with $M_S > 3$ were assumed to be highly emotional. Upon thresholding, 28 faces were found to be highly emotional, while 32 faces belonged to the 'mildly/moderately emotional' category. Faces with extensive deformations around the nose and mouth were perceived to be highly emotional by observers in general. We then studied eye movements on these face images upon automatically determining the ROI rectangles corresponding to the upper and lower face halves employing the Viola-Jones face detector [30] and the neural-network based Rowley eye detector [21]. The top row of Fig.7 presents the valence and median scores for six exemplar faces, along with the automatically determined eye, nose and mouth ROI rectangles.

Upon analyzing eye tracking data for the emotional faces, we made the following observations. Usually, eyes are most salient in the face, and a majority of the fixations appear around the eyes for mildly/moderately expressive faces. However, with increase in the emotional intensity, more and more fixations start appearing in the lower half of the face when significant deformations are observed around the nose and

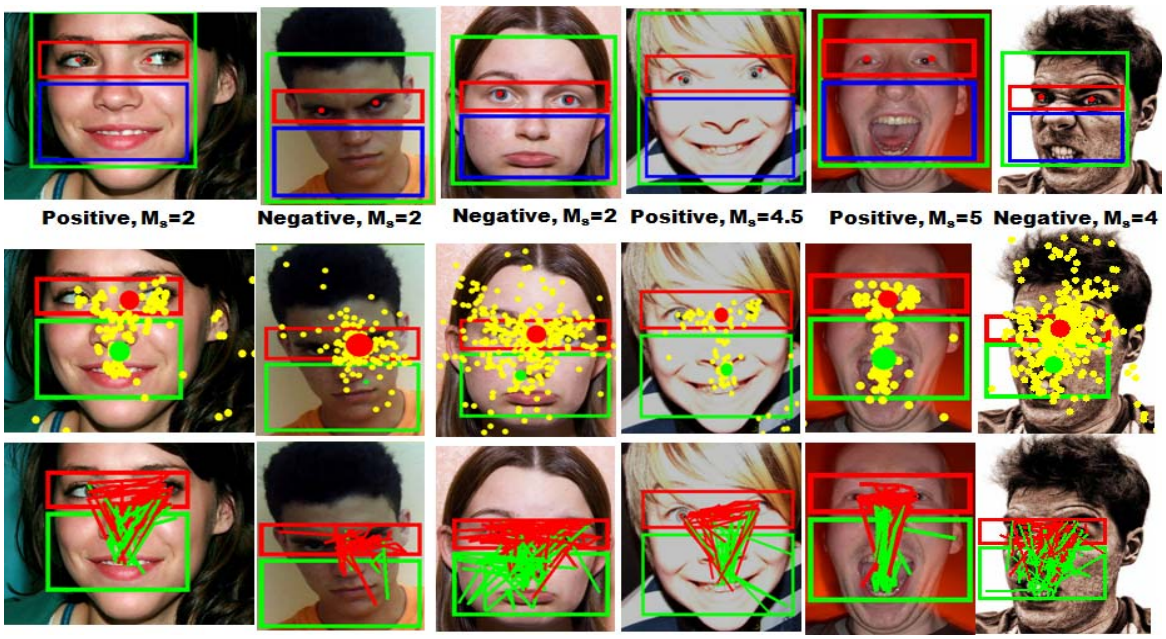


Figure 7: (Top row) Exemplar expressive faces and their median scores. The ROI rectangles for the upper and lower face halves are automatically determined employing face and eye detectors. (Middle row) Distribution of fixations among the two face halves. The red and green circles denote the center of the fixation clusters, and their size denotes cluster membership. Bottom row shows distribution of saccades. More saccades appear on the lower half of the face with increasing emotional intensity.

the mouth (Fig.7, middle row). Earlier works [26] have reported classification of expressive faces upon computing the density of eye fixations in the upper and lower face halves. We obtained better results by also employing saccades in addition to fixations, and built our analysis on the following hypotheses.

- Salient entities are fixated for longer times by viewers. Denoting attentional-bias for the i^{th} entity as B_i, we observe that $B_i \propto |F_i|$, where $|F_i|$ denotes total number of fixations (fixation density) in the i^{th} ROI.
- Viewers' eye movements are exploratory in nature as they attend to interesting scene details. Therefore, for a highly salient entity one can expect a large number of intra-saccades (saccades within the same ROI). Conversely, the likelihood of viewers saccading to a less salient entity from a more salient entity is low. Therefore, $B_i \propto |S_{(i,i)}|$ and $B_i \propto \frac{1}{|S_{(i,j)}|}$, where $|S_{(i,j)}|$ denotes number of saccades from the i^{th} to the j^{th} ROI. $B_i \propto \frac{|S_{(i,i)}|}{|S_{(i,j)}|}$, which we term relative saccadic ratio or $SR_{(i,j)}$.

Table 2 shows the results for classification of expressive faces using (a) only fixation density, (b) using only saccades and (c) using the combination of fixations and saccades. We observe that saccade features in the form of saccadic ratios are more effective for classification compared to using fixation densities, and accuracy increases by 5%. The combination of saccades and fixations adds little to using saccades alone.

	Features	Acc	MSE
Social scenes	Saccades	0.778	0.085
Affective scenes (faces)	Fixations	0.675	0.105
	Saccades	0.720	0.096
	All features	0.728	0.097
Affective scenes (nudes)	Saccades	0.835	0.066
	Fixations	0.861	0.058
	All features	0.854	0.063

Table 2: Automatic classification results for social and affective scenes.

Fig.7 illustrates why saccades are good features for determining the emotional intensity of faces. There are more intra-saccades within eyes regions in mildly emotive faces ($SR_{(eyes,mouth)} > 1$), while in intensely emotional faces, there are more saccades directed towards the nose and mouth regions ($SR_{(mouth,eyes)} > 1$). We believe that this finding is significant as it means that ***apart from the amount of time we spend on observing salient entities (as captured by fixations), the frequency with which we observe them (as given by the number of intra-saccades) is also important.***

3.2.2 *Detection of nude/erotic content*

Another area where eye data can be useful for inferring semantics is with respect to the detection of nude/erotic pictures. State-of-the-art pornography detection approaches [15] are only moderately successful, and we believe eye-movement recordings can serve as a useful complement to

minimize false detections and verify true positives in the automated analysis of erotic content. Our task was designed to validate the hypothesis that 'Humans understand nude/erotic content best', and based on the on the observation that faces strongly attract visual attention in portraits (images showing face and body), while most fixations occur on the body in nudes.

As with faces, we asked 12 observers to assign a score concerning the degree of nudity/eroticism for a total of 100 portraits. Each portrait contained a person in a frontal pose such that the face and part of the body (at least up to the chest portion) are visible to viewers. As with the faces, we computed the median score of all viewer ratings as the representative score for a portrait. All portraits that corresponded to median score M_S greater than 3.5 were considered as nudes. 44 of the 100 portraits were rated as nude/erotic.

Using the human upper body [6] and the Viola Jones [30] face detectors, we automatically determined the face and body ROIs in the portraits. The top row in Fig.8 shows the discovered face and body ROIs along with the median user score for each portrait. As evident from the eye fixation patterns shown in the middle row, visual attention is biased towards the face in normal portraits, and heavily skewed towards the body in nudes. Again, we used both saccades and fixations to automatically predict if an image is a nude or not. Table 2 shows the results. In this case, fixations alone predict nudes better than only saccades or a combination of fixations and saccades. This is probably because viewers spend considerable time attending to salient body parts rather than frequently shifting their focus of attention.

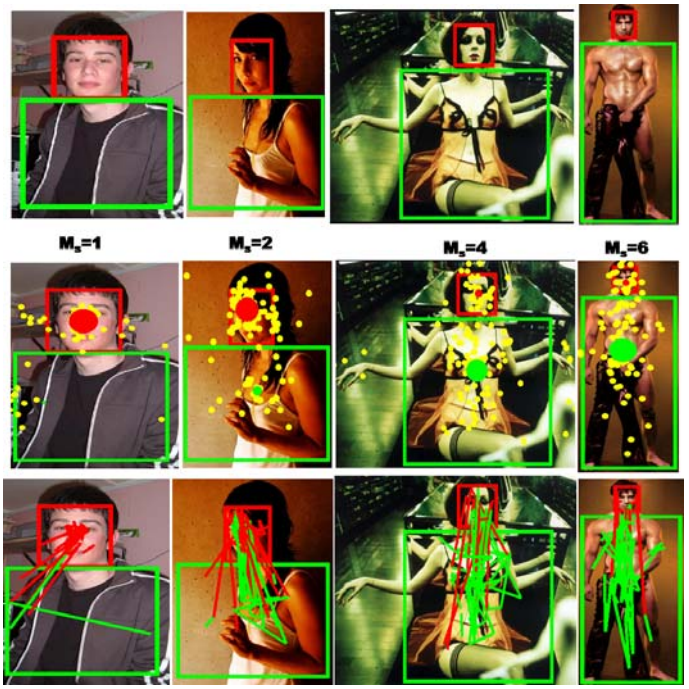


Figure 8: (Top row) Sample portraits with viewer scores for degree of nudity/eroticsm. Outputs of the face and person detectors are used to automatically determine the face and body ROIs. Eye fixation and saccade patterns are shown in the middle and bottom rows respectively. Fixations and saccades are heavily skewed towards the body in nudes.

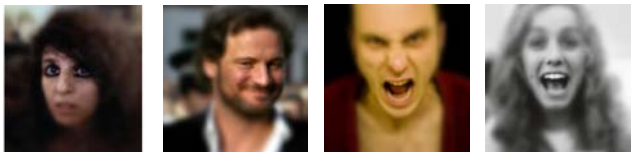


Figure 9: Foveated rendering of faces. The foveation point is placed at the center of the salient ROI.

4. EXPLOITING INFERRED SEMANTICS FOR INTELLIGENT APPLICATIONS

In the previous section, we showed how classification of certain semantically different scene classes can be achieved using eye movements. This, in itself, is highly useful for applications like image retrieval. In this section, we demonstrate a few more examples where knowledge of semantics can enhance the performance of contemporary applications by considering the paradigms of (a) adaptive scene rendering and (b) automatic scene understanding.

4.1 Adaptive scene rendering

Today, there exist ubiquitous multimedia rendering devices, but they have varying capabilities. Different devices have different display resolutions, and not all devices are endowed the processing power to instantaneously render high resolution images. In such cases, it is important to *adaptively* render detail so that the user does not perceive information loss. Techniques such as foveated image rendering [31] and seam carving [1] can be used to preserve regions identified as salient from human fixations. We present how faces can be adaptively rendered so that the user can still perceive can satisfactorily perceive the portrayed facial emotion in Fig.9.

Foveated rendering retains scene details around a foveation point (center of visual attention) while coarsening the rest of the scene. In Fig.9, the foveation point is placed at the center of the *salient* ROI (around the eyes for mildly expressive faces and around the nose and mouth for highly expressive faces). The essence of the facial emotion can be understood in this manner, even if the rendered content is of low resolution. Also, once entities central to the scene meaning can be identified, they can be preserved so that user still understands the scene in re-targeting applications. An example is shown in Fig.10. Once the interacting persons in the scene are identified, they can preserved to perform a more 'semantically correct' image resizing.

4.2 Automated scene understanding- Detection of interacting people

As vacillating saccades enable discovery of interactions in social scenes, combining this information with a person detector allows for identification of interacting people in social scenes. This is a very useful tool to have in tasks like image retrieval. Fig.11 shows two scenes with the same set of objects. However, semantically they are completely different.

We demonstrate some examples for interaction detection in social scenes. We use the state-of-the-art person detector [6] for detecting people in social scenes and also cluster eye fixations at different spacial scales as proposed in section 3.1. Then, we map detected persons to fixation clusters based on

(a) (b) (c)

Figure 10: (a) Fixation map overlaid on original 1024x685 image. (b) Seam-carved 800x600 image preserving salient objects (c) Typical seam-carved output.

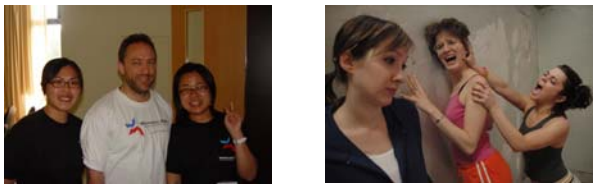


Figure 11: Exemplar images containing the same set of objects (three persons) but with different semantic meanings ('Three people pose for the camera' vs 'Two women are fighting as the third stands on the side').

two rules: (1) The cluster centroid should be located within a bounding box and (2) the cluster should contain a sufficient number of fixations. If vacillating saccades are observed between clusters corresponding to different people in at least one of the spatial scales, then we consider these two persons to be interacting. Fig.12 illustrates several examples, people enclosed in *red bounding boxes* are identified as *engaged in interactions*, while *no interaction* is detected for those labeled using *white bounding boxes*. The yellow dots depict eye-fixations while arrows denote the direction of strongly directed saccades within the scene.

People appearing in images (a-c) are not found to be interacting with others. Interestingly, scenes (a) and (b) involve interaction between people and objects: two boys are holding a sign, and a person is working on the laptop. Images (d-f) contain interacting persons. Overall, this application demonstrates how information from eye movements can be combined with object detectors for semantic interpretation in complex scenes.

5. CONCLUSION

This paper demonstrates how eye movement recordings are useful data for automated scene understanding when combined with content analysis tools (object detectors). Although impressive progress has been recently achieved in object detection and predicting salient scene content, it is currently impossible for purely computational techniques to infer scene semantics such as object interactions from images. Eye movements in the form of fixations and saccades enable discovery of salient entities and entity-relations.

We have shown how eye-movements can be employed for distinguishing interactive and non-interactive social scenes, mild- and high-intensity facial expressions as well as portraits of clothed persons and nudes. With advances in science and technology, it should be possible to non-invasively compile eye movements on a large scale for media browsed on the Internet. In future, we also envision that eye movement recordings will have a critical role to play in automated scene understanding.

6. ACKNOWLEDGMENTS

This work was supported by the Glocal FP7 IP and the S-PATTERNS FIRB projects.

7. REFERENCES

[1] S. Avidan and A. Shamir. Seam carving for content-aware image resizing. *ACM Transactions on Graphics*, 26(3), 2007.

[2] E. Birmingham, W. F. Bischof, and A. Kingstone. Saliency does not account for fixations to eyes within social scenes. *Vision research*, 49(24):2992–3000, 2009.

[3] W. Einhauser, M. Spain, and P. Perona. Objects predict fixations better than early saliency. *Journal of Vision*, 8(14):1–26, 11 2008.

[4] L. Elazary and L. Itti. Interesting objects are visually salient. *Journal of Vision*, 8(3):1–15, 2008.

[5] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2010 (VOC2010) Results.

[6] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 99, 2009.

[7] D. A. Forsyth and J. Ponce. *Computer Vision: A Modern Approach*. Prentice Hall, 2002.

[8] A. Hanjalic and L.-Q. Xu. Affective video content representation and modeling. *IEEE Transactions on Multimedia*, 7(1):143–154, 2005.

[9] J. Henderson. Human gaze control during real-world scene perception. *Trends in Cognitive Sciences*, 7(11):498–504, November 2003.

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

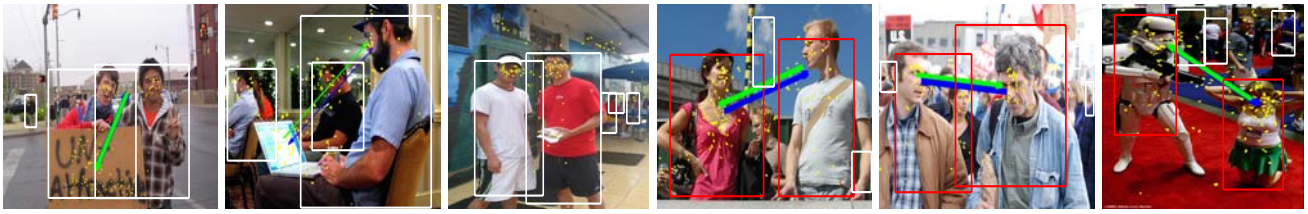
[11] R. Jain and P. Sinha. Content without context is meaningless. In *ACM International Conference on Multimedia*, 2010.

[12] T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. In *International Conference on Computer Vision*, 2009.

[13] A. Kembhavi, T. Yeh, and L. S. Davis. Why did the person cross the road (there)? scene understanding using probabilistic logic models and common sense reasoning. In *European Conference on Computer Vision*, pages II: 693–706, 2010.

[14] G. Kuhn, B. Tatler, and G. Cole. You look where I look! Effect of gaze cues on overt and covert attention in misdirection. *Visual Cognition*, 17(6-7):925–944, 2009.

[15] J.-S. Lee, Y.-M. Kuo, P.-C. Chung, and E.-L. Chen. Naked image detection based on adaptive and extensible skin color model. *Pattern Recognition*, 40:2261–2270, 2007.



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

Figure 12: Discovering interactions in social scenes.

[16] A. D. Lykins, M. Meana, and G. Kambe. Detection of differential viewing patterns to erotic and non-erotic stimuli using eye-tracking methodology. *Archives of sexual behavior*, 35(5):569–575, 2006.

[17] J. Machajdik and A. Hanbury. Affective image classification using features inspired by psychology and art theory. In *ACM International Conference on Multimedia*, pages 83–92, 2010.

[18] C. M. M. Masciocchi, S. Mihalas, D. Parkhurst, and E. Niebur. Everyone knows what is interesting: Salient locations which should be fixated. *Journal of vision*, 9(11), 2009.

[19] S. Ramanathan, H. Katti, N. Sebe, M. Kankanhalli, and T. Chua. An eye fixation database for saliency detection in images. In *European Conference on Computer Vision*, pages IV: 30–43, 2010.

[20] R. A. Rensink, J. K. O'Regan, and J. J. Clark. To See or not to See: The Need for Attention to Perceive Changes in Scenes. *Psychological Science*, 8(5):368–373, 1997.

[21] H. Rowley, S. Baluja, and T. Kanade. Neural network-based face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(1):237–28, 1998.

[22] 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.

[23] A. Santella and D. DeCarlo. Robust clustering of eye movement recordings for quantification of visual interest. In *Proceedings of the symposium on Eye tracking research & applications*, pages 27–34, 2004.

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

[25] M. Spain and P. Perona. Measuring and predicting object importance. *International Journal of Computer Vision*, 91(1), 2011.

[26] R. Subramanian, H. Katti, R. Huang, T.-S. Chua, and M. Kankanhalli. Automated localization of affective objects and actions in images via caption text-cum-eye gaze analysis. In *ACM International Conference on Multimedia*, 2009.

[27] A. Torralba, A. Oliva, M. S. Castelhano, and J. M. Henderson. Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychological review*, 113(4):766–786, 2006.

[28] R. Valenti, N. Sebe, and T. Gevers. Image saliency by isocentric curvedness and color. In *International Conference on Computer Vision*, 2009.

[29] R. Valenti, J. Staiano, N. Sebe, and T. Gevers. Webcam-based visual gaze estimation. In *International Conference on Image Analysis and Processing*, 2009.

[30] P. Viola and M. J. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, May 2004.

[31] Z. Wang, L. Lu, and A. C. Bovik. Foveation scalable video coding with automatic fixation selection. *IEEE Transactions on Image Processing*, 12(2), 2003.

[32] A. Whiten. Evolutionary and developmental origins of the mindreading system. In *Evolution and Development*. Lawrence Erlbaum, 1997.

[33] Y. Yang, M. Song, N. Li, J. Bu, and C. Chen. V: What is the chance of happening: A new way to predict where people look. In *European Conference on Computer Vision*, pages 631–643, 2010.

[34] A. Yarbus. *Eye Movements and Vision*. Plenum Press, 1967.

[35] Z. Zeng and Q. Ji. Knowledge based activity recognition with dynamic bayesian network. In *European Conference on Computer Vision*, pages VI: 532–546, 2010.

[36] Q. Zhao and C. Koch. Learning a saliency map using fixated locations in natural scenes. *Journal of Vision*, 11(3):1–15, 2011.