Human Visual Perception, study and applications to understanding Images and Videos

HARISH KATTI

National University of Singapore
2012

For my parents ...

Acknowledgements

I want to thank my supervisor Prof. Mohan Kankanhalli and co-supervisor Prof. Chua Tat-Seng for their patience and support while I made this eventful journey. I was lucky to not only have learnt the basics of research, but also some valuable life skills from them. My interest in research was nurtured further by my interactions with Profs. Why Yong-Peng, K. R. Ramakrishnan, Nicu Sebe, Zhao Shengdong, Yan Shuicheng and Congyan Lang through collaborative research. Prof. Low Kok Lim's support for our eye-tracking studies was both liberal and unconditional. I want to thank Dr. Ramanathan for the close interaction and fruitful work that became an important part of my thesis. The administrative staff at the School of Computing have been supportive throughout my time as a PhD student and then as a Research Assistant, I take this opportunity to thank Ms Loo Line Fong, Irene, Emily and Agnes in particular for their commitment and responsiveness time and again.

PhD has been a long, sometimes solitary and largely introspective journey. My labmates and friends played a variety of roles ranging from mentors, buddies and critics, at different times. I want to thank my friends Vivek, Shweta, Sanjay, Ankit, Reetesh, Anoop, Avinash, Chiang, Dr. Ravindra, Shanmuga, Karthik and Daljit for the interesting discussions we had. I also crossed paths with some wonderful people like Chandra, Wu Dan and Nivethida and grew as a person because of them.

An overseas PhD comes at the cost of being away from loved ones. I thank my parents Dr. Gururaj , Smt. Jayalaxmi and sister Dr. Spandan for being understanding, tolerant and supportive through my long post-graduate stint through a Masters and now a PhD degree. To my dear wife Yamuna, I am more complete and happy for having found you and am looking forward to seeing more of life and growing older by your side.

On research...

I almost wish I hadn't gone down that rabbit-hole,
and yet,
and yet,
it's rather curious,
you know, this sort of life!
-Alice, "Alice in the Wonderland".

The sole cause of man's unhappiness is that he does not know how to stay quietly in his room.
-Blaise Pascal, "Pensées", 1670

Two kinds of people are never satisfied,
ones who love life,
and ones who love knowledge.
-Maulana Jalaluddin Rumi

On exploring life and making choices, right and wrong...

Two roads diverged in a yellow wood,
And sorry I could not travel both
And be one traveler, long I stood
And looked down one as far as I could
To where it bent in the undergrowth;

Then took the other, as just as fair,
And having perhaps the better claim
Because it was grassy and wanted wear,
Though as for that the passing there
Had worn them really about the same,

And both that morning equally lay
In leaves no step had trodden black.
Oh, I marked the first for another day!
Yet knowing how way leads on to way
I doubted if I should ever come back.

I shall be telling this with a sigh
Somewhere ages and ages hence:
Two roads diverged in a wood, and I,
I took the one less traveled by,
And that has made all the difference. *-Robert Frost*

Abstract

Assessing whether a photograph is interesting, or spotting people in conversation or important objects in an images and videos, are visual tasks that we humans do effortlessly and in a robust manner. In this thesis I first explore and quantify how humans distinguish interesting photos from Flickr in a rapid time span (<100ms) and the visual properties used to make this decision. The role of global colour information in making these decisions is brought to light along with the minimum threshold of time required. Camera related Exchangeable image file format (EXIF) parameters are then used to realize a global scene-wide information based model to identify interesting images across meaningful categories such as indoor and outdoor urban and natural landscapes. My subsequent work focuses on how eye-movements are related to the eventual meaning derived from social and affective (emotion evoking) scenes. Such scenes pose significant challenges due to the abstract nature of visual cues (faces, interaction, affective objects) that influence eye-movements. Behavioural experiments involving eye-tracking are used to establish the consistency of preferential eye-fixations (attentional bias), allocated across different objects in such scenes. This data has been released as the publicly-available eye-fixation NUSEF dataset. Novel statistical measures have been proposed to infer attentional bias across concepts and also to analyse strong/weak relationships between visual elements in an image. The analysis uncovers consistent differences in attentional bias across subtle examples such as expressive/neutral faces and strong/weak relationships between visual elements in a scene. A new online clustering algorithm "binning" has also been developed to infer regions of interest from eye-movements for static and dynamic scenes. Applications of the attentional bias model and binning algorithm to challenging computer vision problems of foreground segmentation and key object detection in images is demonstrated. A human-in-loop interactive application involving dynamic placement of sub-title text in videos has also been explored in this thesis.The thesis also brings forth the influence of human visual perception on *recall*, *precision* and the notion of *interest* in some image and video analysis problems.

Contents

1 Introduction **30**

1.1 Visual media as an artifact of Human experiences . . 30

1.2 Brief overview of work presented in this thesis 31

1.3 The notion of *Goodness* in visual media processing . . 33

1.4 Human in the loop, HVA as versatile ground truth . . . 34

1.5 Human visual attention and eye-gaze 35

1.6 Choice of Eye-gaze to investigate Visual attention . . . 37

1.7 Factors influencing Visual Attention 38

1.8 The role of Visual Saliency 42

1.9 Semantic gap in visual media processing 43

1.10 Organization of the Thesis 45

1.11 Contributions 47

2 Related Work **49**

2.1 Human Visual Perception and Visual Attention 49

2.2 Eye-gaze as an artifact of Human Visual Attention . . 49

2.3 Image Understanding 52

2.4 Understanding video content 56

2.5 Eye-gaze as a modality in HCI 57

3 Experimental protocols and Data pre-processing **59**

3.1 Experiment design for pre-attentive interestingness discrimination 59

3.1.1 Data collection 60

3.2 Experiment design for Image based eye-tracking experiments 63
3.2.1 Data collection and preparation 64
3.2.2 Participants 66
3.2.3 Experiment design 66
3.2.4 Apparatus 66
3.2.5 Image content 67
3.3 Experimental procedure for video based eye-tracking experiments 71
3.4 Summary 73

4 Developing the framework 74

4.1 Pre-attentive discrimination of interestingness in the absence of attention 76
4.1.1 Effectiveness of noise masks in destroying image persistence 79
4.2 Eye-gaze, an artifact of Human Visual Attention(HVA) 84
4.2.1 Description of eye-gaze based measures and discovering Attentional bias 87
4.2.2 Bias weight 92
4.2.3 Attentional bias model synthesis from fixation data 94
4.2.4 A basic measure for interaction in Image concepts 96
4.3 Estimating Regions of Interest in Images using the ‘Binning’ algorithm 97
4.3.1 Performance analysis of the binning method . . 102
4.3.2 Evaluation of the binning with a popular baseline method 108

4.3.3 Extending the binning algorithm to infer Interaction represented in static images 112

4.4 Modeling attentional bias for videos 115

4.4.1 video-binning : Discovering ROIs and propagating to future frames 116

4.5 Summary 125

5 Applications to Image and Video understanding 126

5.1 Automatically predicting pre-attentive image interestingness 126

5.2 Applications of Attentional bias to image classification 130

5.3 Application to localization of key concepts images . . . 132

5.3.1 Steps followed 133

5.3.2 Results 134

5.4 Applications of interaction discovery to image classification 136

5.5 Application of ROIs discovered using the binning method, Image segmentation 136

5.6 Application of ROIs discovered using the binning method, guiding object detectors 144

5.6.1 Using eye-gaze based ROIs 146

5.6.2 ROI size estimation to reduce false positives and low relevance detections 147

5.6.3 Multiple ROIs as parts of an object 149

5.6.4 Experimental results and Discussion 149

5.7 Applying *video binning* to *Interactive* and *online* dynamic dialogue localisation onto video frames 151

5.7.1 Data collection 154

5.7.2 Experiment design 154

5.7.3 Evaluation of user attention in captioning . . . 156

5.7.4 Results and discussion 157

5.7.5 The online framework 158

5.7.6 Lessons from dynamic captioning 158

5.7.7 Effect of captioning on eye movements 160

5.7.8 Influence of habituation on subject experience . 161

6 Discussion and Future work 163

6.1 Discussion of important results 163

6.2 Future work 165

Bibliography167

List of Figures

1 Panel illustrates the distribution of rod and cone cells in the retinal wall of the human eye. The highest acuity is in the central region (fovea centralis) with maximum concentration of Cone cells. The blind spot corresponds to the region devoid of rods or cones, here the optical nerve bundle emerges from the eye. http://www. uxmatters.com/ mt/archives/2010/07/ updating-our-understanding-of-perception-and-cognition-part-i.php. 36

2 A standard visual acuity chart used to check for reading tests. Humans can distinguish characters at a much smaller size at the center than at the periphery. http://people.usd.edu/ schieber/ coglab/IntroPeripheral.html 37

3 Comparing attributes of different non-conventional information sources. Representative values for important attributes of each modality are obtained from [103] (EEG), [70] (Eye-Gaze), [99][76] (Face detction expression analysis) and [98] (Electrophysiological signals). . 38

4 Additional attributes of different non-conventional information sources, continued from Fig. 3. Representative values for important attributes of each modality are obtained from [103] (EEG), [70] (Eye-Gaze), [99][76] (Face detction expression analysis) and [98] (Electrophysiological signals). 39

5 Different factors that can affect human visual attention and hence, subsequent understanding of visual content. 40

6 Some results from Yarbus's seminal work [53]. Subject gaze patterns from 3 minute recordings, under different tasks posed prior to viewing the painting "An unexpected visitor" by I.E. Repin. The original painting is shown in the top left panel. The different tasks posed are as follows, (1) Free examination with no prior task (2) A moderately abstract reasoning task, to gauge the economic situation of the family (3) To find the ages of family members (4) Another abstract task, to find the activity that the family was involved in prior to arrival of the visitor (5) To remember the clothes worn by people (6) To remember positions taken by people in the room (7) A more abstract task, to infer how long the visitor had been away from the family. 41

7 The semantic gap can show up in more than one way. The *Intent* of an *Expert* or *Naive* content creator can get lost or altered either during encoding into visual content, or in conversion between media types during the *(encode,store,consume)* cycle. Effects of the Semantic gap are more pronounced in situations where *Naive* users generate and consume visual media. . . . 45

8 The schema represents information flow hierarchy and chapter organization in the thesis. The top layer lists different input modalities that are then analysed in the middle layer to extract features and semantics related information. 46

9 The figure highlights the scope of this chapter in the overall schema for the thesis, input data is captured via Image/Video content, eye-tracking, manual annotation. 59

10 Illustration of image manipulation results, (a) intact image, (b) scrambling to destroy global order, (c) removal of color information (d) blurring to remove local properties, the color is removed as well as it can contains information about global structure in the image. . . . 62

11 The short time-span image presentation protocol for aesthetics discrimination is visualized here, an image pair relevant to the concept *apple* is presented one after another in random order. The presentation time for each image in the pair is same and chosen between 50 to 1000 milliseconds. Images are alternated with noise masks to destroy persistence, a forced choice input records which of the rapidly presented images was perceived as more aesthetic by the user. 62

12 The long time-span image presentation protocol for aesthetics discrimination is visualized here, an image pair relevant to the concept *apple* is presented side-by-side. The stimulus is presented as long as the user needs time to decide whether an image clearly has more aesthetic value than the other. 63

13 Exemplar images corresponding to various semantic categories. (a) Outdoor scene (b) Indoor scene (c) Face (d) *World* image comprising living beings and inanimate objects (e) Reptile (f) Nude (g) Multiple human (h) Blood (i) Image depicting *read* action. (j) and (k) are examples of an image-pair synthesized using image manipulation techniques. The damaged/injured left eye in (j) is restored in (k). 65

14 Experimental set-up overview. (a) Results of 9 point gaze calibration, where the ellipses with the green squares represent regions of uncertainty in gaze computation over different areas on the screen. (b) An experiment in progress. (c) Fixations patterns obtained upon gaze data processing. 67

15 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. 70

16 Illustration of the interactive eye-tracking setup for video. (a) Experiment in progress (b) The subject looks at visual input. (c) The on-screen location being attended to. (d) An off-the-shelf camera is used to establish a mapping between images of the subject's eye while viewing the video. 72

17 The schema visualizes the overall organization of the thesis and highlights the components described in this chapter. The current chapter deals with analysis and modeling of visual content, eye-gaze information and meta-data. 75

18 The panel illustrates how the arrangemnt of different visual elements in images can give rise to rich and abstract semantics. Beginning from simple texture in (a), the meaning of an image can be dominated by low level cues like color and depth in (b), shape and symmetry in (c) and (d). The unusual interaction of *cat* and *book* gives rise to an element of surprise and rich human interaction and emotions are conveyed through inanimate *paper-clips*. 75

19 Image on the left is a relevant result for the query concept *apple*. The image on the right illustrates an image for the same concept that has been viewed preferentially in the Flickr database. 77

20 A visualization of the lack of correlation between image ordering based on mere semantic relevance of tags Vs *interestingness* in the Flickr system. Semantic relevance based ordering of 2132 images is plotted against their ranks on *interestingness*. This illustrates the need for methods that can harness human interaction information. 79

21 Impact of noise masks in reducing the effect of persistence of visual stimulus. The two plots are agreement between user-decisions made for short term and long term image pair presentation. It can be seen that image persistence in the absence of the noise mask significantly increases the overall discrimination capability of the user. 80

22 Improvement in user discrimination as short-term presentation span is varied from 50 milliseconds to 1000 milliseconds. As expected, users make more reliable choices amongst the image pairs presented. A presentation time of about 500 milliseconds appears to be the minimum threshold for reliable decisions by the human observer and can be used as a threshold for display rate for rapid discrimination of interestingness. 81

23 Improvement in user discrimination as short-term presentation span is varied from 50 milliseconds to 200 milliseconds. A binomial statistical significance test reveals agreements between short and long term decisions starting from 50 millisecond short term decisions. 82

24 The panel illustrates changes in pre-attentive discrimination of image interestingness as image content is selectively manipulated. Removing color channel information results in a 20 % drop in discrimination capability. A drop of about 15 % in short-term to long-term agreement when global information is destroyed by scrambling image blocks 83

25 Agreement of short-term decisions made at 100 millisecond presentation with long-term decisions. Though loss of colour information or loss of global order in the image result in a similar drop of about 7% in agreement, removal of local information reduces agreement significantly by more than 20%, this is surprising as literature suggests a dominant role of global information in pre-attentive time spans. 84

26 Different parameters extracted from eye-fixations corresponding to an image in the NUSEF [71] dataset. Images were shown to human subjects for 5 seconds. (a) Fixation sequence numbers, each subject is color-coded with a different color, fixations can be seen to converge quickly to the key concepts (*eye*,*nose+mouth*) (b) Each gray-scale disc represents the fixation duration corresponding to each fixated location, gray-scale value represents fixation start time with a black disc representing 0 second start time and completely white disc representing 5 second fixation start time.(c) Normalized saccade velocities are visualized as thickness of line segments connecting successive fixation locations. Gray-scale value codes for fixation start time. . . 86

27 Visualization of manual annotation of key concepts and their sub-parts for the NUSEF [71] dataset. The annotators additionally label any clearly visible sub-parts of the key-concepts. That way a labeled *face* would also have *eye* and *mouth* regions labeled, if they were clearly visible. This can be seen in Figure 27 (a)(d)(e)(f), where as the annotators omitted *eye* and *mouth* labels for (b) and (c). 88

28 A visualization of some well-supported meronyms relationships in the NUSEF [71] dataset. Manually annotated pairs of (bounding-box, semantic label) are analysed for *part-of* relationships, also described in Eqn. 4. 89

29 Automatically extracted ROIs for (a) *normal* and (b) *expressive* face, (c) *portrait* and (d) *nude* are shown in the first row. Bottom row (e-h) show fixation distribution among the automatically obtained ROIs. 90

30 The figure visualizes how the total fixation time over a concept D_i can be explained in terms of time spent on individual, non overlapping sub-parts. The final ratios are derived from combined fixations from all viewers over objects and sub-parts in an image. 91

31 Panel (b) visualizes fixation transitions between important concepts in the image (a). The transitions are also color coded with gray scale values representing fixation onset time, black represents early onset and white represents fixation onset much later in a 5 second presentation time. Visualized data represents eye-gaze recordings from 22 subjects and is part of the NUSEF dataset [71].(c) Red circles illustrate the well supported regions of interest, green dotted arrows show the dominant, pair-wise $P(m/l)_I$ and $P(l/m)_I$ values between concepts *m* and *l*, thickness of the arrows is proportional to the probability values values. 92

32 Attentional bias model. A shift from blue to green-shaded ellipses denotes a shift from preferentially attended to concepts having high w_i values, to those less fixated upon and have lower w_i. Dotted arrows represent action-characteristic fixation transitions between objects. The vertical axis represents decreasing object size due to the object-part ontology and is marked by *Resolution*. 95

33 Panel (a) visualizes fixation transitions between important concepts in the image, transitions are color coded with gray scale values representing fixation onset time, black represents early onset and white represents fixation onset much later in a 5 second presentation time. Visualized data represents eye-gaze recordings from 22 subjects and is part of the NUSEF dataset [71].(b) Red circles in the cartoon illustrate the well supported regions of interest, green dotted arrows show the dominant, pair-wise $P(m/l)_I$ and $P(l/m)_I$ values between concepts *m* and *l*, thickness of the arrows is proportional to the probability values values. (c) Visualization of normalized $Int_{(l,m)I}$ values depicting the dominant interactions in the given image, a single green arrow marks the direction and magnitude of inferred interaction. 97

34 *Action* vs multiple non-interacting entities. Exemplar images from the *read* and *look* semantic categories are shown in (a),(b). (c),(d) are examples of images containing multiple non-interacting entities. In (e)-(h), the green arrows denote fixation transitions between the different clusters. The thickness of the arrows are indicative of the fixation transition probabilities between two given ROIs. 98

35 The *binning* algorithm. Panels in the top row show a representative image (top-left) and eye-fixation information visualized as described earlier in 26, followed by abstraction of the key visual elements in the image. The middle row illustrates how inter-fixation saccades can be between the same ROI (red arrow) or distinct ROIs (green arrow). The bottom row illustrates how isolating inter-ROI saccades enables grouping of fixation points potentially belonging to the same ROI into one cluster. The right panel in the bottom row is an output from the *binning* algorithm for the chosen image, ROIs clusters are depicted using red polygons and the cluster centroid is illustrated with a blue disc of radius proportional to the cluster support. Yellow dots are eye-fixation information that is input to the algorithm. . . . 99

36 Panels illustrate how ROIs identified by the the *binning* method correspond to visual elements that might be at the level of objects, gestalt elements or abstract concepts. (a) ROIs correspond to the faces involved in the conversation and the apple logo on the laptop.(b) Key elements in the image *solitary mountain* and the two *vanishing points* one on the left where the road curves around and another where the river vanishes into the valley. Vanishing points are strong perceptual cues. (c) Junctions of the bridge and columns are fixated upon selectively by users and are captured well in the discovered ROIs. 102

37 Visualization of eye-gaze based ROIs obtained from binning and the corresponding manually annotated ground truth for evaulation (a) Original image (b) Eye-gaze based ROIs (c) Manually annotated ground truth for the corresponding Image. 5 annotators were given randomly chosen images from the NUSEF dataset [71], the annotators assign white to foreground regions and the black to the background. 104

38 Visualization manually annotated ground truth for randomly chosen images from the NUSEF dataset [71]. The images can have one or more ROIs. 104

39 Performance of the binning method for 50 randomly chosen images from the NUSEF dataset. The binning method employs a conservative strategy to select fixation points into bins, large proportion of fixation points fall within object boundary. This results in higher precision values as compared to recall. An f-measure of 38.5% is achieved in this case. 105

40 Performance of the binning method as the number of subjects viewing an image is increased from 1 to 30. The neighbourhood value is chosen to be 130 pixels to discriminate between intra-object saccades and inter-object saccades. The precision, recall and consequently f-measure are approximately even at 20 subjects. . . . 106

41 Panels illustrate precision, recall and fmeasure of the binning method for 1 subject with (a) eye-gaze information alone, and (b) when eye-gaze ROI information are grown using *active segmentation* [60]. A simple fusion of segmentation based cues with eye-gaze ROIs gives an improvement of over 230% in f-measure as shown underlined with the dotted red lines above the graphs. 107

42 Small *neighborhood* values result in the formation of very small clusters and few of those have sufficient membership to be considered as an ROI. The clusters are well within the object boundary resulting in high precision > 70% and low < 30% recall. The cross over point for $neighbourhood = 80$ is due to a combination of factors including stimulus viewing distance, natural statistics of the images and typical eye-movement behavior. Larger neighborhood values result in large, coarse ROIs which can be bigger than the object and include noisy outliers. This causes reduction in precision as well as that in recall. 109

43 The binning method (a) orders existing bins according to distances to their centroids from s_j and then finds the bin containing a gaze point very close to s_j. On the other hand, the mean-shift based method in [77] replaces the new point s_j with the weighted mean of all points in the specified *neighbourhood*. 112

44 A comparison of precision, recall and fmeasure variations between (a) the meanshift based method in [77] and the binning method presented in this thesis. The behavior of both methods for smaller neighborhood values is similar. It changes significantly for larger neighborhood values, where ROI sizes are preserved in the binning method and result in preservation of precision scores. On the other hand, recall values fall in the binning method as compared to [77]. 113

45 Clusters obtained with varying neighborhoods over image with weak interactions 114

46 Clusters obtained with varying neighborhoods over image with strong interactions 114

47 (a),(b) and (c) Illustrate shift in *HVA* shown by the *red dot*, as the prominent speaker changes in a video sequence. (d),(e) and (f) show the same in a different video sequence. An interesting event is depicted in (g),(h),(i), where the *HVA* shifts from prominent speaker in (h) to the *talking-puppet* in (i), which is actually more *meaningful* and *compelling* in the scene. 117

48 The graph illustrates spatio-temporal eye-gaze data, it indicates good agreement of human eye-movements over 3 different subjects while viewing a video clip, clip height and width form two axes and a third one is formed by the video frame display time. Eye fixations are aligned according to the onset time and each subject is depicted using distinct colors. The colored blobs depict the eye-fixation duration on a ROI in the video stimulus. 118

49 Good agreement of human eye-movements over successive views of a video clip, clip height and width form two axes and a third one is formed by the video frame display time. Eye fixations are aligned according to the onset time and each viewing session is depicted using distinct colors. The colored blobs depict the eye-fixation duration on a ROI in the video stimulus. 118

50 Panels in the top row illustrate important stages in the interactive framework. (a) A frame from the video stream. (b) Eye-gaze based ROIs discovered using the *video binning* method [46]. The red circle shows current location being attended and yellow circles show past ROIs, the arrows show dominant eye movement trajectories. (c) An example image region overlayed with motion saliency computed using motion vector information in the encoded video stream. (d) Face saliency map constructed by detecting and tracking frontal and side profile faces. Panels in the middle row visualize stages in the dialogue captioning framework. (e) Regions likely to contain faces are combined with ROI and likely eye movement paths shown in (f) to compute likely concepts of human interest. Video frame taken from the movie *swades*©UTV Motion Pictures. 125

51 The figure highlights the components described in this chapter. The current chapter deals with image and video understanding applications using the framework developed in chapter 4. 126

52 Normalized frequency of occurrence of different EXIF attributes in our database. Important EXIF attributes that encode global image information directly or indirectly are highlighted(boxed) in red. 128

53 Appropriate subsets of the dataset can be chosen as positive and negative samples to trian individual preferences and community preferences. 128

54 Color-homogeneous cluster (red) obtained from original fixation cluster (green) on (a) *cat face* and (b) *reptile*. Fixation points are shown in yellow. 134

55 Affective object/action localization results for images with captions (a) A dog's *face* $.aoi's : eyes, nose{+}mouth, face$ (b) Her *surprised face* said it all! $aoi's : eyes, nose + mouth, face$(c) *Two girls* posing for a photo. $aoi's : face1, face2)(d)$ *Birds* in the park. $aoi's : bird$ (e) *Lizard* on a plate. $aoi's : reptile$ (f) *Blood-stained* war victim rescued by soldiers. aoi's:blood (g) *Two ladies looking* and laughing at an *old man*. $aoi's : face_1,\ face_2,\ face_3$ (h) *Man reading* a *book*. $aoi's : human, book$ (i) Man with a *damaged* eye. $aoi's : damage$ (h) Fixation patterns and face localization when the damaged eye is restored. $aoi's : face$ 135

56 Discrimination obtained by the cluster profiling method, the vertical axis plots accumulated scores for different images measured using equation 14. Distinct images have grouped under each of the 4 themes, the plot represents values over more than 100 images. The method separates out images with strong visual elements and interactions *affective*-red,*aesthetic*-green and *action*-blue from those which have low interaction or weak visual elements (magenta). *action* and *affect* images are grouped together by the measure described earlier in 14, this needs to be investigated further. . . . 137

57 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[60]. 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. 140

58 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.15. 141

59 The pseudocode describes details of steps (a) to (d). . 142

60 F measure plot for 80 images showing the improvement brought about by using multiple fixation seeds for segmentation (d) in comparison to the baseline (a) using equal number of random locations within the object as segmentation seeds. The legend is as follows - *red* baseline and *green* - Integration of segments obtained from multiple sub-clusters. 144

61 The schema for guiding sliding window based object detectors using visual attention information. Image pyramid (a) is obtained by successively resizing the input image *I* over L levels. Features corresponding to areas covered by sliding, rectangular windows at each level l_i are combined with a template based filter (b) to generate scores indicating presence of the object. These are combined over all levels that indicate the presence of the object. Eye-gaze information is used to extract Regions of attention (ROIs) (d), which then restrict the image region for object search. The number of scales (c) are restricted to a small fraction of possible levels, using scale information from ROIs (e) is the output from our method and (f) from a state-of-art detector [23]. . . 145

62 Illustration of the significant reduction in computation time achieved by constraining the state-of-art object classifier in [23] using eye-gaze information. 150

63 Illustration of the improvement in $fmeasure$ of over 18 % achieved by constraining the object classifier in [23] using eye-gaze information. $fmeasures$ are recorded from our method *VA* and that of [23] attempting to find the concept *person* over 150 images. The images were chosen to capture diversity in number of instances, size, activity and overall scene complexity. 151

64 The panel illustrates outputs at every stage of our attention driven method. (a) Original image of a crowded street-scene. (b) Manual ground truth annotation boxes for key objects. (c) Clusters identified from eye-gaze information, centroids marked by red circles. (d) ROIs generated based on cluster information. (e) Detected instances of *person* class within ROIs, using detector from [23] marked by yellow boxes. (f) Finally result detections after filtering for ROI size, marked by red boxes. (g) Results for the same image from the baseline detector. 152

65 (a),(b) Cases where visual attention greatly enhances performance of the detection system, (e),(f) are the corresponding results for (a),(b) from the multi-scale, sliding window method in [23]. (c) A case where attention directs ROIs away from non-central, but seemingly important *person*s. This problem is not faced by the baseline as seen in (g). (d) Generated ROIs are not good enough to permit detection, the baseline outperforms our method in this case as seen in (h). 152

66 Panels in the top row illustrate important stages in the interactive framework. (a) A frame from the video stream. (b) Eye-gaze based ROIs discovered using the *video binning* method [46]. The red circle shows current location being attended and yellow circles show past ROIs, the arrows show dominant eye movement trajectories. (c) An example image region overlayed with motion saliency computed using motion vector information in the encoded video stream. (d) Face saliency map constructed by detecting and tracking frontal and side profile faces. Panels in the middle row visualize stages in the dialogue captioning framework. (e) Regions likely to contain faces are combined with ROI and likely eye movement paths shown in (f) to compute likely locations to place the dialogue currently in progress. (g) Video sequences are dynamic and object motion as well as camera motion cause change in position of the dominant objects over successive video frames, this combined with noisy eye-gaze ROIs in turn gives rise noticeable and annoying jitter. (h) A history of dialogue placement locations is maintained and smoothed over to obtain smooth movements of overlayed dialogue boxes across the screen. Video frame taken from the movie *swades*©UTV Motion Pictures. 157

67 Group A part of the subject pool, changes its decision as the dialogues are restricted to the locations where they are initialized. On the other hand, subjects in the larger pool, Group B, do not their change their preference and consistently report better comprehension and viewing comfort with static captions. One reason for such response could be the familiarity and habituation to static captions through long exposure to current captioning. 162

List of Tables

1 Typical tasks accomplished in Automated Image understanding and relevant references. 53

2 Details of Flickr images collected for 5 of the 14 image themes chosen. 61

3 Image distribution in the NUSEF dataset, organised according to semantic categories. 68

4 A brief comparison between datasets in [43], [6] and [11] with NUSEF [71]. 69

5 Computation of w_i for a_i's corresponding to the semantic image categories shown in Figure29. 94

6 A comparison of the salient features of [24] with those of the binning method proposed in this thesis. 111

7 The accuracy achieved by a personalized model trained for individual user's aesthetics preference. 129

8 Combining concept detectors and fixations to classify *face* and *person* images. 131

9 Using eye-gaze information to classify for *Action* and *No Action* social scenes. 132

10 Performance evaluation for segmentation outputs from (a), (b), (c) and (d). 142

11 Evaluation of the visual attention guided *ROIs* against human annotated ground truth. The object detector is not run in *ROIs* and instead the entire ROI is considered to be a detection, this experiment illustrates the meaningfulness of the ROIs generated by our method against human annotated ground truth boxes for different concepts in our database. This is especially significant in cases like *bird and cat/dog*, where the baseline detector fails completely. 153

12 Description of the video clips chosen for evaluation of the online framework and applications. The clips were obtained from the public domain and normalized to a 5 minute duration. The clips were chosen from amongst social scenes, to have variety in the theme, indoor and outdoor locations, spoken language and extent of activity in the video clip. 155

13 Different modes in which video clips were shown to subjects during evaluation of the online, interactive captioning framework. The first 8 participants were shown captions with opaque or semi-transparent blurbs, subsequent participants saw text-only dialogue captions. . 156

14 Changes in clip comprehension when text caption placement in constrained in different ways. A clip is counted only once for the mode in which it is shown for the first time in a subject's viewing list. Floating dialogue captions were found to be very annoying and subjects also report that it hindered their comprehension. This is also visible in the *comprehension* value for the first row. The clip comprehension and overall subject feedback improved as the dialogue captions were restrained to the initialization locations. An additional *manual placement* mode was generated by using the mouse as a proxy for eye-movements and this improved the user feedback and comprehension slightly. 160

15 Dynamic caption strategies draw significant amounts of user attention as can be seen from the fraction of eye movements spent on exploring dialogue boxes. This can be seen in the high fraction of gaze points taken up by dialogue boxes in column 2. 162

1 Introduction

1.1 Visual media as an artifact of Human experiences

Huge volumes of images and video are being generated as a result of human experiences and interaction with the environment. These can vary from personal collections containing thousands of videos and images, to millions of video clips in communities such as *YouTube* and billions of images on repositories such as *Flickr* or *Picasa*. It becomes useful and necessary to automate the process of understanding such content and enable subsequent applications like indexing and retrieval [69][84] and query processing, re-purposing for devices with different form factors [80]. This thesis focuses on the hypothesis that looking at media and human perception together, is a more holistic way to look at problems relating to image and video understanding than to try and understand visual content alone in isolation.

A growing body of research is correlating human understanding of scenes to the underlying semantics [86][34][46], affect [70][71] and aesthetics [45]. Early research in image and video analytics focused almost entirely on low level information to understand visual content, the shortcomings of such approaches have been discussed elaborately in [83]. A more recent survey has pointed out the importance of modeling higher level abstractions [69]. This thesis also shows how understanding abstract information such as semantics and affect, can lead to improvements in signficantly hard problems in computer vision[71][45], multimedia indexing and retrieval[70][46] and aspects of human-media interaction.

1.2 Brief overview of work presented in this thesis

The focus of this thesis is to get a better understanding of visual perception and attention as people interact with digital images and video. Chronologically, the first problem was on finding how low level global and local information in images influence category discrimination and aesthetic value in images [45]. This work identifies the important role of color in aesthetics discrimination in pre-attentive time spans and also established that humans can distinguish simple notions of aesthetics even at very short presentation times $< 100ms$. We also establish a minimum presentation time threshold for aesthetics discrimination in images. Modeling using global color based features and SVM classifier training is used to identify aesthetic images from the publicly available Flickr dataset (manuscript in preparation).

Subsequent work investigates how semantic and affective cues relating to objects and their interactions influence scene semantics in static and dynamic scenes [70][46]. Eye-tracking is used as proxy for human visual attention. Preliminary work on free viewing affective images resulted in *world model*, that quantifies *attentional bias* amongst common and important concepts in social scenes [70]. The attentional bias is measured in terms of fixation duration and frequency across different concepts in the image. Our dataset named NUSEF, has now been made public. NUSEF contains images with a diverse set of visual concepts like faces, people, animals along with a variety of objects with varying degrees of action/interaction commonly encountered in social scenes. It has already been adopted and cited by some of the leading research groups in vision science [42][107].

Novel measures are developed to infer interaction, location and scale of visual concepts in images using eye-gaze data. These have been reported in a series of publications [70][46]. Furthermore, an efficient and robust clustering method for eye-gaze data is developed in the form of the "binning" [70] algorithm. The *binning* algorithm can give a good prior on the location and scale of regions of human interest (ROIs) in images and video. It compares well against state-of-art in terms of computational cost, precision and recall over the underlying visual concepts. The binning method and fusion with image content based features has given interesting insights as to how a few dominant regions of interest can influence the theme and overall interestingness and semantics of an image (manuscript in preparation). A framework is developed for fusion of eye-gaze data with bottom-up low level visual cues like motion and saliency in video streams and top-down semantic cues from object detectors (manuscript in preparation). We have encouraging results for video and image content. Furthermore, we have extended this framework for online human interest prediction as the subject watches a video stream and demonstrate applications for intelligent on-the-fly placement of foreign language captions onto video streams.

The usefulness of the quantitative attentional bias model, interaction measures and binning algorithm is shown via challenging applications. Examples include detection of key objects and interactions for image summarisation, foreground object segmentation for difficult natural scenes, localisation of text keywords into image content. More recently we have also demonstrated the usefulness of the quantitative models and *binning* for interactive video repurposing.

1.3 The notion of *Goodness* in visual media processing

A very important problem at this juncture is that of *Goodness*, a notion that can encompass popular measures such as *precision* and *recall* in a reliable manner and to human satisfaction. In this thesis I have explored how human perception can determine the interpretation of each of these measures. The meaning of *interest* is explored both in the context of rapid image categorization[45] and image understanding over longer time spans via eye-movements[70]. *Precision* and *recall* are explored in the applications to key object detection[46] and foreground object segmentation[71]. This can also be seen in the fact that most mid and high level multimedia problems relating to *image categorization*, *object segmentation and detection* [19], *action recognition* and *event detection* in images and video remain open challenges. More abstract problems such as inferring *affect*, *aesthetics* and *interactions* in images and video are even harder. This could be due to the sensory nature of visual media, evolving mathematical formulation and lack of complete knowledge of Human visual perception.

In most images and videos, statistical properties of the scene [88] and the optical properties of light [26] lend far easily to measurement and modeling as compared to the human perception of the same. I have studied some aspects of human visual perception such as rapid *pre-attentive* processing in the absence of attention and relationship between eye-movements and scene semantics. Rapid processing is shown for scene classification in basic categories such as *(indoor, outdoor, man-made and natural)* [52][88], detection of basic categories such as *(faces, animals, cars)* [52] and rapid discrimination of *interestingness* between images of the same semantic theme [45]. Subsequent scene-understanding tasks such as identifying key objects and their relationships, rely on exploration of the Visual scene, and the phenomenon of *Human visual attention (HVA)* is an important part of this process. Eye-movements allow selection of regions of interest in a scene and detailed understanding of the same. This is a complex phenomenon and is influenced by learning from past experiences, current visual input and task-at-hand amongst other factors. Most image and video analysis methods run on different computational principles and need not take into account biological phenomenon as long as it performs satisfactorily. Current state-of-art though, lags significantly in producing satisfactory results in realistic image and video understanding problems [19][66].

1.4 Human in the loop, HVA as versatile ground truth

An approach to close the gap between an end-user's notion of correctness and that provided by an Image or Video based application, is to put the Human visual system into the interaction loop at an appropriate stage. Attempts are made in this thesis to demonstrate the feasibility and effectiveness of involving *Human Visual Attention (HVA)* within media processing methods for low, mid and abstract Image and Video processing tasks.

A similar principle has been employed in the form of *Relevance Feed-*

back [106] using manual user feedback, or the use of user interaction logs to improve content based image retrieval (CBIR) [62]. Machine learning methods such as active learning [49] have been used for this purpose.

HVA is a less explored option, both due to lack of knowledge of exact functional role in attention and novelty of eye-gaze in context to Computer science and Media processing. *HVA* is an intricate strategy with an observable *overt* component that lends to measurement and a *covert* component which is hard to measure non-invasively. In this thesis, *Eye-gaze* measurements that can be considered as a proxy for *overt* (observable) shifts in *HVA* [37], have been used to enhance automated scene-understanding.

1.5 Human visual attention and eye-gaze

The central portion of the retinal wall of the eye is called the *fovea* and is lined with cone cells meant for color and fine detail processing, the central region rich in *cone cells*, or *fovea centralis*, is surrounded by *rod cells* that used for sensing intensity information. Measured density curves for the rods and cones on the retina as shown in the Figure. 1 show an enormous density of cones in the *fovea centralis*. Both color vision and the highest visual acuity are attributed to cone cells in this region. Visual examination of small detail involves focusing light from that detail onto the fovea centralis. On the other hand, the rods are absent from the fovea. Cortical regions in the brain, responsible for higher visual processing have disproportionate resources allocated to inputs from the foveal region. Thus the spatial resolution at the

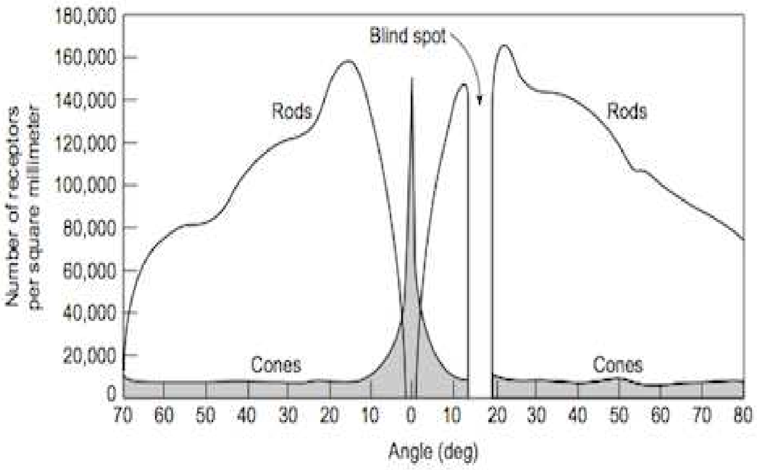


Figure 1: Panel illustrates the distribution of rod and cone cells in the retinal wall of the human eye. The highest acuity is in the central region (fovea centralis) with maximum concentration of Cone cells. The blind spot corresponds to the region devoid of rods or cones, here the optical nerve bundle emerges from the eye. http:// www. uxmatters.com/ mt/archives/2010/07/ updating-our-understanding-of-perception-and-cognition-part-i.php.

center of the visual field is quite high. This can be understood from the visual acuity chart illustrated in Figure 2, frequently used in optometry. Human's can identify much smaller characters at the center of the chart, than at the periphery. The high resolution area corresponding to the fovea is about the size of one's thumb nail, as one extends his/her arm out fully.

HVA is the strategy employed to execute ballistic eye-movements to align a Region of Interest with the *fovea centralis*. Each ballistic movement of the eye called a *saccade*, is followed by a brief duration of *fixation* where the eye-gaze holds steady for a short while to inspect the Region of Interest. This strategy is critical to the limited capabilities of the Human visual system, as the total visual input from the environment is too overwhelming and oftentime redundant in large parts. Though eye-movements appear seemingly involuntary,

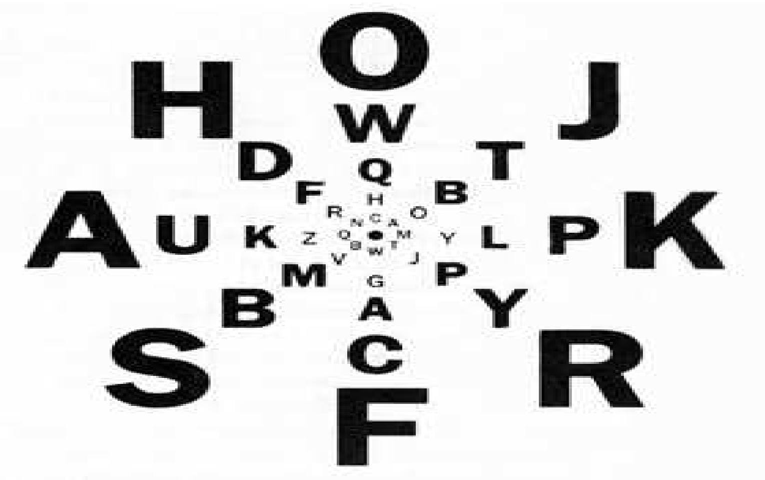
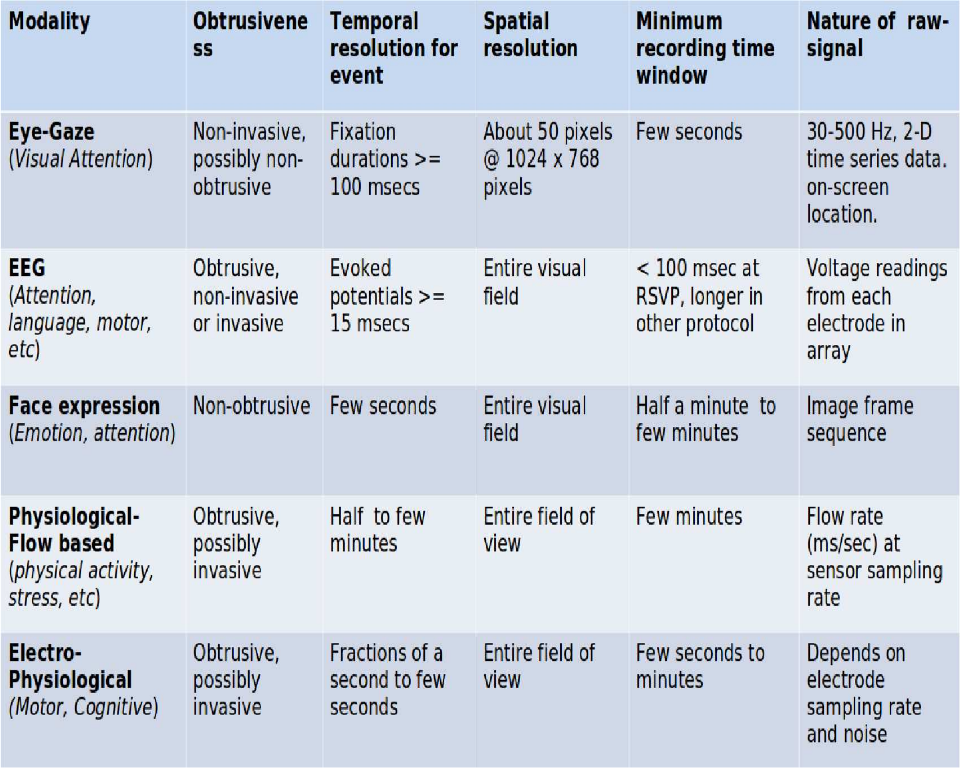


Figure 2: A standard visual acuity chart used to check for reading tests. Humans can distinguish characters at a much smaller size at the center than at the periphery. http:// people.usd.edu/ schieber/ coglab/IntroPeripheral.html

they can be influenced by (a) high level over-rides, like an explicit instruction not to look in a particular direction or Inhibition-of-return phenomenon, where a larger cost is associated with an already explored spatial location. This again can be over-ridden by top-down information like semantic importance of a location in the visual field. This would cause attention to repeatedly revert to the some locations.

1.6 Choice of Eye-gaze to investigate Visual attention

HVA in general and eye-gaze in particular is one of the many non-conventional sources of information being considered in the research community. Key attributes of different modalities are presented in Figs. 34. Being non-invasive spatio-temporal information that is well correlated to the human visual process, eye-gaze makes an ideal candidate for Image and video understanding problems.

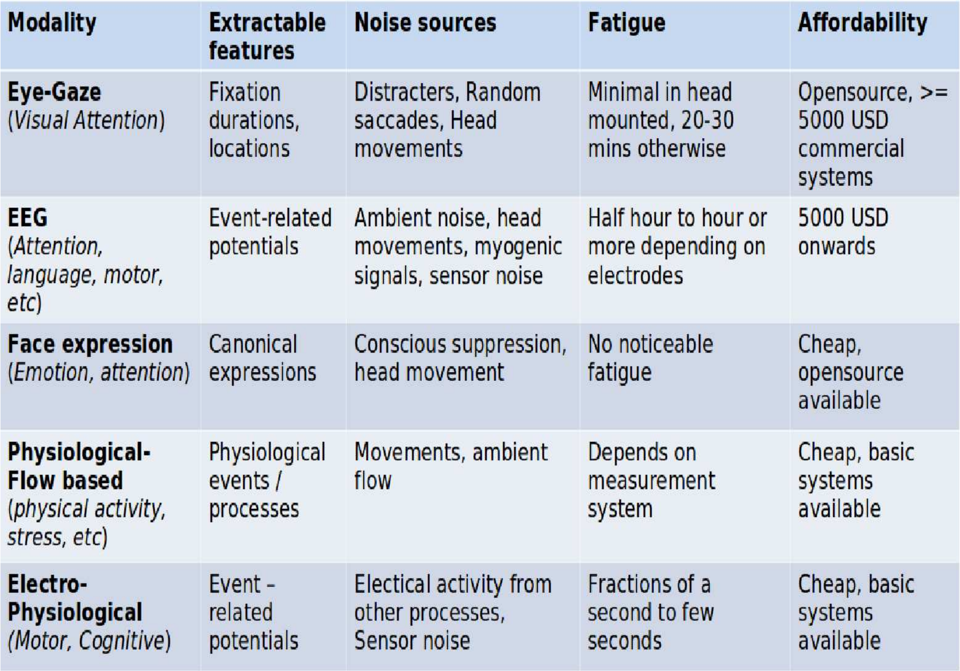


Modality	Obtrusivene ss	Temporal resolution for event	Spatial resolution	Minimum recording time window	Nature of raw-signal
Eye-Gaze (*Visual Attention*)	Non-invasive, possibly non-obtrusive	Fixation durations >= 100 msecs	About 50 pixels @ 1024 x 768 pixels	Few seconds	30-500 Hz, 2-D time series data. on-screen location.
EEG (*Attention, language, motor, etc*)	Obtrusive, non-invasive or invasive	Evoked potentials >= 15 msecs	Entire visual field	< 100 msec at RSVP, longer in other protocol	Voltage readings from each electrode in array
Face expression (*Emotion, attention*)	Non-obtrusive	Few seconds	Entire visual field	Half a minute to few minutes	Image frame sequence
Physiological-Flow based (*physical activity, stress, etc*)	Obtrusive, possibly invasive	Half to few minutes	Entire field of view	Few minutes	Flow rate (ms/sec) at sensor sampling rate
Electro-Physiological (*Motor, Cognitive*)	Obtrusive, possibly invasive	Fractions of a second to few seconds	Entire field of view	Few seconds to minutes	Depends on electrode sampling rate and noise

Figure 3: Comparing attributes of different non-conventional information sources. Representative values for important attributes of each modality are obtained from [103] (EEG), [70] (Eye-Gaze), [99][76] (Face detction expression analysis) and [98] (Electrophysiological signals).

1.7 Factors influencing Visual Attention

In addition to the aspects of visual attention addressed in this proposal, current research in HCI, Behavioral science, media research has found that beyond the visual properties of any semantic category, factors such as our real world model, short term memory and task/intent in the user's mind significantly influence Visual attention and subsequent understanding of an image [44]. This is illustrated in the following figure, Each of these factors has been shown to influence visual attention. Visual attention is an important mechanism to isolate, selected regions of the visual field for more detailed observation. Hence it is important to acknowledge and if possible control the



Modality	Extractable features	Noise sources	Fatigue	Affordability
Eye-Gaze (*Visual Attention*)	Fixation durations, locations	Distracters, Random saccades, Head movements	Minimal in head mounted, 20-30 mins otherwise	Opensource, >= 5000 USD commercial systems
EEG (*Attention, language, motor, etc*)	Event-related potentials	Ambient noise, head movements, myogenic signals, sensor noise	Half hour to hour or more depending on electrodes	5000 USD onwards
Face expression (*Emotion, attention*)	Canonical expressions	Conscious suppression, head movement	No noticeable fatigue	Cheap, opensource available
Physiological-Flow based (*physical activity, stress, etc*)	Physiological events / processes	Movements, ambient flow	Depends on measurement system	Cheap, basic systems available
Electro-Physiological (*Motor, Cognitive*)	Event - related potentials	Electical activity from other processes, Sensor noise	Fractions of a second to few seconds	Cheap, basic systems available

Figure 4: Additional attributes of different non-conventional information sources, continued from Fig. 3. Representative values for important attributes of each modality are obtained from [103] (EEG), [70] (Eye-Gaze), [99][76] (Face detction expression analysis) and [98] (Electrophysiological signals).

influence of these factors in human-image interaction. For example, the Short term memory allows remembering most recently encountered concepts. The practical necessity of modeling the Short term memory can be seen in past and recent multimedia research focusing on mouse-click sequences [12][58], recent pages visited and even recent user context from recent user activity in the system [2].

The influence of Task/Intent was illustrated in the seminal research by Yarbus [53], in which he illustrated significant changes in visual attention when subjects were shown the same visual stimulus, but with different task while viewing the image. The results from this seminal research are shown in the following Figure. 6.

The influence of a subject's *Real world model* is harder to gauge,

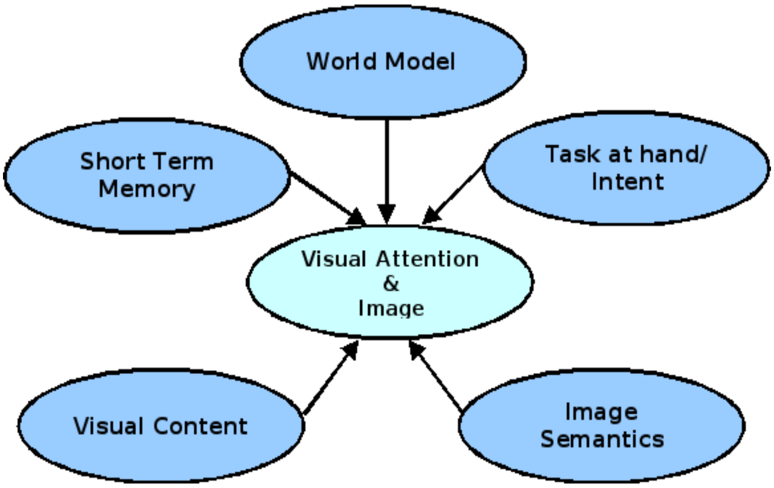


Figure 5: Different factors that can affect human visual attention and hence, subsequent understanding of visual content.

as it spans his or her experiential learning, familiarity or expertise in a domain, cultural and gender based biases and behavioral traits. Some intuitive scenarios are that attention patterns of a domain expert or familiarity with a given image will result in very distinct visual patterns, and will be different from a naïve user. We could measure an example of this influence in the eye-gaze patterns of male and female subjects as they viewed images of *nude* men and women, or *nudes* in intimate positions. The gender-bias was observed in terms of the image regions attended. We also observed familiarity/expertise based bias in attention patterns as subjects were presented the same visual stimulus in successive trials. Increased subject attention is observed over most salient/interesting concepts and also to image details that are ignored in initial presentations. The influence of clinical disorders is out of the scope of this research and is not discussed here, the focus is on healthy subjects with normal or corrected vision. We could quantitatively measure and model this for a general image understanding task in our experiments. We try and control for the influ-

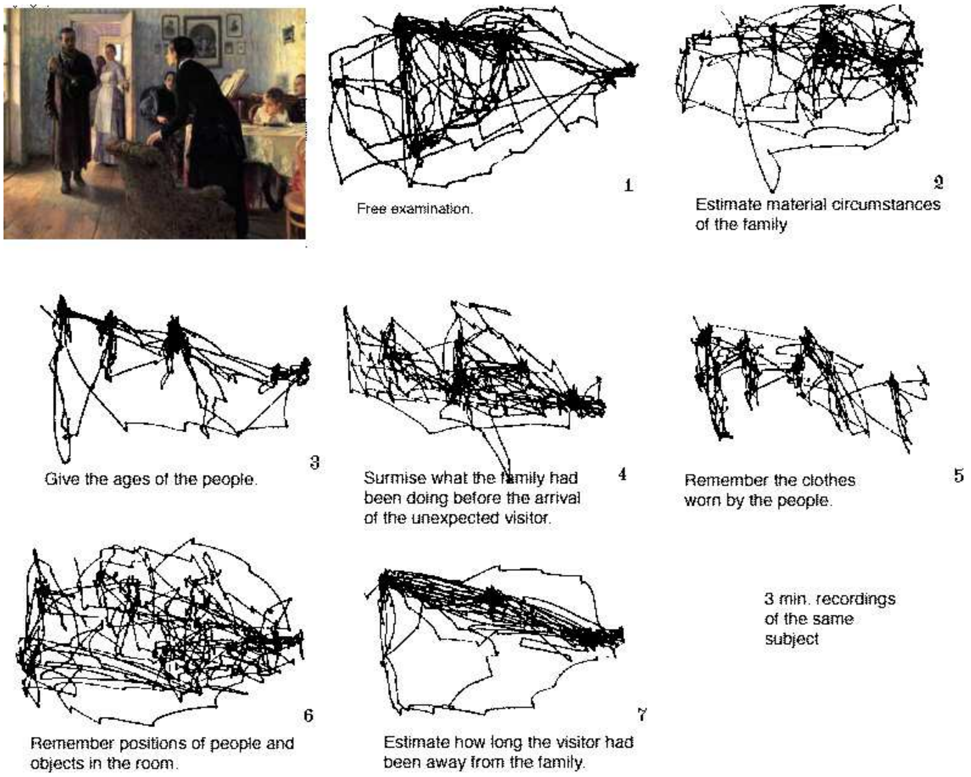

1
Free examination.
2
Estimate material circumstances of the family
3
Give the ages of the people.
4
Surmise what the family had been doing before the arrival of the unexpected visitor.
5
Remember the clothes worn by the people.
3 min. recordings of the same subject
6
Remember positions of people and objects in the room.
7
Estimate how long the visitor had been away from the family.

Figure 6: Some results from Yarbus's seminal work [53]. Subject gaze patterns from 3 minute recordings, under different tasks posed prior to viewing the painting "An unexpected visitor" by I.E. Repin. The original painting is shown in the top left panel. The different tasks posed are as follows, (1) Free examination with no prior task (2) A moderately abstract reasoning task, to gauge the economic situation of the family (3) To find the ages of family members (4) Another abstract task, to find the activity that the family was involved in prior to arrival of the visitor (5) To remember the clothes worn by people (6) To remember positions taken by people in the room (7) A more abstract task, to infer how long the visitor had been away from the family.

ence of *Short term memory*, *Real world model* of different people and *task-based biases* in our experiment. The influence of low-level and semantic information in images was then brought out through analysis and modeling. The results were also demonstrated by a semantic concept localization application involving short text captions [70].

1.8 The role of Visual Saliency

Saliency is popularly used with the notion of *Visual Conspicuity* like in the popular Visual saliency model proposed in [38] and more recent and comprehensive models for videos [3] and images [92]. Earlier notions of Saliency as being bottom-up and being equivalent to *pop-out* effects of *color* or *contrast* [38] for example, have given way to more integrated models that bring in cues like *curves* and *shapes* [92] or higher level cues such as *key objects* [18] and faces [43].

For scenes which lack clear semantics, such as arbitrary shapes, *HVA* seems to be motivated by maximizing *Information Gain* by successively processing high entropy locations in the Visual field [72]. For visual input with richer semantics, it has also been shown that Humans give unequal importance to different objects in natural scenes [86], the authors also show correlation of this *attentional bias* with recall statistics after the stimulus is removed. Similar, consistent bias for object categories in *social* and *affective* scenes is shown in this thesis[70]. A combination of this *bottom-up* and *top-down* which can take *low and mid level visual cues*, influence of objects and some abstract semantics such as *affect* and *action* has been shown to account significantly for *HVA* strategies [44].

A parallel line of research has focused on modeling *scene context*. *Context* estimation has been shown to be useful in scene understanding tasks such as scene classification [88], object detection [91] and Saliency estimation for image re-targeting and collage generation [75]. In this thesis, the emphasis is on Eye-movement driven modeling and notion of the best *foreground* or *Region of Interest* estimate for a given task, Context estimation is not dealt with in detail. With richer *semantics* and *affective* semantic content, it becomes harder for automated methods to reliably estimate *visual conspicuity*. This thesis also looks into important categories derived from computer vision literature, photography and personal media collections, with rich underlying semantics. Methods are proposed to bring a human-in-loop to estimate notions such as *saliency*, *semantics* and *affect* in visual content.

1.9 Semantic gap in visual media processing

The *Semantic gap* continues to present an important and challenging problem in image and video analysis, it typically arises from the disparity between text based and visual descriptions of a concept. Humans commonly adhere to well formed vocabularies and grammatical frameworks while describing visual content using text. Most people can be considered as trained individuals or experts expressing themselves. Visual media on the other hand, can be sensory information representing statistical properties of the environment or intuitive patterns of scene composition. Effects of the semantic gap become more prominent with increasing abstraction represented by the visual concept, this has resulted in most of current image and video analysis to be focused on concepts with small semantic gap [55]. Though it is attractive to work on simple concepts such as simple inanimate objects (CALTECH 101, 256)[27], faces[56], people[17], etc, many social scenes exhibit meaningful interaction [71], affect [70] and aesthetics [14] and present challenging problems for scene understanding.

The difference in understanding visual and textual representations of the same concept can lead to the *Semantic Gap*. Another way that the *Semantic Gap* can manifest is between understanding of visual content at the time of *Content-Creation* and *Content-Consumption*; Figure 7. Visual media based communication will be less ambiguous and relatively unaffected by semantic gap when content generators as well as content consumers are expert users having shared domain knowledge, training, controlled vocabularies, guidelines and prior experience. On the contrary, naive users interacting through visual media will face significant problems due to lack of aforementioned aspects in Expert-to-Expert communication. Experimental and algorithmic methods are proposed in this thesis along with applications to demonstrate to enhanced image and video understanding. An interdisciplinary mix of psycho-physical, behavioral experiments are used to gain insights into human vision and to improve upon state-of-art in a variety of important problems relating to understanding of low, mid level and abstract information in *images* and *videos*.

An attempt to address some aspects of the semantic gap has been made in this thesis the preferences of individual users and communities for interesting images is studied and modeled in Section 5.1, this application closes the expert to naive user gap between the Flickr®system and users. The naive user to naive user gap is addressed through use of eye-movements to extract scene semantics of human interet in the text caption localizationa application Section 5.3 and detecting foreground regions Section 5.5 and key objects Section 5.6 of human interest.

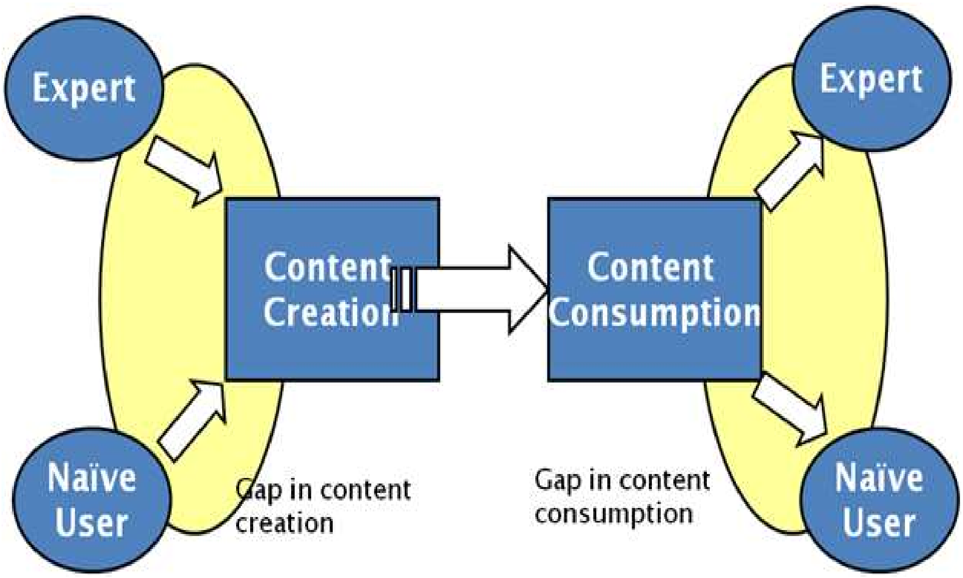


Figure 7: The semantic gap can show up in more than one way. The *Intent* of an *Expert* or *Naive* content creator can get lost or altered either during encoding into visual content, or in conversion between media types during the *(encode,store,consume)* cycle. Effects of the Semantic gap are more pronounced in situations where *Naive* users generate and consume visual media.

1.10 Organization of the Thesis

The *Introduction* Chapter 1, is followed by the important *Contributions* from this thesis in Section 1.11. Relevant literature and state-of-art is covered as *Related Work* in Chapter 2. Chapter 2 covers important references and background literature for aspects of *Visual content understanding*, *Human Visual Attention* and some important multimedia applications that are addressed in this Thesis. Organization of subsequent chapters follows the schematic shown in Fig. 8.

Data pre-processing and experimental protocols to acquire eye-tracking data and meta-data are explained in chapter 3. The methods described psychophysical experiments, eye tracking protocols and choices made for dataset and experiment design. Chapter 4 describes analysis and modeling done using visual content as well

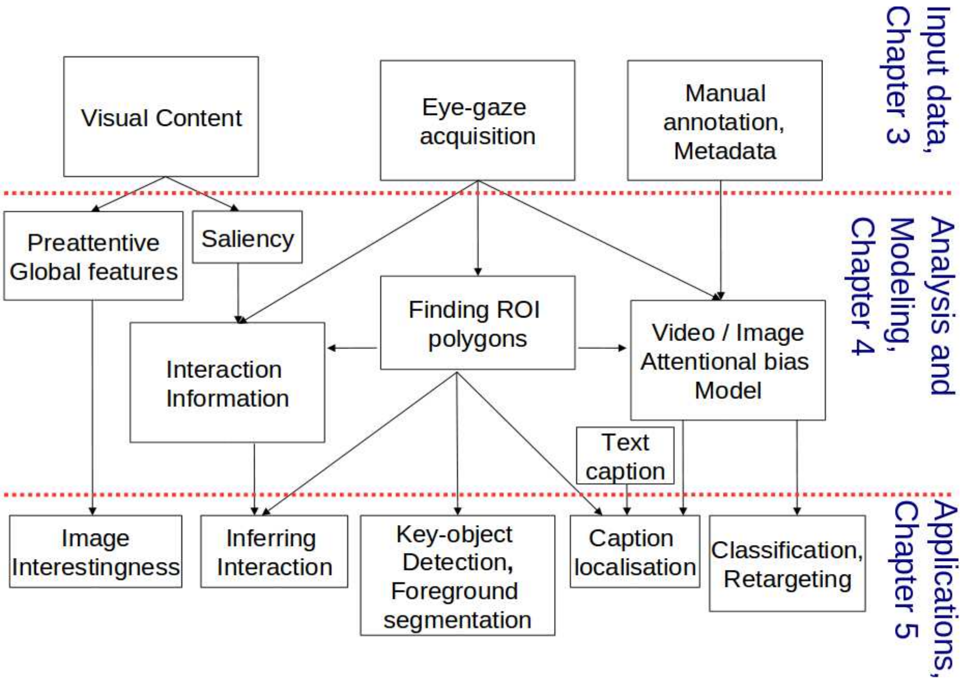


Figure 8: The schema represents information flow hierarchy and chapter organization in the thesis. The top layer lists different input modalities that are then analysed in the middle layer to extract features and semantics related information.

as eye-gaze information. Global image information involved in rapid aesthetics discrimination in some important categories of images is also described in chapter 4. This chapter also introduces eye-tracking methodology and statistical measures to analyses eye fixation information. The close relationship of Human Visual Attention (HVA) to semantics and affect in images is demonstrated by constructing an *Attentional bias model*, which captures preferential attention biases towards some concepts and their relationships. This chapter also introduces the novel *binning* method to discover ROIs and its extension to discover dominant interactions in images.

Applications to low and mid level problems in image understanding are described in chapter 5. An application of text caption localisation, which exploits attentional bias is presented. Use of eye-gaze and HVA to bring about improvements in problems such as object segmentation and key-object detection are also demonstrated.An interactive, caption-localisation application is demonstrated for videos.

1.11 Contributions

The main theme of this thesis is understanding the role of perceptual mechanisms such as visual attention(HVA) and pre-attentive information processing for image and video content. Some interesting phenomena relating to understanding scenes and the concepts therein are taken up in this context. Eye-gaze is exploited to better understand human perception, behaviour and address computer vision problems such as segmentation, object detection, inferring interactions and classifying images. There exists a *semantic gap* between sensory representation of scenes and the semantics inferred subsequently by a human.To summarize, the key contributions of the thesis are,

- Bringing forth the capability of human vision to rapidly categorize interesting images in very short time spans.
- Methodology to infer scene semantics in images and videos using a combination of eye movements and content analysis and the novel *binning* algorithm to cluster eye-gaze information, along with extentions for content fusion.
- Release of experimental data for use by the research community as the publicly accessible NUS Eye fixation (NUSEF) dataset. Our effort has already been recognised by leading researchers

in vision community [42][107] .

- Using eye-gaze to exploring the possibility of interactive applications with a human-in-loop.

2 Related Work

2.1 Human Visual Perception and Visual Attention

This chapter explains and summaries phenomena related to Human Visual perception and relevant to the problems taken up in this thesis. The Pre-attentive stage and its importance in rapid understanding of Visual content is discussed. The phenomenon of Human Visual Attention is then presented along with its relevance to Image and Video understanding. Human Visual attention (HVA) is a mechanism employed by primates to identify a subset of the visual input for further processing. HVA is believed to optimize the search inherent in vision. Posner [68] concluded that that visual attention is the result of processing by a network of anatomical structures including the brain. It does so by selectively tuning the visual processing network. This is accomplished by a hierarchy of winner-take-all processes embedded in the visual processing pathway [40]. Compensation for the slow speed of neuronal circuitry and reduction of scene complexity could be amongst the important reasons for visual attention [38].

2.2 Eye-gaze as an artifact of Human Visual Attention

Eye-gaze is believed to be a mechanism that brings selected regions of the visual field to coincide with the foveal region on the retina, so that finer details can be processed. Many popular models for *HVA* inspect image content at multiple scales and information channels [38][63] and more recent approaches that analyze contour and color information [92] to compute notions of *novelty, surprise* or *vi-*

sual conspicuity to fine spatio-temporal regions in visual input that will be chosen by *Human Visual Attention (HVA)*. Similar models for video, inspect successive frames in a video sequence to compute *surprise* [36]. Though these models do not account for top-down influences of *task-at-hand, behavioral influences* and *personality traits*, similar models have been successfully applied in applications such as saliency based image retargeting [75] and surveillance video synopsis [102]. HVA is responsible for shifts in spatial attention that are either observable as eye-movements (*overt*) or hidden (*covert*). A significant proportion of attentional shifts are in tandem with *overt* shifts, measurable as eye-gaze as explored and exploited usefully in this thesis. It is also possible to have *covert* shifts in absence of eye-movements as demonstrated in [32].

The time course of Human image understanding can be categorized roughly into an early pre-attentive phase $(<= 100ms)$ involving rapid, global information processing followed by HVA involving eye movements for exploration of the scene. This phase has been shown to be important in context estimation [64] and precede understanding of local information [57]. The role of global and local features in image classification has been explored in [101] where usefulness of global image properties have been verified experimentally and computational modeling is done to bring forth the importance of color information. Another work that is similar in spirit is [64] where a low-dimensional representation of global image information shows discrimination for basic scene categories like streets, highways, coast, etc the authors show how second order statistics can discriminate perceptual qualities such as naturalness, openness, roughness, expansion, ruggedness. Rapidly processed global information has been shown to act as a prior scene-context for subsequent HVA based exploration, an example is [65] where authors show how global information based contextual priming is a good indicator of key object locations. Human capability to discriminate between *interesting* images from ordinary ones amongst consumer photos in a rapid, pre-attentive time span has been demonstrated in [45].

Eye-gaze has shown to be significant in local information processing [33]. More interestingly, eye-gaze is influenced significantly by key objects in a scene as shown in [18], making it very valuable in image understanding. Furthermore, top-down attention appears to dominate while viewing semantically rich and affective images [8], where authors show that humans preferentially attend to emotional content (both *pleasant* and *unpleasant*). This has been shown to be true even during the first 500 milliseconds, when emotional and neutral stimuli are presented simultaneously. Strong correlation of eye-gaze with abstract concepts such as *affect* and key-object *interactions* has been show in [70]. Since eye-gaze is employed to explore key objects and their parts, it is possible to estimate the scale of key objects that are gazed at [46]. Eye-gaze analysis can yield rich information like salient regions of interest in Images [70][46] and video [87]. The sensitivity of eye movements to motion and low-level saliency in videos has been exploited in [87] to propose a calibration free method for eye-gaze estimation.

2.3 Image Understanding

Using computers to extract meaning from an uncontrolled set of digital images remains a significant challenge. The aim of such algorithms is to enable computers to *“See what we see, and understand as we do”* from a digital image or video. This can be seen in recent image understanding challenges such as the PASCAL VOC [19] with the following results image classification (MAP: 0.4695), object detection (Mean Accuracy: 17.52 %), object segmentation (Mean Accuracy: 16,77 %) over 20 classes, and in ImageCLEF [67] with the following results photo annotation (MAP: 0.2812) and wikipedia based image retrieval (MAP: 0.1388). Automated Image understanding deals with issues all the way from pixel-level representation, low-level analysis of local information, mid-level shape information, contours to high-level tasks of concept detection, context analysis.

Early research focused on low-level feature based Image indexing and retrieval based primarily on local geometry, color and texture based image processing leading to features such as salient points, shapes, global features, etc [83]. Though there has been significant progress in the last 10 years in Content based Image Retrieval (CBIR), the Sensory-to-Semantics gap is yet to be bridged satisfactorily [69]. In computational modeling, an image can be broken down to its *context*, *objects involved*, *ontological relationships* between objects, *inter-object interactions* and more *abstract concepts* like affect and aesthetics. Key references related to these tasks are listed in Tab. 3.

The global context of an image has been estimated using low-level,

Task	Type of information & Region of Interest	References and methods
Global	**Entire image**	**[64]**
Local	**Small patches**	**[54],[1]**
Saliency	**Image Segments at Multiple scales**	**[38],[92],[87]**
Object Detection	**Image Segments**	**[13],[22],[94]**
Aesthetics	**Image Segments & Entire image**	**[14][51]**
Interaction	**Multiple Image regions**	**[70]**

Table 1: Typical tasks accomplished in Automated Image understanding and relevant references.

global image statistics in [64], where the authors use Fourier spectrum analysis followed by Principal component analysis (PCA) to discriminate between basic natural scene types and underlies the GIST operator. The method relies on characteristic changes in image gradient information with changes in the depth, contents and scene-context of natural and man-made scenes. This kind of low dimensional representation of images has been shown to be scalable in inferring context in millions of images [90]. Global features have also been shown to be useful in automated image categorization in [100].

Interest points are another interesting and useful construct computed from low-level pixel information analysis. Local changes of intensity information in image patches yields features such as Harris corner points [29], SIFT [54] is the current the state-of-art in interest point detection. SIFT interest points are locations that show high entropy at multiple scales, the descriptor is a histogram around the location over multiple scales and orientations. The generalized Hough transform is another interest point definition [4] and has been used for object detection in [25]. Encouraging results on eye-gaze based enhancements in interest point based representation of objects are shown later in this thesis.

An important complementary problem in Object detection is that of segmentation and identification of the object contour. Low-level color analysis has been used in [16] to identify object and part segments in a scene. Semi-automated image segmentation has been proposed in [60] using a single seed location input by a human user. The seed location must lie within the object boundary, the problem is then modeled as a background separation problem using the graph cuts [5] method. Our contribution proposes the framework for bringing in eye-gaze based multiple fixation seeds to [60] and demonstrate superior segmentation results in challenging scenes.

Another successful low dimensional representation is the histogram-of-gradients (HOG) approach [13] based on binning gradient information in image regions along pre-determined orientations. Computing the HOG over small image patches characterizes local properties, and has been used for person [13] and state-of-art generic object detection [22]. Another sliding-window based approach is taken in [94] where a 3-stage cascade employs increasingly powerful classifiers starting from linear classifiers to multiple kernel [93] based learners to reject windows in which objects are not present. Both [22] and [94] have been state-of-art detectors in the PASCAL VOC detection task [19]. Generic object detection has also been done using generalized Hough based interest points in [25]. A per-category codebook of generalized Hough based interest points is created along with object-centroid information from the training images, matches during the detection phase vote for likely locations of object centroids. The votes are then summed up into a Hough image and the peaks are then considered to be successful detections. One significant drawback of the state-of-art in Object detection is the need for controlled and well chosen training data like the Caltech-256 [27] which has centered, normalized, single-instances of object categories in images. Real-world scenes with complex background and variations in scale, pose and depth often confuse state-of-art detection methods and furthermore, the detectors are agnostic to key-concepts that are important in scene understanding and recall by humans [86].This thesis attempts to address some of these issues in object detection by fusing eye-gaze information with state-of-art detection in [46]. Encouraging results have been shown in identifying key objects as well as increasing the accuracy and speed of a baseline state-of-art detector [22].

Holistic understanding of a concepts in their inter-relationships can be captured in the form of ontologies such as the popular WordNet [59]. More recent and relevant to image understanding is the Imagenet ontology [15] which contains over a million images arranged according to the WordNet hierarchy. This thesis also explores a semi-automated method of building rich visual concept ontologies that not only encode *part-of* relationships, but also preferential *attentional-bias* between visual concepts, *interaction* information between key objects in an image and also affective information [70]. Our data has also been made publicly available in [71]. Eye-gaze based analysis for Top-down influence of image semantics manifested in inter-object *interaction* and *affect* is also demonstrated in this thesis [70], [46].

2.4 Understanding video content

Digital video has become a ubiquitous way of capturing and storing the myriad variety of events and experiences we encounter. Increasing afford-ability of video capture devices like mobile phones, laptops, surveillance cameras, hand-held video cameras,etc has resulted in massive quantities of video content. For example, upload rates for user-generated videos in *Youtube* are currently topping 24 hours a minute, which in other words could be a complete day's life-blog for an individual. Broadcast and surveillance video are other systems that generate huge amount of video on a hourly and daily basis. Once the basic storage and transmission frameworks are in place, the next and natural challenge is to make sense of the content, inferring *actions*, *concepts* and more abstract notions such as *events* and *affect* remain open challenges and active research topics. Video summarization involves identification of audio-visual cues that Video understanding methods can be categorized based on type of information source into internal methods that use information only from within the video stream, external methods that rely completely on other type of media that is not part of the video stream and lastly hybrid methods that perform a fusion of information sources including the video stream being summarized [61].

A popular scenario for video processing are automatic annotation of video content with concept and event related meta-data. Such annotation then enables improved indexing, and in turn, applications such as video search and summarization [61]. The importance of these tasks can be seen in the increasing complexity and challenging queries in competitions like TRECVID [82] which focuses on quantitative measures such as precision and recall, this complemented by efforts like VideOlympics [85] which also brings evaluates interactivity and visualization of video search systems. Another important video processing task is that of content retargeting to change the resolution, aspect-ratio and bitrate of video content to suite different delivery mechanisms. A popular and recent approach is seam-carving [79] which inserts or removes connected pixel rows and columns to resize image or video content. Another approach is to use low level saliency to decide which pixels to keep and which to remove [78]. Saliency in videos has been explored using low level contrast, gradient and other features [38] and also using motion cues [3]. An effort has been made in this thesis to use eye-gaze based saliency for video annotation using dynamically placed text captions.

2.5 Eye-gaze as a modality in HCI

The use of eye-gaze as an input for operating computing interfaces has been an active topic of interest in the HCI community [41]. Studies on the utility of eye-gaze as an indicator of visual attention have been published by Vertegaal *et al.* [96, 95]. As the eye muscles are one of the fastest in the human body and users often tend to look at a target before initiating manual action. This makes eye gaze one of the fastest inputs available for computer systems to process. Also, since users can produce thousands of eye movements without any fatigue, eye gaze as a communication or control signal can reduce risk of physical strain or injury. To avoid interruptions from ubiquitous computing devices and to enable computers to communicate better with humans, Vertegaal proposed Attentive User Interfaces (AUIs) [97]. Key properties of AUIs include sensing, reasoning and communicating user attention, determining user availability for interruption and augmenting user attention by magnifying regions of interest while attenuating peripheral detail. Example of an AUI is interactive attentive art [30], where plasma screens displaying artwork in museums can highlight areas receiving user attention while darkening areas receiving little attention. AUIs are particularly useful in reducing the cognitive load of visual information on large displays, as they dynamically filter user interest information.

3 Experimental protocols and Data pre-processing

This chapter describes experimental methodology used for different experiments in this thesis, it covers different ways in which input data is captured as visualized in Figure 9. Section 3.1 describes the data collection involved in the problem of interestingness discrimination. The subsequent section 3.2 describes data collection, preparation for eye-tracking experiments.

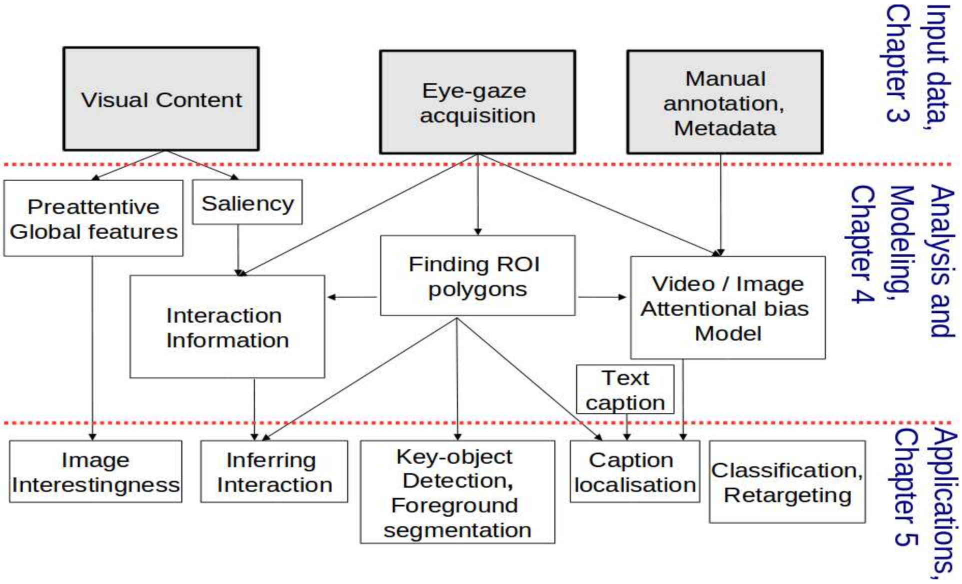


Figure 9: The figure highlights the scope of this chapter in the overall schema for the thesis, input data is captured via Image/Video content, eye-tracking, manual annotation.

3.1 Experiment design for pre-attentive interestingness discrimination

Rapid scene understanding in the absence of attention is explored first in this thesis, these experments do not involve eye-tracking and are behavioural experiments involving human responses to well chosen iamge stimuli.

3.1.1 Data collection

Image data is crawled from the Flickr ®collection based on keywords relevant to selected semantic categories. Semantic categories are drawn from literature and the ones that are well represented in Flickr are chosen. Using Flickr's public API, we queried images with keywords belonging to one of 14 categories, 7 natural, 7 man-made as per [101] and [64]. The chosen categories are, *beach, city-view, coast, field, forest, high-building, highway, indoor scene, man-made object, mountain, natural object, portrait, street*. A list of keywords is created by using a *bag of words* approach using synsets from WordNet [59]. The control set of images was downloaded using descending order of *relevance* as the ordering criterion, and the interesting set of images was downloaded using descending order *interestingness* as the sorting priority. The relevance order and interstingness order are computed within the Flickr system and the rank order are treated as ground truth. In total, we downloaded 9,137 interesting and 16,244 relevant images. Table 2 shows the bag-of-words approach for a few example categories along with the number of images retrieved and the number of images that were concluded to be noise and had to be purged.

We find that subjectivity of users introduced lot of noise in publicly available social networked media as the words relating to a concept are often used in different ways, for example "woods" is frequently used as a name for people, buildings, etc. Such images were filtered out manually. During the experiment it was also found that the nature of images is also of concern as it could be offensive or unpleasant to users. Details of the images collected and the bag-of-words for 5 representative image categories is shown in Table 2.

Category	Bag of words	Number of Images	Purged images
Forest	*woods timberland woodland timber grove jungle*	531	63
Mountain	*mount highland hill ridge alp volcano peak*	679	110
Field	*clearing grassland crop harvest paddy cultivation*	623	76
Beach	*shore plage sand seaside*	500	76
Indoor Scene	*interior bedroom office dining kitchen library*	753	115

Table 2: Details of Flickr images collected for 5 of the 14 image themes chosen.

The contribution of global and local information towards pre-attentive discrimination of interestingness is investigated by presenting manipulated versions of images to users and recording user decisions on whether they find the image interesting. An example manipulation is illustrated in Figure 10, (a) shows the intact image, (b) global order in the image is removed by scrambling image blocks, (c) removal of colour information and (d) removal of local information by blurring the entire image. This protocol has been followed in [101] to study the influence of global and local information in image categorization by humans. Image pairs belonging to a category, one rated as highly aesthetic and another as ordinary, are retrieved from the image database and one of the three manipulations mentioned, is applied to both.

The pair is then shown in succession, in random order for the same time (time spans ranging from 50 to 1000 milliseconds were experimented with). This is illustrated in Figure 11. A forced choice input then records the user decision about which of the two images is found

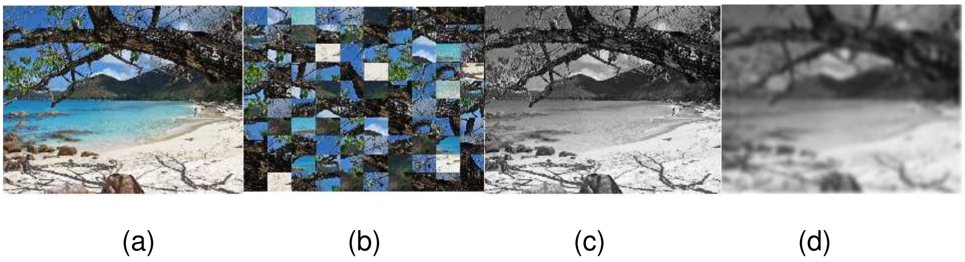


Figure 10: Illustration of image manipulation results, (a) intact image, (b) scrambling to destroy global order, (c) removal of color information (d) blurring to remove local properties, the color is removed as well as it can contains information about global structure in the image.

to be interesting. The forced choice ensures input for rapid image presentations as we find that users might feel they cannot discriminate between images in a pair, but are unaware that they are capable of making decisions that are statistically significant.

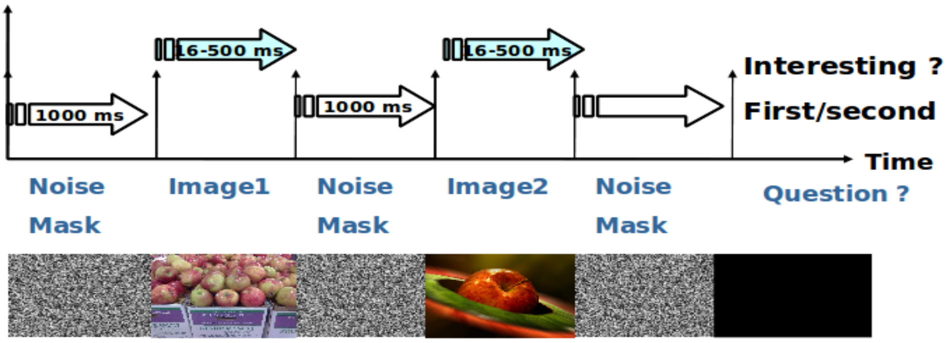


Figure 11: The short time-span image presentation protocol for aesthetics discrimination is visualized here, an image pair relevant to the concept *apple* is presented one after another in random order. The presentation time for each image in the pair is same and chosen between 50 to 1000 milliseconds. Images are alternated with noise masks to destroy persistence, a forced choice input records which of the rapidly presented images was perceived as more aesthetic by the user.

Once the image pairs are presented once in succession at a short time span, ranging from 50 to 1000 milliseconds, each pair is presented side-by-side again later as illustrated in Figure 12. The pair is then displayed for as long as the user needs to decide on which of the two is more aesthetic. The user is allowed to reject image pairs which do not seem relevant to the same concept or are difficult to discriminate based on aesthetic value. We reject such image pairs for which the subjects are undecided about interestingness.

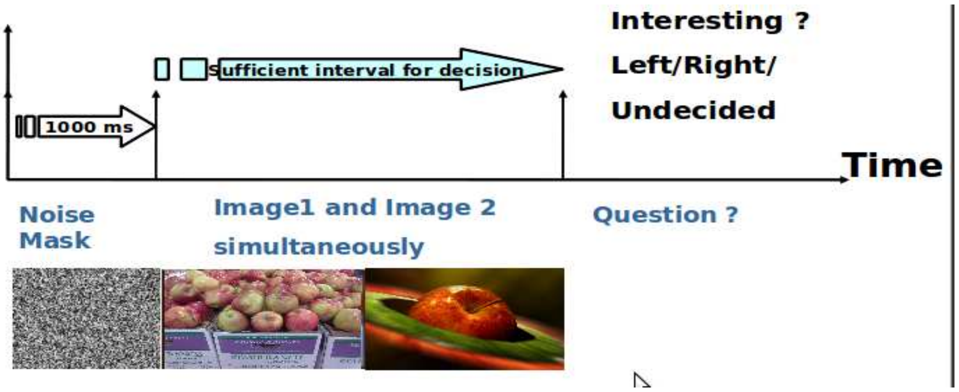


Figure 12: The long time-span image presentation protocol for aesthetics discrimination is visualized here, an image pair relevant to the concept *apple* is presented side-by-side. The stimulus is presented as long as the user needs time to decide whether an image clearly has more aesthetic value than the other.

The data collection and experiment method described above is used to record human response times and decision error rates and subsequently to model the role of global and local image information in rapid image aesthetics discrimination; Sec. 4.1. The following section describes data collection methodology for eye-tracking based experiments.

3.2 Experiment design for Image based eye-tracking experiments

The objective of the next set of experiments is to acquire sufficient eye-gaze patterns for each of the diverse semantic image categories, so as to reliably conclude about human visual attention characteristics. Since many factors like the task on hand and subject's professional and behavioral background influence attention during human-image interaction, we need to carefully address them while designing our experiments. The experimental data collected during the course of this thesis is made publicly available as the NUSEF [71] dataset. The data is also used to quantify the preferential attention to selected objects and actions in the scene termed as *attentional bias* in this thesis.

3.2.1 Data collection and preparation

We chose a diverse and representative set of images with 1024×768 resolution, close to 800 gray-scale and color images from publicly available data sets for our experiments. The images contain scenes and objects captured at varying scale, lighting conditions and viewing profiles. From among the many images acquired using descriptor tag-based internet search (where we employed synonym-based query expansion), the experimental image set was selected based on quality, resolution and aspect ratio constraints. Images include everyday scenes from Flickr, aesthetic content from Photo.net, Google images and emotion-evoking IAPS pictures [50]. Semantic categories include (i) indoor and outdoor scenes, (ii) close-up and mid-range human and mammal faces, (iii) *portrait* images showing face and torso of humans and mammals, (iv) images containing multiple humans/mammals, (v) reptiles, (vi) injury and blood, (vii) nudes, (viii) *world* images containing living beings and inanimate objects (*sky*, *sand*, *building*, *etc.*) and (ix) images depicting action (*look*, *read*, *etc.*). Some exemplar images used in our experiments are presented in Fig.13, these illustrate the diversity of our dataset.

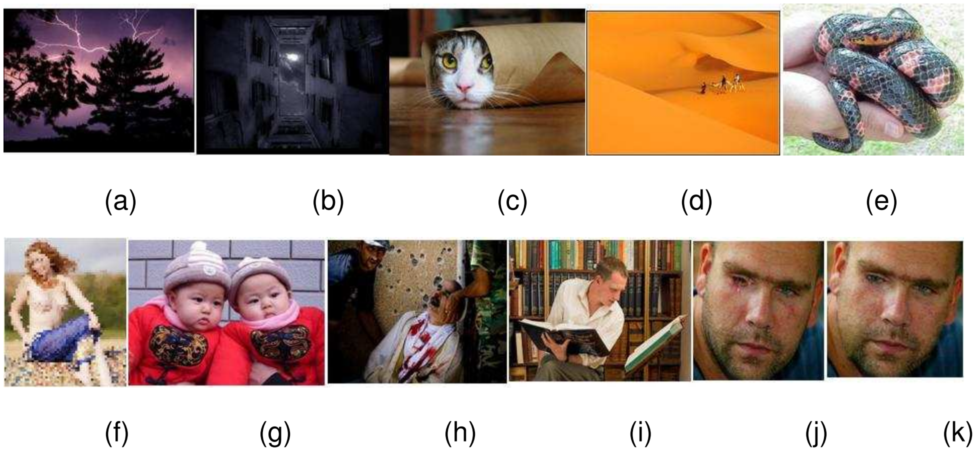


Figure 13: Exemplar images corresponding to various semantic categories. (a) Outdoor scene (b) Indoor scene (c) Face (d) *World* image comprising living beings and inanimate objects (e) Reptile (f) Nude (g) Multiple human (h) Blood (i) Image depicting *read* action. (j) and (k) are examples of an image-pair synthesized using image manipulation techniques. The damaged/injured left eye in (j) is restored in (k).

Furthermore, to study changes in visual attention due to the addition/deletion of *interesting* objects to/from an image, we synthesized a number of image-pairs where affective objects are added or removed using an image editing tool Figure 13(j),(k). Images from Flickr and Photo.net are representative of amateur and semi-professional photographs and are rated by respective user communities and include popular themes such as *landscape*, *urban scenes*, *portrait* and *personal events*. Google images are chosen to get popular images worldwide for our query concepts. IAPS capture the canonical emotions such as *fear*, *anger*, *happy*, etc. Together, these images cover the popular gamut of image categories that a human user or an automated image processing system may encounter. Furthermore, the image categories are also a combination of many prior studies on scene understanding [100], aesthetics rating [14], person detection

[13] and image retrieval [69].

3.2.2 Participants

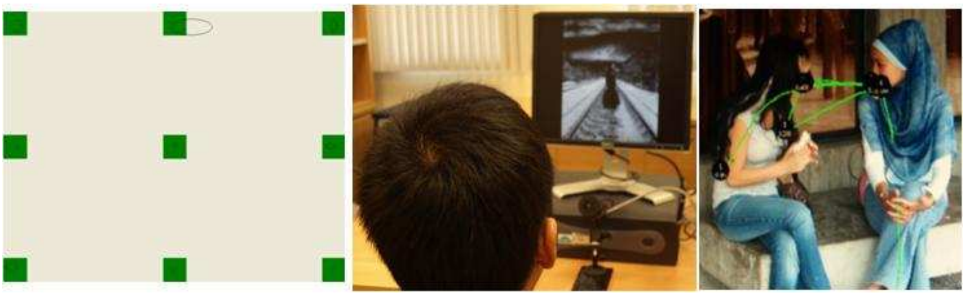
Over 75 subjects including research staff, graduate and undergraduate students, aged between 18-35 years (μ =24.9, σ =3.4) were recruited for our visual attention experiments. All participants had normal or corrected to normal eyesignt and were paid a token fee for participation.

3.2.3 Experiment design

To avoid task-based priming of visual attention, subjects were simply asked to view a set of images with no specific objective. This was done to ensure that any observed attentional-bias is exclusively due to underlying image semantics. 350 randomly chosen images were presented for subject viewing, over two passes separated by a 10 minute break to avoid fatigue. Each image was presented for 5 seconds followed by a gray mask for 2 seconds to destroy image persistence.

3.2.4 Apparatus

We used the Erica eye-tracker for our experiments. The system consists of an infra-red sensing camera, placed alongside the computer monitor, at about 30 inches from the subject. Images were presented using a 17" LCD monitor with a screen resolution of 1024 x 768 pixels (96 dpi). Upon 9-point gaze calibration, the eye-tracker is accurate within the nearest 1^o visual angle at 3 feet viewing distance, translating into an error radius of 5-10 pixels on screen. A dimly lit, sound-proof room was used for the experiments to avoid distractions. The screen coordinates representing locations gazed by subjects over time are sampled at 30 Hz, and processed to generate a sequence of fixation points. Each fixation point represents a screen location where the eye-gaze remains within 2^o visual angle for at least 100 milliseconds. Illustrations depicting the calibration, data collection and visualisation of eye-tracking data are shown in Fig.14 (a),(b) and (c) respectively.



(a) (b) (c)

Figure 14: Experimental set-up overview. (a) Results of 9 point gaze calibration, where the ellipses with the green squares represent regions of uncertainty in gaze computation over different areas on the screen. (b) An experiment in progress. (c) Fixations patterns obtained upon gaze data processing.

3.2.5 Image content

The NUSEF database was compiled from images that were viewed by at least 13 subjects (containing a minimum of 50 fixations). Table 3 presents NUSEF's semantic category-based image distribution, while Table 4 compares our database to eye-tracking data in [43] and the *Fixation in Faces* dataset [11]. 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

Table 3: Image distribution in the NUSEF dataset, organised according to semantic categories.

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

ground truth for image understanding applications.

Figure 15 shows the fixation patterns for various semantic image categories. Fixations are denoted by circles of varying sizes and gray-levels. The circle sizes 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 (Figure 15(a,b)). As also noted in [43], fixations are observed around specific regions like the eyes, nose and mouth for *faces* (Figure 15(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

Table 4: A brief comparison between datasets in [43], [6] and [11] with NUSEF [71].

Database	# images	Average # viewers per image	Semantics	Remarks
MIT [43]	1003	15	Everyday scenes from *Flickr* and *LabelMe*	Fixations are found around faces, cars and text. Many fixations are biased towards the center.
CalTech [11]	303	8	Colour and grayscale images of faces	Fixations are predominantly centred around faces and parts therein.
NUSEF [71]	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.

the lower half in highly expressive (*angry*, *surprise*, *disgust*) faces ((Figure 15(d)) (fixation statistics in [70]).

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 mammal *portraits* (Figure 15(i,j,k)), whereas most fixations occur on the body for *nudes* (Figure 15(l)). *Action* images (Figure 15(g,h)) are characterized by frequent fixation transitions between interacting objects, with more

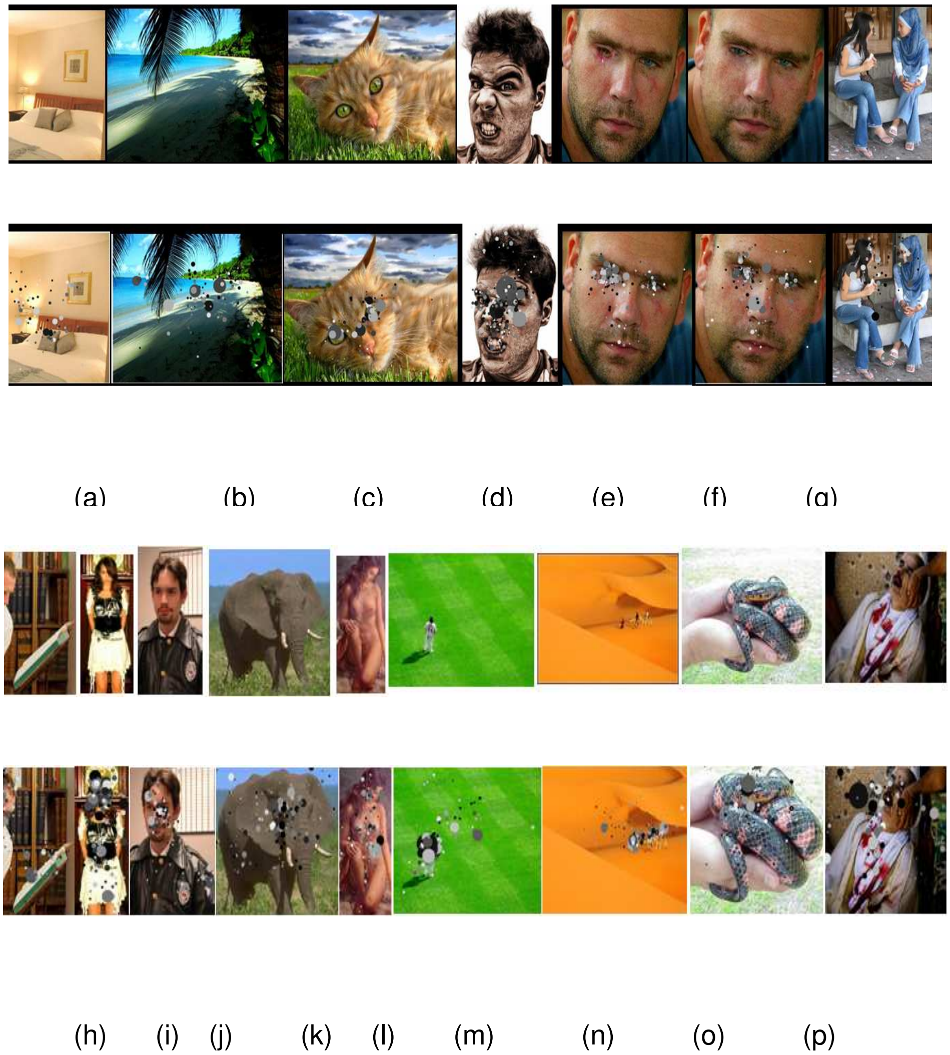


Figure 15: Exemplar images from various semantic categories (top) and corresponding gaze patterns (bottom) from NUSEF. Categories include Indoor (a) and Outdoor (b) scenes, *faces*- mammal (c) and human (d), *affect-variant group* (e,f), *action-look* (g) and *read* (h), *portrait*- human (i,j) and mammal (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.

transitions occurring from the *action recipient* to the *action source* [70] (*e.g.* Man and book are *action source* and *recipient* respectively in Figure 15(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 Figure 15(e) attracts the most attention, while the fixation distribution is more typical when the missing eye is replaced using image manipulation techniques in Figure 15(f). Fixations are observed around living beings in *world* images Figure 15(l,m), as well as unpleasant concepts such as *reptile* (Figure 15(o)) and *injury* (Figure 15(p)).

3.3 Experimental procedure for video based eye-tracking experiments

Two types of experiments have been done in this thesis with video stimuli, the first is with a standard non-interactive setup involving a calibration step followed by a session where 3 subjects watch videos while their eye-movements are tracked. Three chosen clips *meeting.avi*, *friends.avi* and *sports.avi* have 350, 500 and 400 frames respectively. The first two clips have social activity with multiple participants and the third is fast paced sports. This information is then available for offline analysis, this is akin to experiments described earlier with images. Experiments done here are also similar to those in [35] where 3 human subjects view 15 video clips of natural scenes, and [9], where video clips are cut into snippets and concatenated together to minimize semantic relatedness across snippets combined together. Though similar in spirit, these experiments were done with the objective of identifying how image information drives HVA. On the other hand, this thesis focuses on accurate and efficient extraction of ROI meta-data from eye-gaze information and subsequent use in applications. We used the Erica eye-tracker for these experiments. As described earlier, the system consists of an infra-red sensing camera, placed alongside the computer monitor, at about 30 inches from the subject. Videos were presented using a 17" LCD monitor with a screen resolution of 1024 x 768 pixels (96 dpi). Upon 9-point or 12-point gaze calibration, the eye-tracker is accurate within the nearest 1^o visual angle at 3 feet viewing distance, translating into an error radius of 5-10 pixels on screen. The screen coordinates representing locations gazed by subjects over time are sampled at 30 Hz, and processed to generate a sequence of fixation points. Each fixation point represents a screen location where the eye-gaze remains within 2^o visual angle for at least 100 milliseconds. This is similar to free-viewing experiments with image stimuli, the major difference is that eye-gaze is now recorded as time series information, with sampling frequency (30 Hz), which is quite close to the video frame rate. This raises challenges for eye-gaze analysis, and is dealt in Chapter 4.

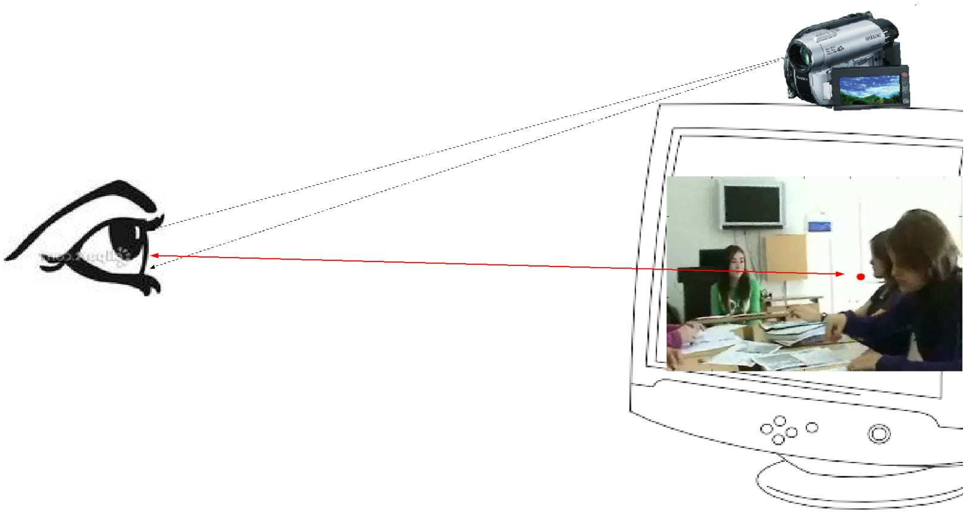


Figure 16: Illustration of the interactive eye-tracking setup for video. (a) Experiment in progress (b) The subject looks at visual input. (c) The on-screen location being attended to. (d) An off-the-shelf camera is used to establish a mapping between images of the subject's eye while viewing the video.

The second type of experiment was to explore a closed loop scenario where eye-gaze information is utilized on-the-fly to manipulate video stimulus being seen by the subject. Low cost eye tracking using open source software and webcams or handycams were explored in the second case. Eye-tracking was done with video clips of 2-5 minute duration, from the theme *Television Sitcoms*. The clips depict individuals and groups of people interacting in a social setting. The clips have *semantically rich* events involving *people* and *objects*. The video clips were resized while preserving the aspect-ratio and displayed at a resolution of 1024×768 pixels on-screen. Figure 16 illustrates the experimental setup for interactive eye tracking, wherein eye-gaze information is acquired and analysed on-the-fly.

3.4 Summary

This chapter describes the data collection protocol including selection of visual stimulus, hardware and software used and psychophysics experiments. Image and video data is chosen in this thesis, to reflect important themes that humans experience. Visual data is collected primarily from publicly accessible sources and using as much automation as possible to counter biases in selection 3.2.1. The experimental setups for data collection also include very low cost hardware and software as described in 3.3 . This is done to explore scenarios where eye-gaze based technology can naturally be made part of existing computing devices such as laptops, cameras and phones. The data collection and experimental methods are geared towards generating results that hold over a diverse collection of visual content.

4 Developing the framework

The grand goal of automatically understanding semantics, aesthetics and affect represented by images and videos is made challenging by the diversity of visual attributes and possibilities of composing different visual elements. Pixel-based information in images can represent rich semantic information relating to objects and their interactions as can be seen in Figure 18 (e) *cat reads book* and (f) *paper clip scene*. Inferring semantic, aesthetic and affective content are basic tasks that most humans are able to perform effortlessly and continuously and yet pose a huge challenge for automated methods. The difficulty arises from the diverse possibilities of semantics, abstraction and affect that can be depicted by an image. Figure 18 depicts how pixel content can depict something as simple as (a) simple texture (b) aesthetic value from colour and depth (c) strong visual patterns (d) symmetry and composition and (e)(f) depicting abstract semantics. This chapter introduces the analysis and modeling visualized in Figure 17, which then support applications on visual content understanding. The current chapter first explores the notion of aesthetics in natural images and presents evidence for aesthetics discrimination in early vision using psychophysical experiments. The role of global color information is also highlighted using results from behavioral experiments and content based modeling. Human bias to selected objects and actions in scenes is modeled as the attentional bias, using statistical analysis over eye-tracking data.

Local image information contained in objects and their relationships

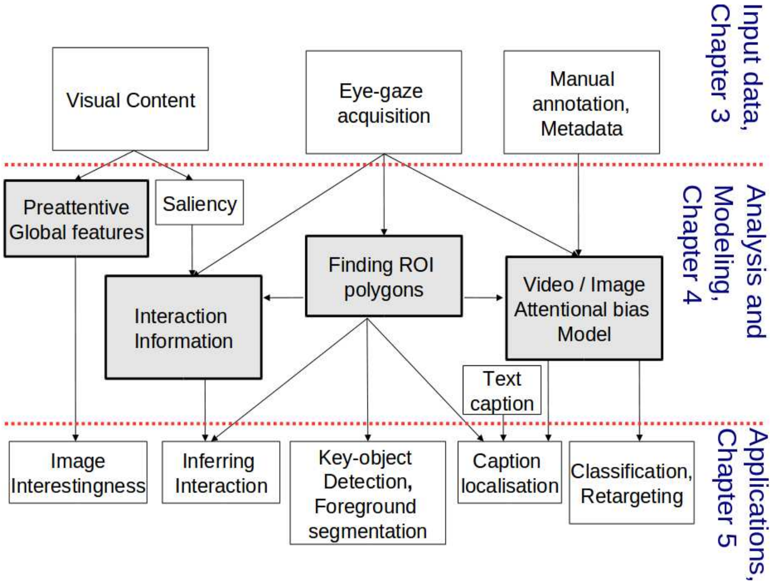


Figure 17: The schema visualizes the overall organization of the thesis and highlights the components described in this chapter. The current chapter deals with analysis and modeling of visual content, eye-gaze information and meta-data.

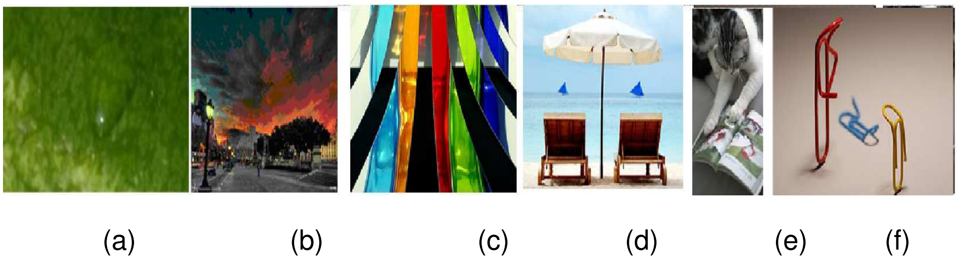


Figure 18: The panel illustrates how the arrangemnt of different visual elements in images can give rise to rich and abstract semantics. Beginning from simple texture in (a), the meaning of an image can be dominated by low level cues like color and depth in (b), shape and symmetry in (c) and (d). The unusual interaction of *cat* and *book* gives rise to an element of surprise and rich human interaction and emotions are conveyed through inanimate *paper-clips*.

in images also influences image aesthetics. Semantic and Affective information is investigated in using image context, key concepts and interactions. Eye-gaze analysis measure for semantics and affect related analysis are introduced. Insights from analysis of eye-gaze on manually and automatically annotated images are used to construct an *Attentional bias* model that capture preferential attention to important concepts and possible interactions or relationships between them. The model is then used to illustrate the close correspondence that key concepts and actions in images have with text descriptions of the image.

4.1 Pre-attentive discrimination of interestingness in the absence of attention

This section brings forth the human capability to identify aesthetically pleasing images even when the image stimulus is presented for a brief time span. The influence of early-vision related global information processing for aesthetics discrimination is also demonstrated. Furthermore, attributes such as *depth*, *lighting* and *saturation* that are related to early vision are captured directly or indirectly in some fields of the Exchangeable image file format (EXIF) meta-data. We exploit EXIF information to model some of the pre-attentive global image attributes and find them to be useful for aesthetics discrimination. Our notion of *aesthetically pleasing* in the current context is that of *interestingness* as defined in the Flickr ®system. *Interestingness* indicates whether an image would be interesting to view amongst others that are relevant to a particular semantic concept [7]. This has been modeled in Flickr by recording user interaction and recommendation statistics for images. An example image pair is shown in Figure 19, though both are relevant to the concept *apple*, Figure 19 (b) has a much higher *interestingness* score and is preferred by Flickr users over Figure 19 (a). Experiments designed for interestingness discrimination involve a significant component of psychophysics, psychophysics is a branch of psychology which quantitatively investigates relations between physical stimuli and the sensations and perceptions they evoke. The aim is to explain and model the processes underlying perception. Initial foundations were laid by the seminal work of the German psychologist Fechner [20] and subsequently his students.

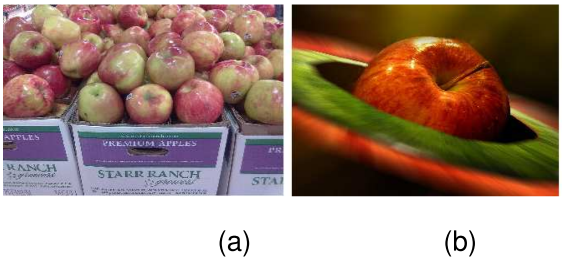


Figure 19: Image on the left is a relevant result for the query concept *apple*. The image on the right illustrates an image for the same concept that has been viewed preferentially in the Flickr database.

We investigate the following problems related to aesthetics in images,

- Can interesting images be discriminated rapidly in pre-attentive time spans ?
- What role does global and local information play in determining image interestingness ?

Though human visual attention is an imporatnt part of scene perception, meaningful eye-fixations usually begin after 100 milliseconds. The time period between stimulus onset and the first fixation is called the pre-attentive time span. Global information processing and rapid scene understanding has been demonstrated during this phase [52]. Global information has been shown to be effective for natural scene categorisation [101][64] and also an important precursor for subsequent visual exploration of the scene via eye fixations [89].

The aesthetic value of an image can be influenced by both bottom-up or top-down visual features. Figure 18 illustrates how bottom-up cues such as color or elements of symmetry can make an image aesthetically pleasing. On the other hand. Figure 18 (e) illustrates top-down cues such as the unusual juxtaposition of concepts *cat, book* and actions *look* and (f) shows how objects can be used as metaphors to depict complex events or emotions in aesthetic images.

Aesthetics may not be well captured in simpler meta-data such as user generated tags or other text content surrounding images on the web. As current image search relies heavily on such textual meta-data, significant mis-match can be observed between system generated ordering for relevant images against ordering by user's notion of *interestingness*. This is illustrated in Figure 20, which visualizes the lack of correlation in ordering images from Flickr®for the theme *beach* when done based on user interaction data, as compared to mere textual relevance. Semantic relevance ranking for 2132 images is plotted against their ranks based on *interestingness*. The images are grouped into high, medium and low ranked images according to *interestingness*. The clear lack of correlation brings forth the fact that text based semantic relevance may not capture interestingness in an effective manner.

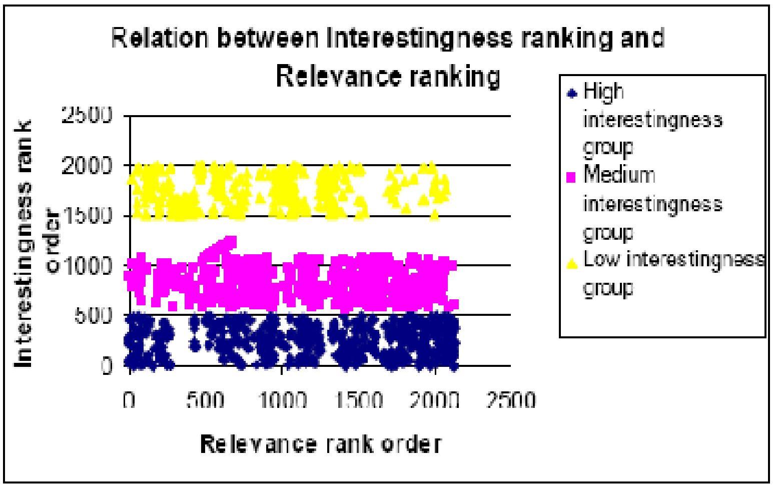


Figure 20: A visualization of the lack of correlation between image ordering based on mere semantic relevance of tags Vs *interestingness* in the Flickr system. Semantic relevance based ordering of 2132 images is plotted against their ranks on *interestingness*. This illustrates the need for methods that can harness human interaction information.

4.1.1 Effectiveness of noise masks in destroying image persistence

Noise masks used in the two experiments ensure that image persistence on the eye does not influence the results for short presentation of images. This can significantly alter the results for short presentation spans. This is illustrated in Figure 21 for a user across different presentation time spans. An overall increase in agreement can be seen when the noise mask is absent. This is made possible by the persistence of images in the visual system after the presentation stimulus is removed.

The discrimination accuracy is measured as the agreement between user decisions made at short-term presentation against a longer-term presentation of the image pair. The decision made on short-term presentation,

$$decision_{i,short} = \begin{cases} 1 & , \ first \\ 0 & , \ second \end{cases} \quad (1)$$

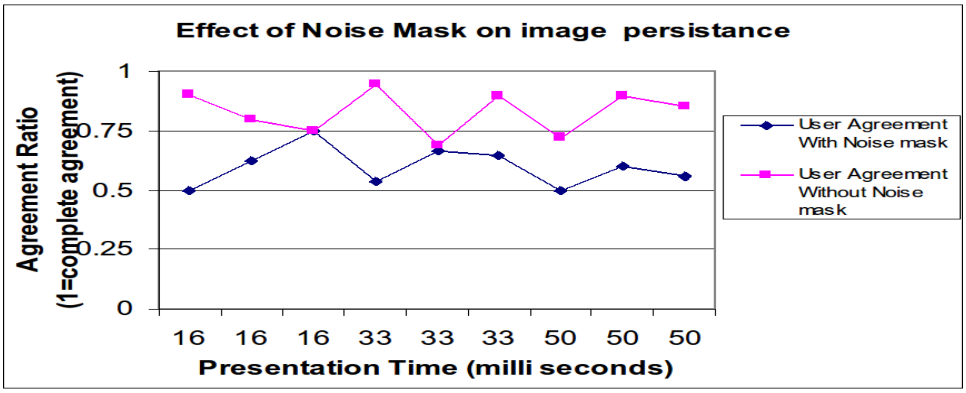


Figure 21: Impact of noise masks in reducing the effect of persistence of visual stimulus. The two plots are agreement between user-decisions made for short term and long term image pair presentation. It can be seen that image persistence in the absence of the noise mask significantly increases the overall discrimination capability of the user.

and long term,

$$decision_{i,long} = \begin{cases} 1 & , \ left \\ 0 & , \ right \\ -1 & , \ undecided \end{cases} \qquad (2)$$

The short-term to long-term decision agreement is modeled as,

$$agreement = \frac{\sum_i \{decision_{i,short} \ = \ decision_{i,long}\}}{\sum_i decision_{i,long} \neq -1} \qquad (3)$$

The agreement is measured only over trials that are not rejected as undecidable for long-term presentation $decision_{i,long} \neq -1$. The short term presentation is forced choice and can take only two values, the participant can also reject an image pair in long term presentations using a third key-press. Longer presentation spans result in more reliable decisions as compared to short-term presentations, this can be seen in Figure 22 which plots the agreement scores for image-pairs as the short-term presentation time is increased from 50 milliseconds to 1000 milliseconds and Figure 23 The short term presentation is kept as a forced-choice experiment, as participants tend to be very unsure of their short term decisions and are likely to reject most trials. Long term decisions on the other hand, give sufficient time to accept the left or right image as more aesthetic or to reject the image pair.

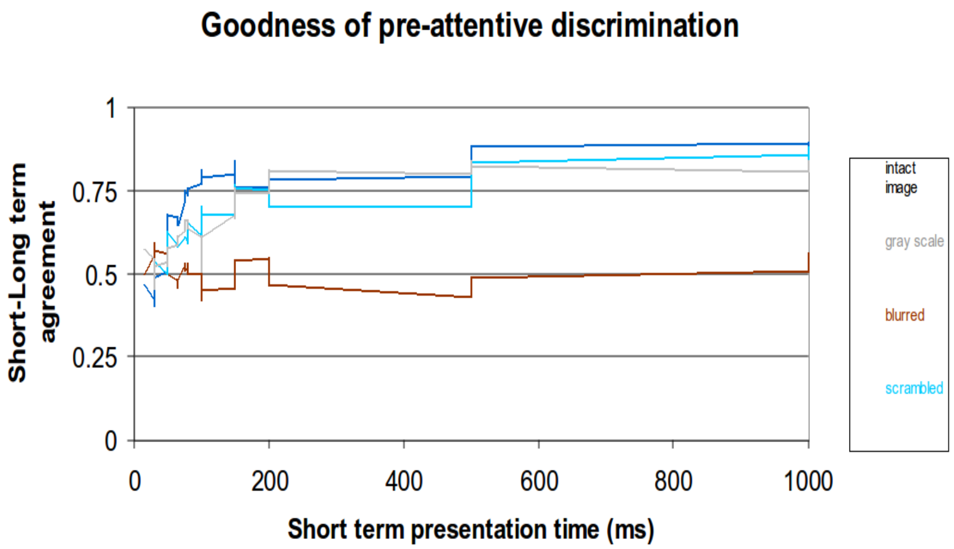


Figure 22: Improvement in user discrimination as short-term presentation span is varied from 50 milliseconds to 1000 milliseconds. As expected, users make more reliable choices amongst the image pairs presented. A presentation time of about 500 milliseconds appears to be the minimum threshold for reliable decisions by the human observer and can be used as a threshold for display rate for rapid discrimination of interestingness.

The agreement value stabilizes for most users by 500 milliseconds and they are able to correctly identify the more aesthetic image from a pair. Pre-attentive decisions made between 30 to 50 milliseconds have agreement values over 75% with those made over longer presentation times, this is well over chance performance. Statistical significance using the binomial test also indicates that short-term decisions made by 50 milliseconds onwards, agree with the long term decisions at $p = 0.01$. Binomial test was done by modeling agreement and non-agreement of short and long term decisions as two outcomes of a coin toss. The significance test is then done to essentially find out

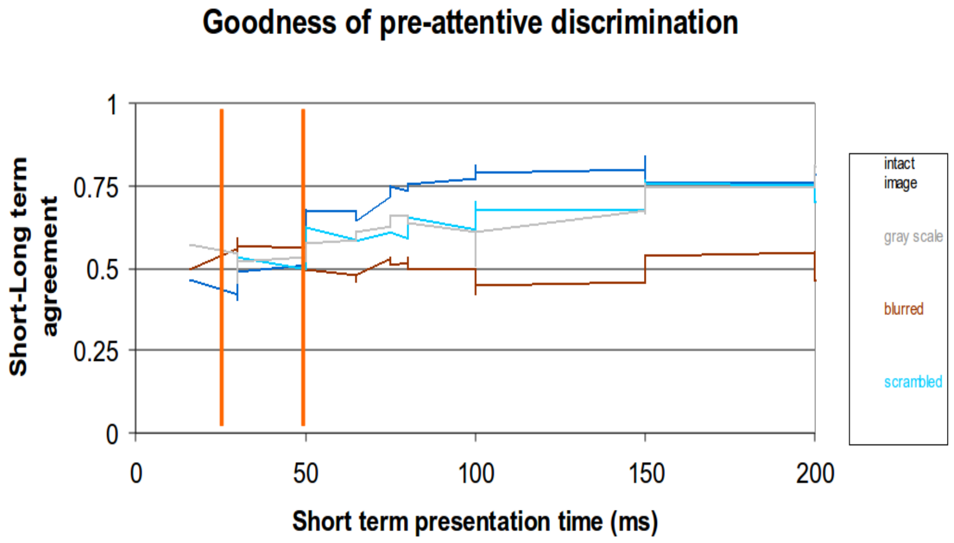


Figure 23: Improvement in user discrimination as short-term presentation span is varied from 50 milliseconds to 200 milliseconds. A binomial statistical significance test reveals agreements between short and long term decisions starting from 50 millisecond short term decisions.

if the coin is biased in favour of agreement. The wide variation at 16 milliseconds indicates lack of discrimination at very short time spans. High values at 50 milliseconds are followed by a drop at 100 milliseconds before converging to the steady or higher value beyond 500 milliseconds. This could indicate different cognitive process responsible for short-term and longer-term discrimination.

Pre-attentive time spans are believed to be of the order of 100 milliseconds or lesser, from presentation of stimulus. The contribution of color towards rapid, aesthetics discrimination is shown in our our experimental results 24. There is a marked drop of 20 % in the discrimination capability as color information is removed from images. Global information also contributes about the same, this can be seen from the drop of about 15 % in short-term to long-term agreement when global information is destroyed by scrambling image blocks. Another interpretation of agreement of short-term decisions made at 100 millisecond presentation with long-term decisions is shown in Figure. 25. Though loss of colour information or loss of global order in the image result in a similar drop of about 7% in agreement, removal of local information reduces agreement significantly by more than 20%, this is surprising as literature suggests a dominant role of global information in pre-attentive time spans.

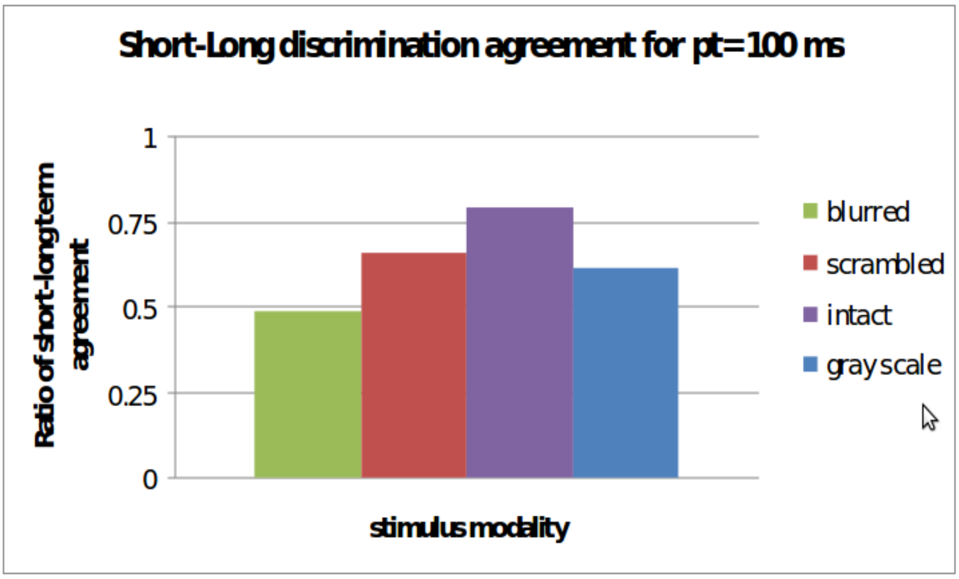


Figure 24: The panel illustrates changes in pre-attentive discrimination of image interestingness as image content is selectively manipulated. Removing color channel information results in a 20 $\%$ drop in discrimination capability. A drop of about 15 $\%$ in short-term to long-term agreement when global information is destroyed by scrambling image blocks

To summarize, the important results contributed from above-mentioned experiments are,

- Color information plays an important role in rapid discrimination of interestingness in images as shown in Figure 24.
- The minimum time required for such rapid decisions is between

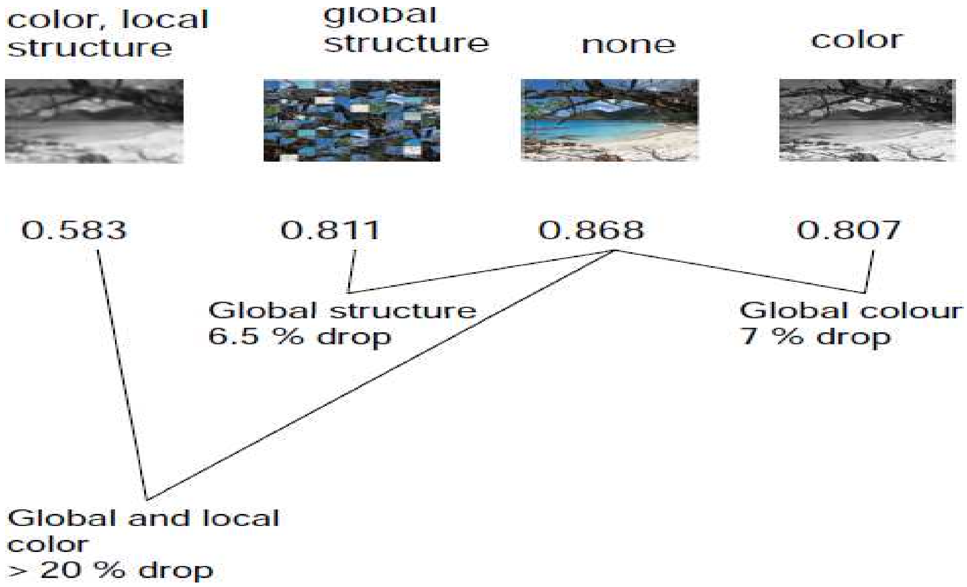


Figure 25: Agreement of short-term decisions made at 100 millisecond presentation with long-term decisions. Though loss of colour information or loss of global order in the image result in a similar drop of about 7% in agreement, removal of local information reduces agreement significantly by more than 20%, this is surprising as literature suggests a dominant role of global information in pre-attentive time spans.

200-300 milliseconds as visualized in Figure 22. This also sets the fastest refresh rates for such systems to be 3-4 Hz. These results on global feature based interestingness hold for a diverse set of image categories such as those in Flickr®.

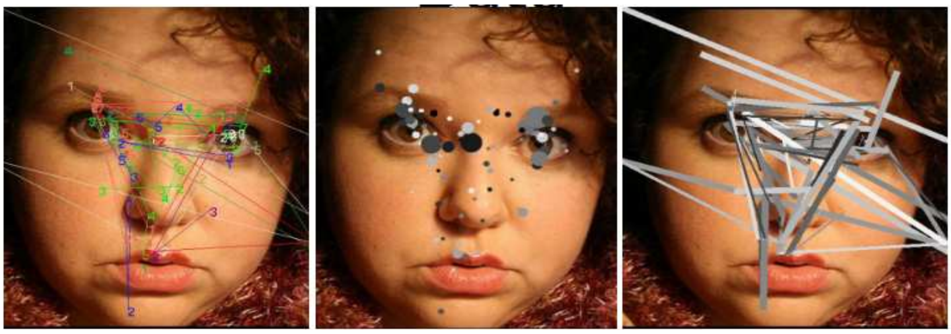
- Though prior research has shown how local and global features influence long term scene categorization, this is the first work to address a more abstract problem of predicting whether an image is interesting or not.

4.2 Eye-gaze, an artifact of Human Visual Attention(HVA)

This section explores the use of eye-gaze to systematically infer scene semantics. The human visual pathway responsible for processing mid-level information such as shapes and increasingly abstract information relating to objects, inter-object interactions, affect and even aesthetics. HVA is an important cognitive mechanism to selectively process portions of the visual abundant visual information available. Eye-tracking allows the visual system to inspect Regions of interest (ROIs), which in turn enables subsequent understanding and inference. In this section, Eye-tracking experiments and content based modeling is used to gain insights into how humans understand semantics, affect and aesthetics in Images. The eye-tracking experiments provide spatio-temporal data indicating where and in what order are Image locations processed by humans. Eye-movements consist of *fixations* during which the eye is relatively stationary, transition from one location of fixation to another is made by a ballistic movement called the *saccade*. Fixations are typically longer than 100 milliseconds and tend to remain within 1 degree of visual angle subtended by the eye. Fixations indicate the processing of detail in the fixated location, there is no visual information intake during saccades. Eye fixation data is spatio-temporal, time series information. Different visualisations of eye-tracking data are shown in Figure 26 for a *human portrait* image.

A caveat at this point is that other cognitive processes like peripheral vision, that are parallel and simultaneous to HVA exist. These enable context estimation [57] and in-parallel, rapid processing of some important stimuli such as faces and animals [52].

This section demonstrates the specificity with which humans attend to key objects and their relationships in images and how this is man-



(a) (b) (c)

Figure 26: Different parameters extracted from eye-fixations corresponding to an image in the NUSEF [71] dataset. Images were shown to human subjects for 5 seconds. (a) Fixation sequence numbers, each subject is color-coded with a different color, fixations can be seen to converge quickly to the key concepts (*eye*,*nose+mouth*) (b) Each gray-scale disc represents the fixation duration corresponding to each fixated location, gray-scale value represents fixation start time with a black disc representing 0 second start time and completely white disc representing 5 second fixation start time.(c) Normalized saccade velocities are visualized as thickness of line segments connecting successive fixation locations. Gray-scale value codes for fixation start time.

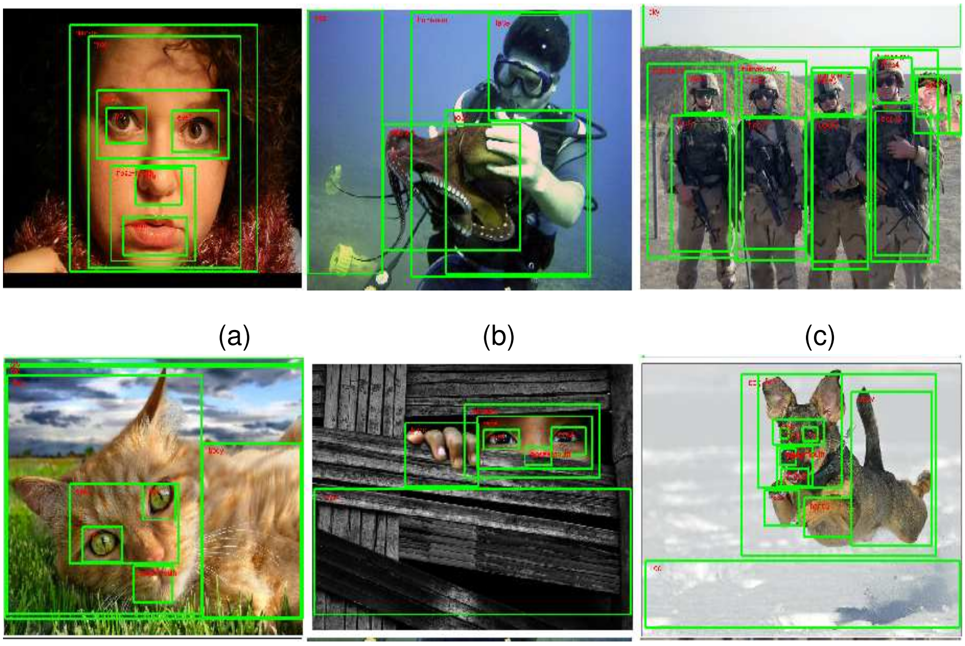
ifested in eye-tracking data. Appropriately defined measures over eye-fixation information are used to discover the *Attentional bias* that humans show while understanding natural scenes with animate and inanimate objects in them. Fusion of eye-gaze analysis with an object-part ontology over our diverse dataset yields quantitative *Attentional bias* values that are consistent across different subjects. The experimental insights are then exploited to extract ROIs and their relationships, abstract notions of affect and aesthetics are also investigated. ROIs discovered by the *binning* method are put to use in applications for interactive object segmentation, key-object detection. An extension of the method also helps in identifying strongly interacting or related visual elements. A hybrid approach using eye-tracking and content analysis is used to classify images as affective, aesthetic or those having noticeable interactions between their elements.

4.2.1 Description of eye-gaze based measures and discovering Attentional bias

Most semantically rich images can be represented using a number of regions termed regions-of-interest (*ROI*s), with each ROI denoting a unique semantic concept. Eye-fixations occur preferentially over important concepts, this phenomenon was verified systematically and also used to quantify the extent this bias and its consistency over human subjects. The NUSEF [71] dataset was then analysed to answer the following questions,

- What are the frequently identified key concepts and sub-parts in themes represented in an image collection such as NUSEF [71] dataset ?
- What are the preferential attentional bias values for such key concepts ?
- Does this model tell something about interactions between concepts in an image ?
- What does the *part-of* ontology for concepts in this dataset look like and what does it tell us about how images are understood by humans ? ie; What is the *world model* for this dataset ?

A group of paid volunteers were asked to manually annotate between 5-10 key concepts in images, this involves indicating bounding boxes and semantic labels as shown in Figure 27. These volunteers did not participate in subsequent eye-tracking experiments. The annotation was done using the Fixplot tool which is part of the Eyenal fixation data analysis software from ASL. The annotators were also instructed to additionally label any clearly visible sub-parts of the key-concepts. That way a labeled *face* would also have *eye* and *mouth* regions labeled, if they were clearly visible. This can be seen in Figure 27 (a)(d)(e)(f), where as the annotators omitted *eye* and *mouth* labels for (b) and (c). Manual annotation can be noisy and this can be seen in the large bounding box for *face* in (d) and smaller than the true size in (e).



(a) (b) (c)

(d) (e) (f)

Figure 27: Visualization of manual annotation of key concepts and their sub-parts for the NUSEF [71] dataset. The annotators additionally label any clearly visible sub-parts of the key-concepts. That way a labeled *face* would also have *eye* and *mouth* regions labeled, if they were clearly visible. This can be seen in Figure 27 (a)(d)(e)(f), where as the annotators omitted *eye* and *mouth* labels for (b) and (c).

The annotated (bounding-box, semantic label) pairs are then analysed to discover well supported *meronym* relations of the form $x \subset y$ between concepts. Here $x \subset y$ means, concept *x* is a sub-part of concept *y*. eg; *eyes* ⊂ *face* and *face* ⊂ *human*.

$$x \subset y \Rightarrow \{area(x) \cap area(y) \approx area(x)\} \wedge \{area(x) < area(y)\} \quad (4)$$

Some frequently occurring meronym relationships in NUSEF [71] are visualized in Figure 28. This ontology of concepts, then forms the basis for quantifying *attentional-bias* amongst concepts.

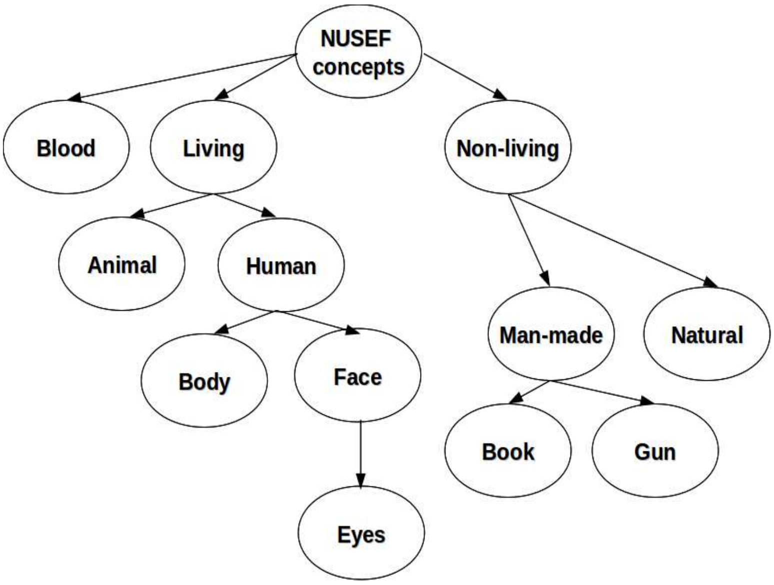
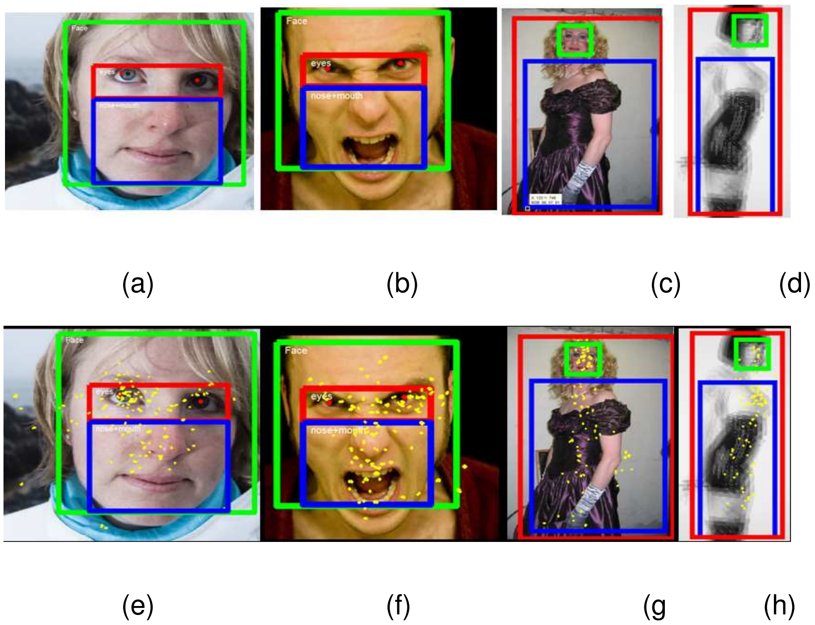


Figure 28: A visualization of some well-supported meronyms relationships in the NUSEF [71] dataset. Manually annotated pairs of (bounding-box, semantic label) are analysed for *part-of* relationships, also described in Eqn. 4.

The first row of Figure29 shows how *face* and *person* images are represented using a number of rectangular ROIs. The second row in Figure29 shows the distribution of fixation points (in yellow) among the ROIs for the various semantic classes for a subject population. Each fixation point is associated with a (x, y) location, which denotes the fixation coordinates, as well as a sequence index, S_j, that denotes the chronological order in which the fixations happened for each image during the 5 second viewing time. Note that a unique set of S_is are associated with every image viewed by each subject. Let n ROIs $\{a_1, .., a_n\}$ constitute image I, such that $\bigcup a_i \subseteq I$. As evident from Figure29, ROIs can overlap, and the $\subseteq$ symbol denotes that some image regions may be unlabeled.



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

(e) (f) (g (h)

Figure 29: Automatically extracted ROIs for (a) *normal* and (b) *expressive* face, (c) *portrait* and (d) *nude* are shown in the first row. Bottom row (e-h) show fixation distribution among the automatically obtained ROIs.

If m subjects have viewed image I, and $D_{i,j}$ denotes the duration for which subject j has fixated on a_i, the representative fixation duration, D_i for concept $a_i \in I$, is given by

$$D_{iI} = \frac{1}{m} \sum_{j=1}^{m} D_{i,j} \tag{5}$$

The intuition for this expression is visualized for a hypothetical concept with three non overlapping sub-parts is visualized in Figure 30.

Generally, the fixation duration D_i is also proportional to the number of fixation points, or the fixation density, within a given ROI. This measure is useful to compute preferential bias to different concepts

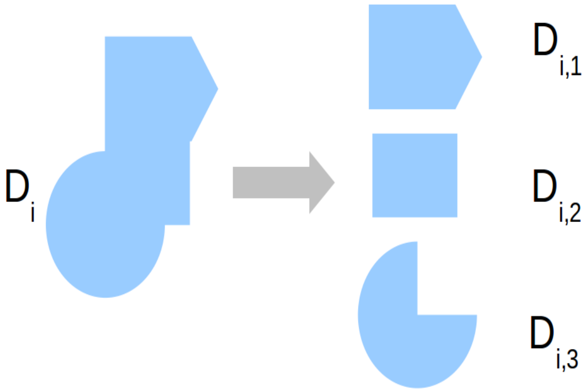


Figure 30: The figure visualizes how the total fixation time over a concept D_i can be explained in terms of time spent on individual, non overlapping sub-parts. The final ratios are derived from combined fixations from all viewers over objects and sub-parts in an image.

as shown later in 4.2.2.

Furthermore, given a pair of ROIs (concepts) $(a_l, a_m) \in I$, let $TC_{l,m,j}$, $NF_{l,j}$ respectively denote the fixation transition count from a_l to a_m and the number of fixations in a_l for subject j. The representative conditional probability $P(m/l)_I$, which models the likelihood of a fixation transition from a_l to a_m following a fixation in a_l is defined as

$$P(m/l)_I = \frac{\sum_{j=1}^{m}(TC_{l,m,j})}{\sum_{j=1}^{m}(NF_{l,j})} \tag{6}$$

Hence, $P(m/l)_I$ values are probability scores in the interval [[0,1]]. The significance of modeling fixation transition probability between ROIs can be seen in Figure 31, where a significant share of inter-fixation transitions is taken up by the strongly interacting ROIs *face* and *laptop*. Interaction amongst concepts results in higher likelihood of fixation transitions between corresponding ROIs. Equation 6 can be used to construct a notion of such interaction as shown later in 4.2.4.

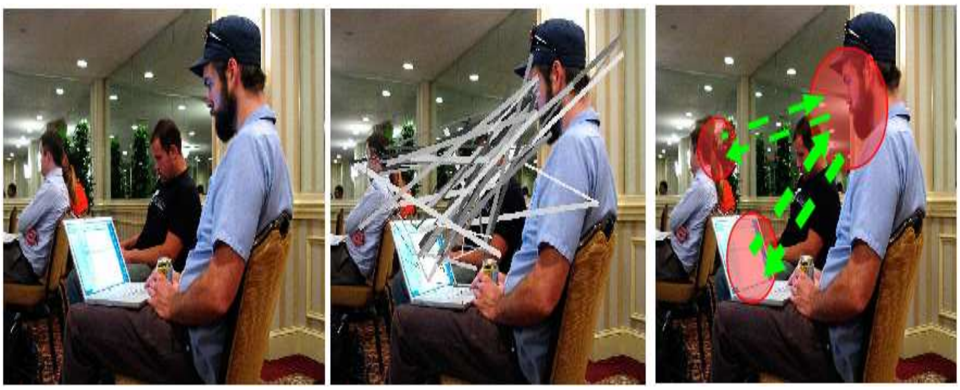


Figure 31: Panel (b) visualizes fixation transitions between important concepts in the image (a). The transitions are also color coded with gray scale values representing fixation onset time, black represents early onset and white represents fixation onset much later in a 5 second presentation time. Visualized data represents eye-gaze recordings from 22 subjects and is part of the NUSEF dataset [71].(c) Red circles illustrate the well supported regions of interest, green dotted arrows show the dominant, pair-wise $P(m/l)_I$ and $P(l/m)_I$ values between concepts *m* and *l*, thickness of the arrows is proportional to the probability values values.

4.2.2 Bias weight

Concepts relevant to an image are determined by the real-world instance that the image captures. For example, *eyes*, *nose* and *mouth* are the most relevant concepts in an *face* image, whereas *face* and *body* are most relevant in a *portrait* image. Depending on the image semantics, we observe a preferential **attentional-bias** towards the *salient* image concepts.

We seek to model this attentional-bias for concept a_i (for the sake of simplicity, we simply refer to concepts through ROIs) through the bias weight measure, denoted by w_i. If P_i is the parent concept for a_i as given by the hierarchical relationship between real-world concepts(*e.g.*, face is the parent concept for eyes), the ROI for P_i contains a_i in I. If S_i denotes the set of N_i images containing a_i, the bias weight, w_i for concept a_i is defined as,

$$w_i = \frac{1}{N_i} \sum_{\forall I \in S_i} \frac{D_{iI}}{D_{P_i I}} \tag{7}$$

Since D_i is essentially characterized by the fixation density within a_i, a large value of w_i indicates that visual attention is biased towards concept a_i. The bias weight w_i for a concept a_i can be learned by sampling the fixation densities corresponding to a_i from a sufficiently large number of training images. Upon learning the w_i's corresponding to many a_i's, it would be possible to *predict* the attentional-bias for each a_i in an image where many a_i's co-occur.

To illustrate, typical w_i values computed from training images for the semantic image categories shown in Figure29 are shown in Table.5. Considering faces as an example, we can see how the w_i's vary for the *eyes* and *nose+mouth* regions for *normal* (neutral and smiling faces) and *expressive* (angry, surprise, disgust) faces. For normal faces, the fixation densities around the *eyes* and *nose+mouth* regions are roughly equal, as observed from training statistics. However, in faces showing pronounced facial deformations around the lower part of the face, the fixation density in the *nose+mouth* region is higher, resulting in a higher w_i value for the *nose+mouth* region in *expressive* faces. This phenomenon can be observed from Figure29(a),(b).

Similarly, while human visual attention is specific to faces in *portraits*, the trend reverses for *nudes*, *i.e.*, considerably higher fixation densities are observed on the *body* in *nudes*. This semantic category-dependent variation in attentional-bias can be exploited to distinguish between different semantic categories corresponding to the same parent class.

Category	**#Images**	**Concept-** w_i
World	30	*living beings*- 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	$mean(P(m/l)_I) - \ 0.4$

Table 5: Computation of w_i for a_i's corresponding to the semantic image categories shown in Figure29.

4.2.3 Attentional bias model synthesis from fixation data

As illustrated in Table 5, we empirically observe that high values of D_i and $P(m/l)_I$ correspond to preferentially attended objects and actions respectively. From labeled image ROIs, we first construct the attentional bias model as an ontology tree incorporating hierarchical relationships between world concepts. For example, a partial ontology for NUSEF [71] is shown in (Figure28). Preferential attention is then learning for a particular concept, from images where it attracts significant eye-fixations, and also co-occurs with other concepts in the world ontology. Each concept in the complete ontology is associated with the *bias weight* w_i, which measures preferential attention to it against other concepts at the same hierarchy level and is depicted in blue in Figure32.

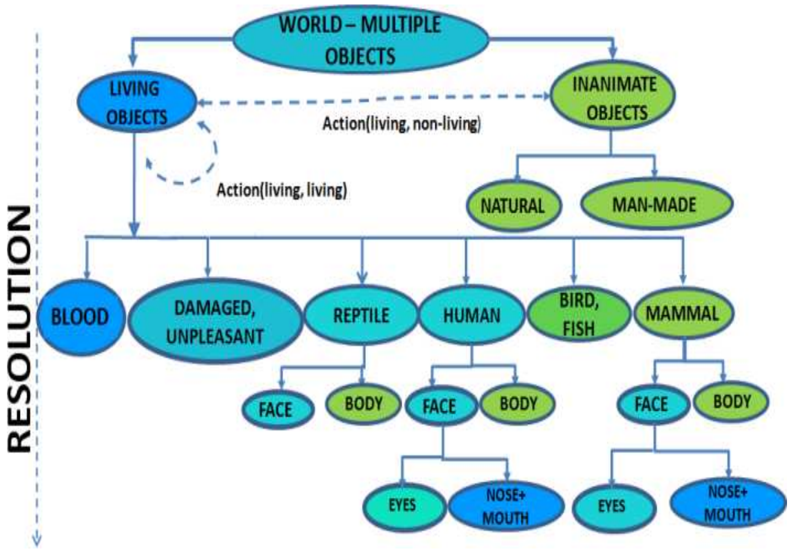


Figure 32: Attentional bias model. A shift from blue to green-shaded ellipses denotes a shift from preferentially attended to concepts having high w_i values, to those less fixated upon and have lower w_i. Dotted arrows represent action-characteristic fixation transitions between objects. The vertical axis represents decreasing object size due to the object-part ontology and is marked by *Resolution*.

World images, which represent a collection of *living* and *inanimate* objects, are used to infer that *living beings* are preferentially attended to. *Face* grabs attention in normal *humans*/*mammals*, while the *body* is substantially affective and hence, preferentially attended in *nude* images. Within the *face*, *nose* and *mouth* correspond to a higher w_i, especially for *expressive* faces. As expected, many preferentially attended concepts like *blood* or *nude* are also affective in nature. The *attentional bias* model is a richer and more complete notion of saliency as it captures the influence of top-down and bottom-up cues in an image. This is also the benchmark to compare against automated saliency detection methods that involve one or more of the following, low-level image information [39], semantic objects and their relative importance [86][43], and possible affect and interaction [70] in the scene. Application of the *attentional bias* model to localize concepts from text captions accompanying images is demonstrated later in this thesis. This model is also useful in any task that requires image saliency and captures the influence of both top-down and bottom-up cues in scenes.

4.2.4 A basic measure for interaction in Image concepts

We define *action* images as those that characterized by a noticeable interaction between the *source* and *recipient* (Figure34(a,b) and Figure31). Let image I with n ROIs the representative interaction measure $Int_{(l,m)}I$, which builds on equation 6 and models the interaction between each key ROI pair a_l, a_m, is defined as,

$$Int_{(l,m)I} = P(m/l)_I + P(l/m)_I \tag{8}$$

When there is a strong interaction observed between a pair of entities (concepts), extensively high number of eye-gaze transitions are observed between the entity-pair as illustrated by the *man looks at laptop* image (Figure 31), resulting in high $Int_{(l,m)I}$ values. The interaction measure is visualised in 33 . A more robust method to detect interaction in images is presented later in 4.3.3.

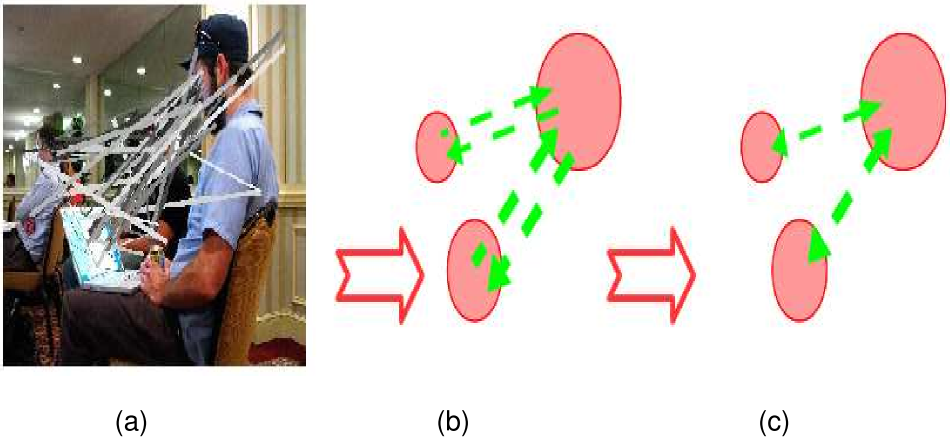
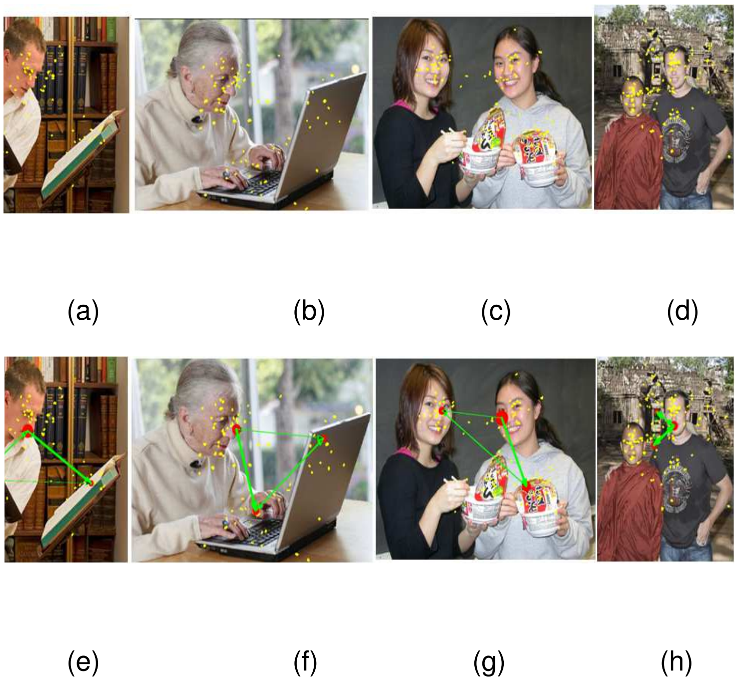


Figure 33: Panel (a) visualizes fixation transitions between important concepts in the image, transitions are color coded with gray scale values representing fixation onset time, black represents early onset and white represents fixation onset much later in a 5 second presentation time. Visualized data represents eye-gaze recordings from 22 subjects and is part of the NUSEF dataset [71].(b) Red circles in the cartoon illustrate the well supported regions of interest, green dotted arrows show the dominant, pair-wise $P(m/l)_I$ and $P(l/m)_I$ values between concepts *m* and *l*, thickness of the arrows is proportional to the probability values values. (c) Visualization of normalized $Int_{(l,m)I}$ values depicting the dominant interactions in the given image, a single green arrow marks the direction and magnitude of inferred interaction.

4.3 Estimating Regions of Interest in Images using the 'Binning' algorithm

In this section, we describe the 'binning' algorithm that we adopt to automatically determine spatially distinct ROIs based on *time-sequence* information. To the best of our knowledge, the proposed algorithm is one of the first attempts to exploit timing information associated with fixations. We demonstrate that the binning algorithm is particularly useful for determining the presence of key objects, and can be extended to discover inter-object interactions using eye-fixation data. Humans exhibit exploratory behavior as they observe scenes. Majority of fixation transitions occur between locations corresponding to distinct ROIs in the image. This gives rise to an almost bi-partite re-



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

(e) (f) (g) (h)

Figure 34: *Action* vs multiple non-interacting entities. Exemplar images from the *read* and *look* semantic categories are shown in (a),(b). (c),(d) are examples of images containing multiple non-interacting entities. In (e)-(h), the green arrows denote fixation transitions between the different clusters. The thickness of the arrows are indicative of the fixation transition probabilities between two given ROIs.

lation between ROIs representing distinct concepts. We exploit this property of eye-gaze to discover and bound ROIs in an image. Our method discriminates saccades made between ROIs against those made within the same ROI and generates clusters having bi-partite relation with each other. The intuition for this algorithm is visualized along with an actual result in Figure 35.

Successive saccades are seldom on the same ROI, this is explained by the inhibition-of-return phenomenon [48] and is used as a constraint to discover the bi-partite relationships mentioned earlier. Inhibition of return (IOR) refers to the observation that the speed and accuracy with which an object is detected are first briefly enhanced for perhaps 100-300 milliseconds after the object is attended, and then detection speed and accuracy are impaired (for perhaps 500-3000 milliseconds). IOR is believed to promote exploration of new, previously unattended objects in the scene during visual search or foraging by preventing attention from returning to already-attended objects.

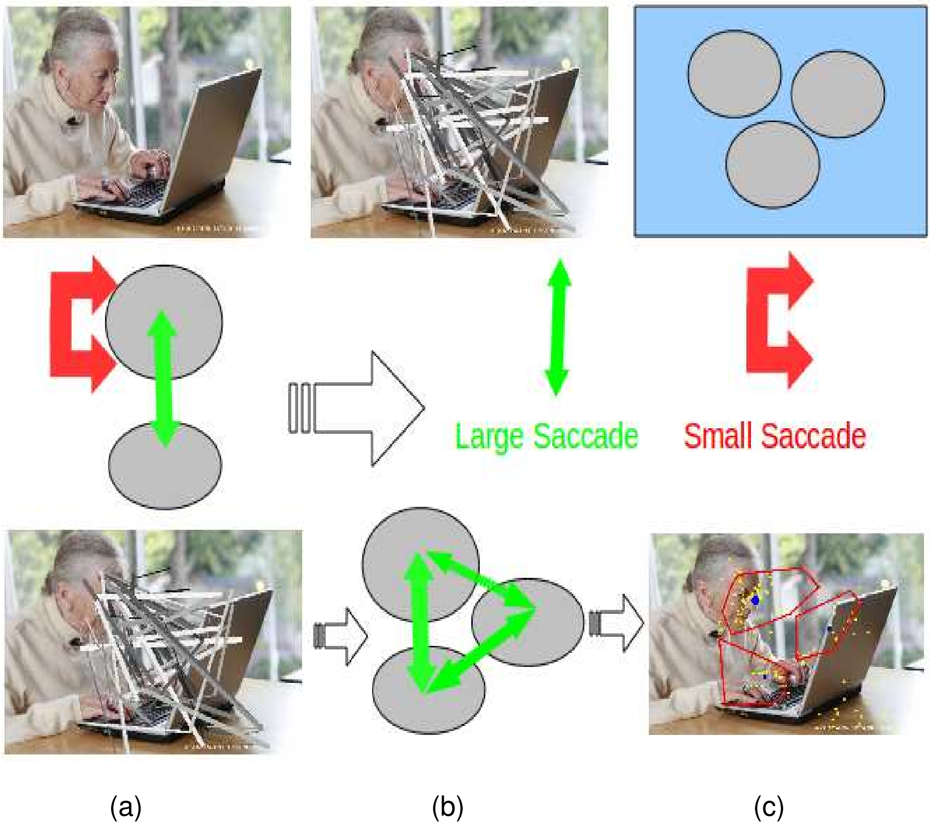


Figure 35: The *binning* algorithm. Panels in the top row show a representative image (top-left) and eye-fixation information visualized as described earlier in 26, followed by abstraction of the key visual elements in the image. The middle row illustrates how inter-fixation saccades can be between the same ROI (red arrow) or distinct ROIs (green arrow). The bottom row illustrates how isolating inter-ROI saccades enables grouping of fixation points potentially belonging to the same ROI into one cluster. The right panel in the bottom row is an output from the *binning* algorithm for the chosen image, ROIs clusters are depicted using red polygons and the cluster centroid is illustrated with a blue disc of radius proportional to the cluster support. Yellow dots are eye-fixation information that is input to the algorithm.

Algorithm 4.1: BinningMethod($FixationData\ S$)

```
$bins \leftarrow$ [NULL]; $BinAdj \leftarrow$ [NULL][NULL]
for each $S_j \in S$
  do
    $prevBin \leftarrow null$
    if $isempty(bins)$
      then
        $bins.create()$
        $bins(1).add(S_j)$
        $prevBin \leftarrow 1$
      else
        $foundBin \leftarrow bin\ closest\ to\ S_j$
        $prevBin \leftarrow foundBin$
        if $dist(foundBin, S_j) > neighbourhood$
          then
            $newBin \leftarrow bins.addNewBin()$
            $bins(newBin).add(S_j)$
            $BinAdj(prevBin, newBin).addEdge()$
            $prevBin \leftarrow newBin$
          else
            if $foundBin \neq prevBin$
              then
                $bins(foundBin).add(S_j)$
                $BinAdj(prevBin, foundBin).addEdge()$
              else
                if $foundBin == prevBin$
                  then
                    $bins(foundBin).add(S_j)$
return $(bins, BinAdj)$
```

The binning procedure is summarized in Algorithm 4.1. The algorithm works as follows, given a set of P fixation points, the binning algorithm assigns them to N bins. The algorithm begins with NULL bins, and bins are created with time, based on the spatial distribution of fixation points, *i.e.*, any fixation point S_j is assigned to it's *closest* bin, based on Euclidean distance to the bin centroid. However, if there is no bin within D_{thresh} distance from S_j, a new bin is created with S_j as its member. Fixation points are added to bins based on the following criterion: If S_j is assigned to bin_k, the algorithm will attempt to assign S_{j+1} one of the remaining $N-1$ bins that are within D_{thresh} distance of S_{j+1}, and will not be assigned to bin_k. If there is no bin within D_{thresh} distance from S_{j+1}, a new bin is created with S_{j+1} as its member. If S_{j+1} is within D_{thresh} of bin_k, membership count of bin_k is incremented and it's assumed that the subject's eye-gaze hasn't transitioned to another ROI in the image. The choice of D_{thresh} is an indicator of the scale of objects occurring in images, in the current implementation, it is set heuristically for typical sizes of objects and their parts. S_j being assigned to bin_l, and $S_j + 1$ to bin_m is equivalent to a *fixation transition* between ROIs a_l and a_m. The $BinAdj$ matrix, which stores the number of transitions between bins $l, m \forall\, l, m = 1..N$, is updated after the assignment of each S_j. Transitions from bin_k to itself are ignored. The bin that contains the most S_j's corresponds to the most *salient* image concept.

ROI discovery over some representative image themes is illustrated in Figure 36. Red polygons outline the ROI clusters, the ROIs often have good overlap with the underlying semantic concept, object or part.

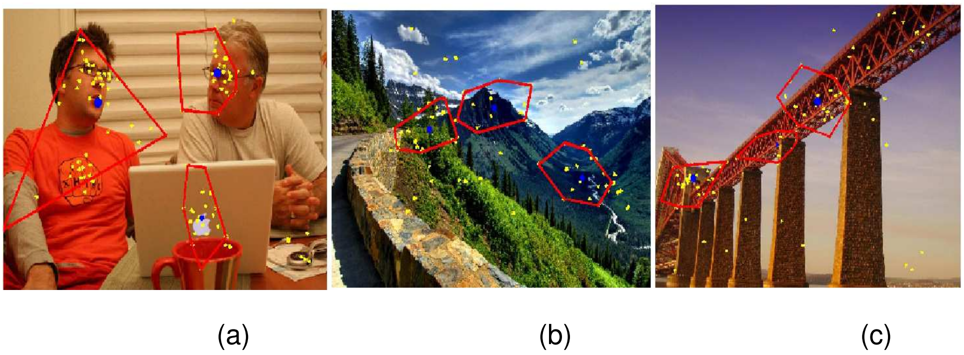


Figure 36: Panels illustrate how ROIs identified by the the *binning* method correspond to visual elements that might be at the level of objects, gestalt elements or abstract concepts. (a) ROIs correspond to the faces involved in the conversation and the apple logo on the laptop.(b) Key elements in the image *solitary mountain* and the two *vanishing points* one on the left where the road curves around and another where the river vanishes into the valley. Vanishing points are strong perceptual cues. (c) Junctions of the bridge and columns are fixated upon selectively by users and are captured well in the discovered ROIs.

4.3.1 Performance analysis of the binning method

The binning method Algorithm 4.1 requires a distance computation to each existing bin, for every fixation point to be binned and can be seen in the nested *for loop* pair in Algorithm 4.1. In the worst case, there can be as many bins as there are fixation points. Hence, it offers an $O(n^2)$ time complexity and $O(n)$ storage complexity for n fixation points.

The performance of the method is influenced by the *neighborhood* chosen in Algorithm 4.1, as well as the number of eye-gaze traces corresponding to an image. The evaluation is done by computing precision, recall and f-measure scores of eye-gaze based ROIs and comparing them with human annotated ground truth. 5 annotators were given randomly chosen images from the NUSEF dataset [71], the annotators assign white to foreground regions and the black to the background. Annotators used the GIMP®tool to identify the object boundaries and assign white to foreground regions and black to background regions respectively. The eye-gaze based ROIs and segmentation ground truth is visualized in Figure 37. Evaluation of the binning method and another mean-shift based method [77], are presented in the following sections. In this thesis, evaluations are done using at least 120 randomly chosen images from amongst a pool of over 1000 images from public datasets [71], [43]. The number of images chosen is comparable or larger than the recent, corresponding reported analysis tasks such as eye-gaze based clustering [70] and computer vision tasks such as segmentation [71] and object detection [46]. The large pool of images and statistically significant number of human subjects gives a reliable insight into the behavior of the chosen methods.

We find that images can contain one or more objects of interest and this was captured during the manual annotation task. Some segmentation maps with one or more ROIs are illustrated in Figure 38. The images chosen for evaluation have a diversity in the number of number ROIs, location and scale of objects.

We combine $precision$ and $recall$ against the baseline using an $fmeasure$ score computed over eye-gaze based ROIs with respect to human annotated ground-truth (*gtruth*) foreground regions as illustrated in Figures 38 and 37 as,

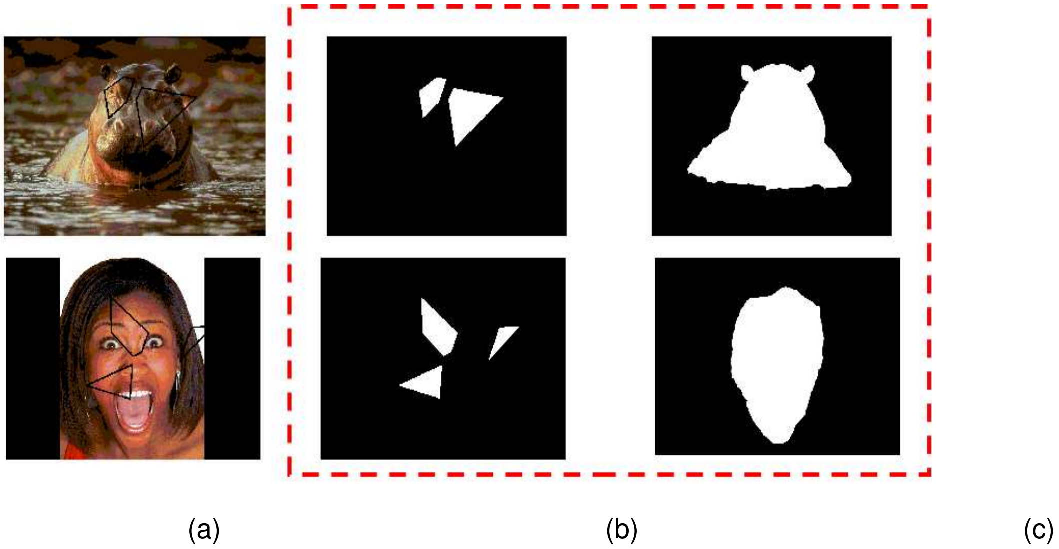


Figure 37: Visualization of eye-gaze based ROIs obtained from binning and the corresponding manually annotated ground truth for evaulation (a) Original image (b) Eye-gaze based ROIs (c) Manually annotated ground truth for the corresponding Image. 5 annotators were given randomly chosen images from the NUSEF dataset [71], the annotators assign white to foreground regions and the black to the background.

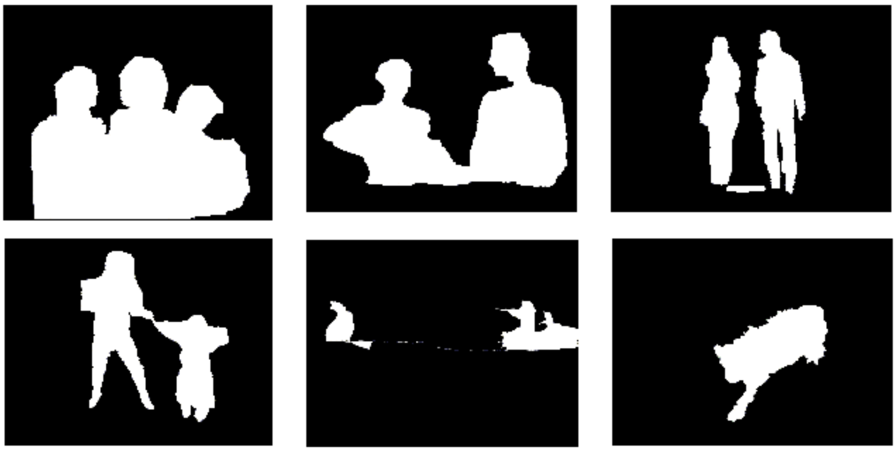


Figure 38: Visualization manually annotated ground truth for randomly chosen images from the NUSEF dataset [71]. The images can have one or more ROIs.

$$fmeasure = \frac{2 * precision * recall}{precision + recall} \tag{9}$$

Precision and recall for each box are,

$$precision = \frac{eye - gazeROIs \cap gtruth}{eye - gazeROIs} \tag{10}$$

$$recall = \frac{eye - gazeROIs \cap gtruth}{gtruth} \qquad (11)$$

Here, $eye - gazeROIs \cap gtruth$ is computed by pixel-wise boolean *AND* operation between the eye-gaze ROIs and manual ground truth as shown in Figure 37 (b) and (c) respectively.

A representative result for 15 subjects per image is shown in Figure 39. The binning method is conservative in assigning fixation points into bins, the chosen points are usually well within the object boundary and so is the corresponding enclosing convex hull ROI. This is a good strategy to reduce false positives in subsequent applications for the ROIs, such as eye-gaze based foreground estimation or guiding object detectors to find key objects. We find that f-measure scores

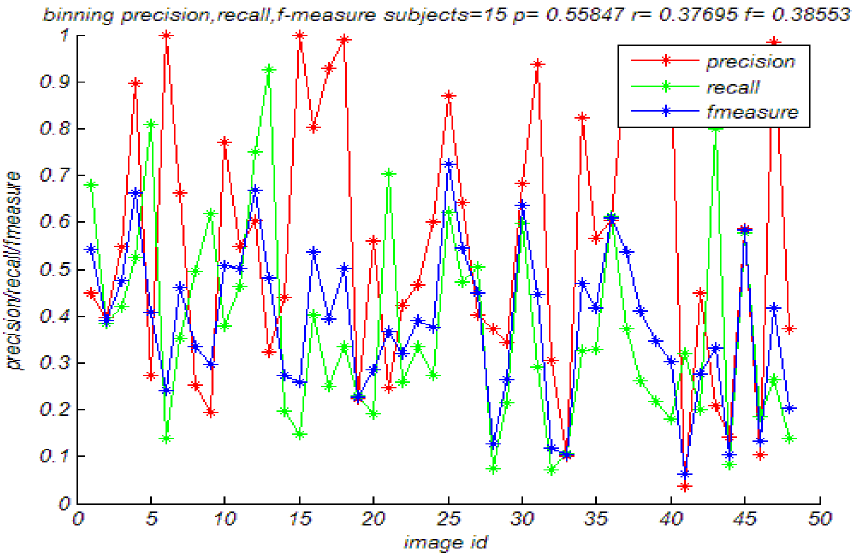


Figure 39: Performance of the binning method for 50 randomly chosen images from the NUSEF dataset. The binning method employs a conservative strategy to select fixation points into bins, large proportion of fixation points fall within object boundary. This results in higher precision values as compared to recall. An f-measure of 38.5% is achieved in this case.

for ROIs stabilise as the number of subjects per image crosses 20 subjects. This can be seen in Figure 40.

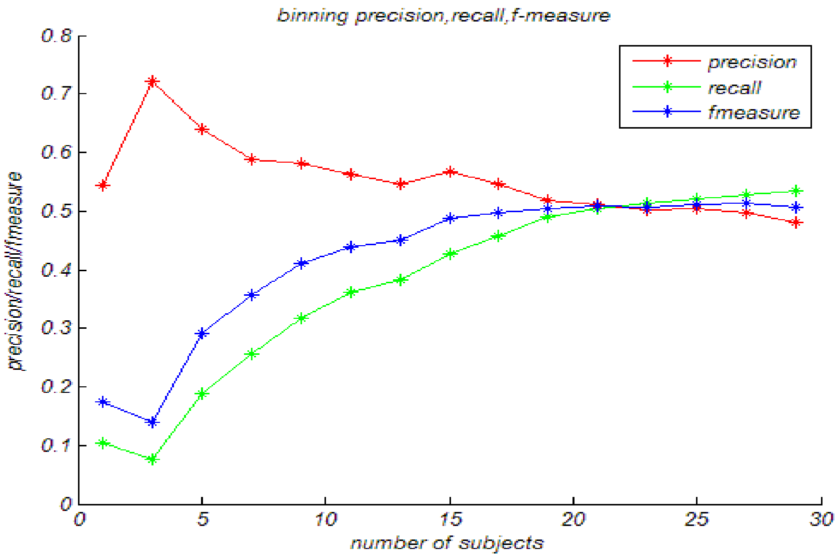
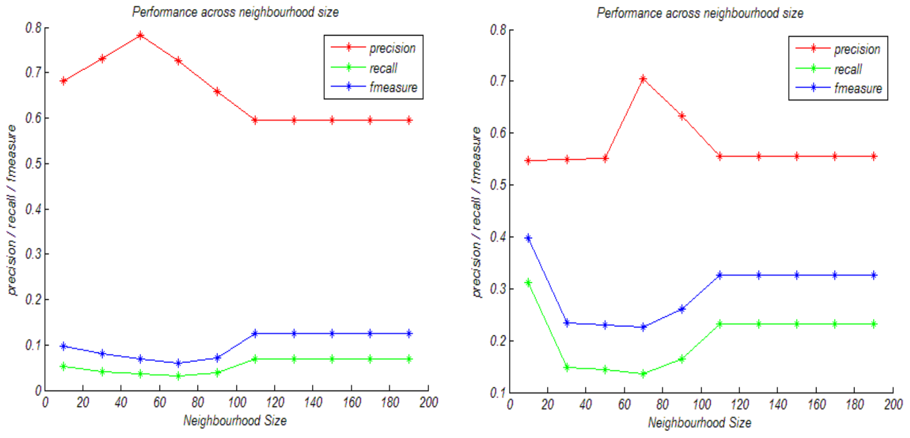


Figure 40: Performance of the binning method as the number of subjects viewing an image is increased from 1 to 30. The neighbourhood value is chosen to be 130 pixels to discriminate between intra-object saccades and inter-object saccades. The precision, recall and consequently f-measure are approximately even at 20 subjects.

Reliable estimates of ROI positions and sizes can be obtained by group statistics over groups of subjects. At the same time, an important question is how precision and recall of eye-gaze based ROIs, when the eye-gaze data available is from one or a few subjects. A single person interacting with video or image content is important for personalization and interactive scenarios. This is indeed true and can be seen in Figure 41 (a). The low score is mainly due to the small areas of eye-gaze ROIs and hence poor recall for the corresponding visual concept. This shortcoming can be improved upon by growing eye-gaze based ROIs using content based image segmentation. The *active segmentation* method [60] is employed for this purpose. Active segmentation relies on an input *fixation seed* to detect the boundary of underlying visual concept. This is an appropriate choice for segmentation as the method does not propose a method to acquire the input fixation seeds. For the current implementation, ROI centroid is chosen as a fixation seed for active segmentation and the resultant segment is then combined with the original eye gaze based ROI using the pixel-wise *OR* operation.

We observe that individual users visit similar ROIs, albeit in different sequence. This means that the binning method can still infer the location of ROIs, but cannot reliably estimate the extent of the regions. Since eye-fixations are due to underlying visual concepts, we use content based analysis to improve estimates of the ROIs. Figure 41 shows the significant increase of more than 2.3 times in f-measure, using this strategy.



(a)

(b)

Figure 41: Panels illustrate precision, recall and fmeasure of the binning method for 1 subject with (a) eye-gaze information alone, and (b) when eye-gaze ROI information are grown using *active segmentation* [60]. A simple fusion of segmentation based cues with eye-gaze ROIs gives an improvement of over 230% in f-measure as shown underlined with the dotted red lines above the graphs.

The *neighborhood* used to discriminate between intra-object and inter-object saccades is an important parameter in the binning method. Larger neighborhoods reveal coarser structures explored by the subject's eye-gaze, this is illustrated in Figure 42. *Neighborhood* values ≤ 25 result in the formation of very small clusters. Few of the small clusters formed, have sufficient membership to be considered as an ROI. These small clusters are well within the object boundary resulting in high precision ($\geq 70\%$) and low ($\leq 30\%$) recall. The cross over point for $neighbourhood = 80$ is due to a combination of factors including stimulus viewing distance, natural statistics of the images and typical eye-movement behavior. Larger neighborhood values result in large, coarse ROIs which can be bigger than the object and include noisy outliers. This causes reduction in precision as well as that in recall.

4.3.2 Evaluation of the binning with a popular baseline method

The mean-shift based clustering method by Santella et. al. [77] is chosen for comparison with the binnnig method. Though different in implementation details, the overall objectives of [77] are similar to the binning method. Namely,(a) to group eye-gaze fixations into meaningful clusters, (b) to be data driven and not depend on initial random guesses and (c) consistency and robustness in the results. The method in [77] groups eye-fixations by assigning labels to each fixation point and then moving them to the nearest dominant group using the mean-shift algorithm [24]. More specifically, each fixation point s is

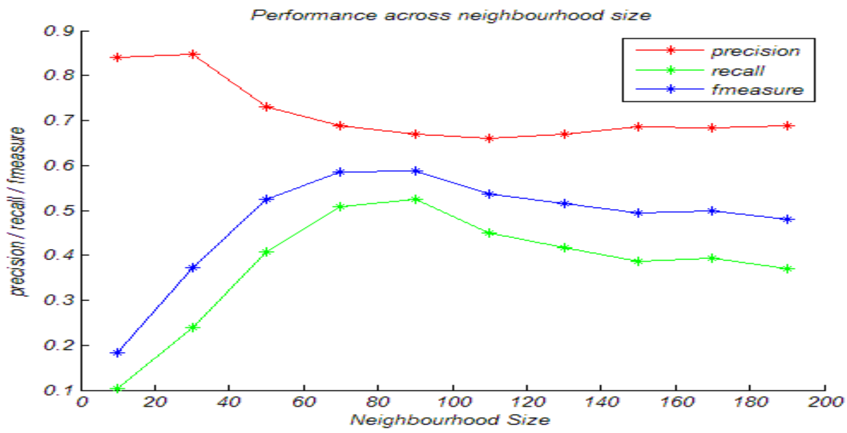


Figure 42: Small *neighborhood* values result in the formation of very small clusters and few of those have sufficient membership to be considered as an ROI. The clusters are well within the object boundary resulting in high precision > 70% and low < 30% recall. The cross over point for $neighbourhood = 80$ is due to a combination of factors including stimulus viewing distance, natural statistics of the images and typical eye-movement behavior. Larger neighborhood values result in large, coarse ROIs which can be bigger than the object and include noisy outliers. This causes reduction in precision as well as that in recall.

moved to a new location $p(s)$ that is a weighted mean of the distances of s to all its neighbors s_j.

$$p(s) = \frac{\sum_{j=1}^{n} \sigma \times |s - s_j| \times s_j}{\sum_{j=1}^{n} \sigma \times |s - s_j|} \tag{12}$$

$|s - s_j|$ being the euclidean distance between s and s_j. The important stages in [77] are presented in Algorithm 4.2,

Algorithm 4.2: MEANSHIFTCLUSTERINGSANTELLA($FixationData\ S$)

```
comment: Initialise one position for each fixation point in S
comment: This is the 0 th iteration
for each $s_j \in S$
  do
  {
    ${y^0}_j \leftarrow s_j$
  }
$k \leftarrow 0$
while convergence
  do
  {
    comment: Apply meanshift iteratively for k th iteration
    for each $s_j \in S$
      do
      {
        ${y^k}_j \leftarrow p({y^{k-1}}_j)$
      }
    $k \leftarrow k + 1$
  }
comment: Return result from most recent iteration
$S \leftarrow \{{y^k}_1, ..., {y^k}_j, ..., {y^k}_N\}$
return $(S)$
```

Salient features of the method in [24] are compared against the binning method in Table. 6.

Feature	Binning method	Santella et. al. [77]
Clustering principle	Discovering bi-partite structures	Meanshift based
Time complexity	$O(n^2)$	$O(iterations \times n^2)$
Storage requirement	$O(n)$	$O(n)$
Notion of neighborhood	near / far saccade threshold	mean-shift neighborhood
Iterative	One pass, not iterative	Iterates till convergence

Table 6: A comparison of the salient features of [24] with those of the binning method proposed in this thesis.

One important difference beteween the two methods is the way of moving or assigning an eye gaze point to its appropriate cluster. This is visualised in Figure 43. The meanshift method (b) effectively replaces the new point s_j with the weighted mean of all points in the specified *neighbourhood*. This process repeats iteratively over all points, till they gather around the mode of each cluster. The process avoids initial random guesses like in the popular k-means clustering method. On the other hand, the binning method (b) orders existing bins according to distances to their centroids from s_j and then finds the bin containing a gaze point very close to s_j.

The two algorithms are compared for precision, recall and f-measure over a diverse variety of images. The number of subjects per image is set to 21, which is sufficient for both algorithms to give consistent and robust eye-gaze based ROIs. Increasing the neighborhood parameter in [77] results in identification of coarse structures and larger ROI sizes. Smaller values of the neighborhood result in smaller ROIs contained within objects and consequently low recall and high precision.

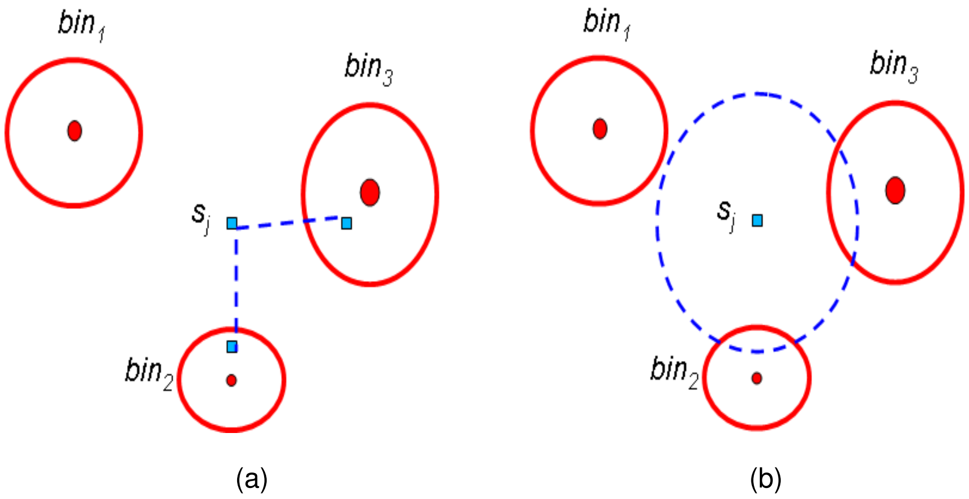


Figure 43: The binning method (a) orders existing bins according to distances to their centroids from s_j and then finds the bin containing a gaze point very close to s_j. On the other hand, the mean-shift based method in [77] replaces the new point s_j with the weighted mean of all points in the specified *neighbourhood*.

A point of difference between [77] and the binning method is for larger neighborhood values, the binning method ROIs do not keep increasing in size unlike the mean-shift based method. This helps keep the ROI size close to object contour and high precision values are maintained. On the contrary, ROI sizes keep increasing and lead to falling precision in [77]. A comparison of the precision, recall and fmeasure performance of the two methods is presented in Figure 44.

4.3.3 Extending the binning algorithm to infer Interaction represented in static images

Since the binning algorithm keeps track of the number of fixation transitions between each of the image ROIs, it is especially useful for automatically inferring strong object interactions from images. This is especially interesting because it is extremely difficult to infer object

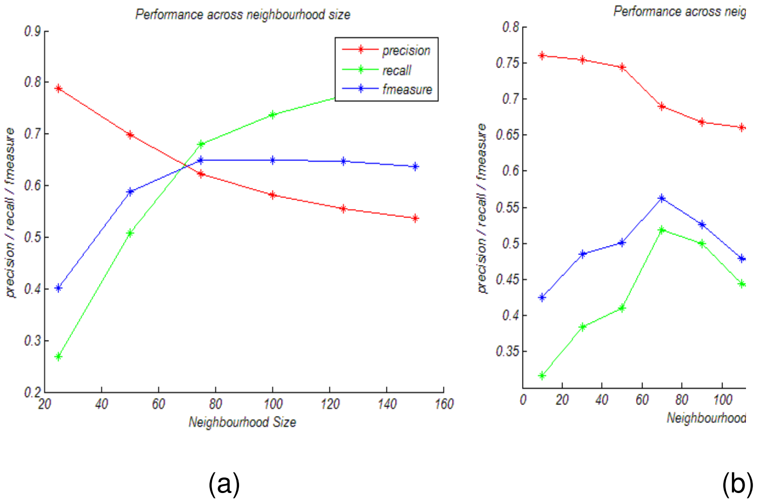


Figure 44: A comparison of precision, recall and fmeasure variations between (a) the mean-shift based method in [77] and the binning method presented in this thesis. The behavior of both methods for smaller neighborhood values is similar. It changes significantly for larger neighborhood values, where ROI sizes are preserved in the binning method and result in preservation of precision scores. On the other hand, recall values fall in the binning method as compared to [77].

interactions such as *look*, by applying computer-vision based techniques on image or video data.

The *binning* method relies on identifying fixation transitions between important concepts in images. Important concepts in images are attention grabbing and human subjects quickly detect their presence and possible relationships. This brings about another interesting and useful property of images having visual elements with strong relationships. Namely, interacting ROIs are preserved even as the neighborhood is changed in the *binning* method. This is a consequence of most saccades being devoted to exploring these key regions and their relationships. An example can be see by contrasting the locations of dominant interacting pairs in Figure 46 with Figure 45. There is significant change in the locations, sizes and number of ROIs discovered when the neighborhood values are changed all the way from a very small neighborhood corresponding to $0.5°$ which can lead to many clusters, to a very large one of $5°$ where all fixation points end up in the same ROI.

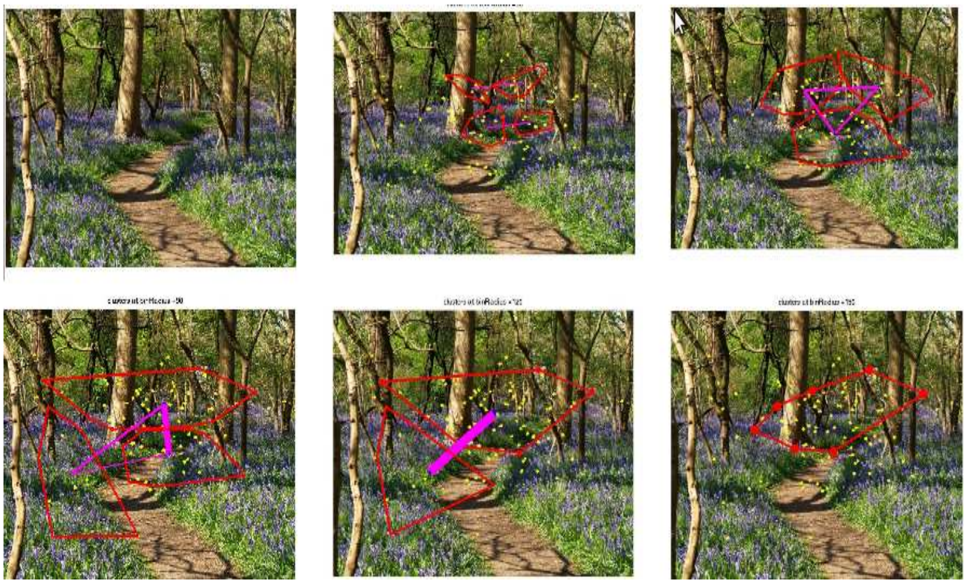


Figure 45: Clusters obtained with varying neighborhoods over image with weak interactions

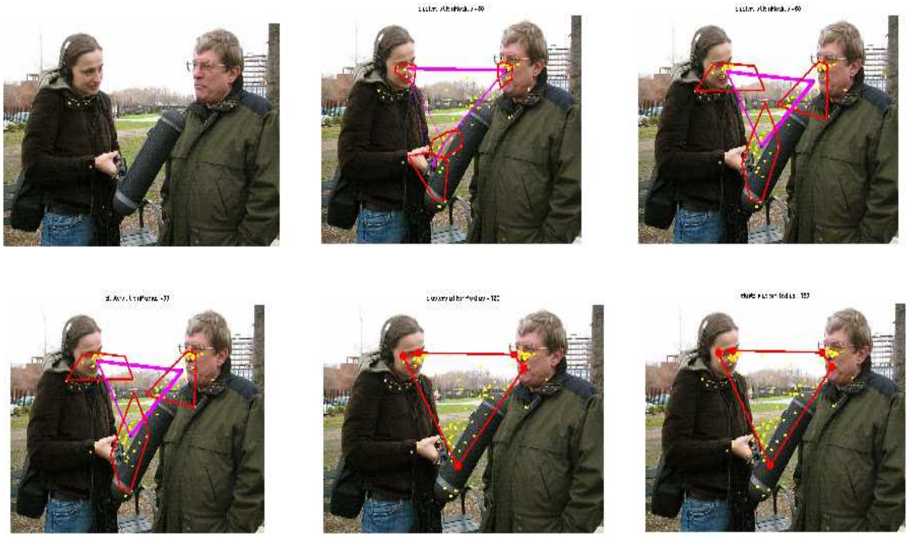


Figure 46: Clusters obtained with varying neighborhoods over image with strong interactions

The change in ROIs and their interactions is accumulated over these neighborhoods and gives a measure of how strong the possible interactions in images are. Images with strongly interacting visual elements lead to minimal changes in the ROIs and fixation transitions between them. An 10×10 *interaction matrix A* is constructed corresponding to fixation transitions between ROIs with large membership of fixation points, at each value of the neighborhood n. Each of the $10x10$ tiles corresponds roughly to the average ROI size that we discover over the NUSEF [71] dataset. Entries in the *interaction matrix* correspond to fixation transitions between the corresponding ROIs as observed from the eye-tracking data. The interactions follow the definition given earlier in Eqn. 8. ie; every $A(i, j)$ is nothing but $Int_{(l,m)I}$ defined as,

$$Int_{(l,m)I} = \overline{P(m/l)}_I + \overline{P(l/m)}_I \qquad (13)$$

The accumulated change in *interaction matrix* is measured as,

$$\sum_{n=0.5}^{5} A(n + step) - A(n) \qquad (14)$$

Where $A(n+step)-A(n)$ is the magnitude of element wise difference between the *interaction matrix* values at successive values of neighborhood *n* and *step* is the increment in neighborhood for the *binning* method.

4.4 Modeling attentional bias for videos

This section shows how eye-gaze can be used as an indicator of HVA when subjects free-view videos. An illustrative experimental result to motivate the idea is shown in Figure 47. Simlar to the attentional bias model for images as described earlier, the model for video is also based on eye-gaze analysis. Our framework is tailored for online estimation of human interest using eye-gaze based ROIs. These are then used in a closed-loop, interactive application. Some interesting questions in this context are,

- Can eye-gaze information be used to indicate reliable and useful ROIs in video content ?
- Are they different from conventional motion based [3] and object-based [43] saliency in videos ?
- Will an eye-gaze based attentional bias model hold over different subjects, or successive views over the same video content ?

The first two questions are addressed by Figure 47, it illustrates how HVA follows key-objects(important actors) and interesting events closely, these are difficult to infer from content analysis. This can be a valuable cue for a video summarization or re-targeting system.

Consistency of eye-gaze over different subjects as well as successive views of the same content is illustrated in Figures 49 and 48 respectively. Three different subjects were shown the same video clip and eye-movements recorded, the corresponding fixations are visualized in Figure 49, good agreement can be seen over preferentially fixated regions (larger circles). A representative result from 5 successive views of a meeting room discussion clip in Figure 48, events of interest in the video result in longer eye-fixations and can be seen in the larger coloured circles.

4.4.1 video-binning : Discovering ROIs and propagating to future frames

Our framework maintains the recent past history of viewing to anticipate ROIs for video frames in the near future. We observe that

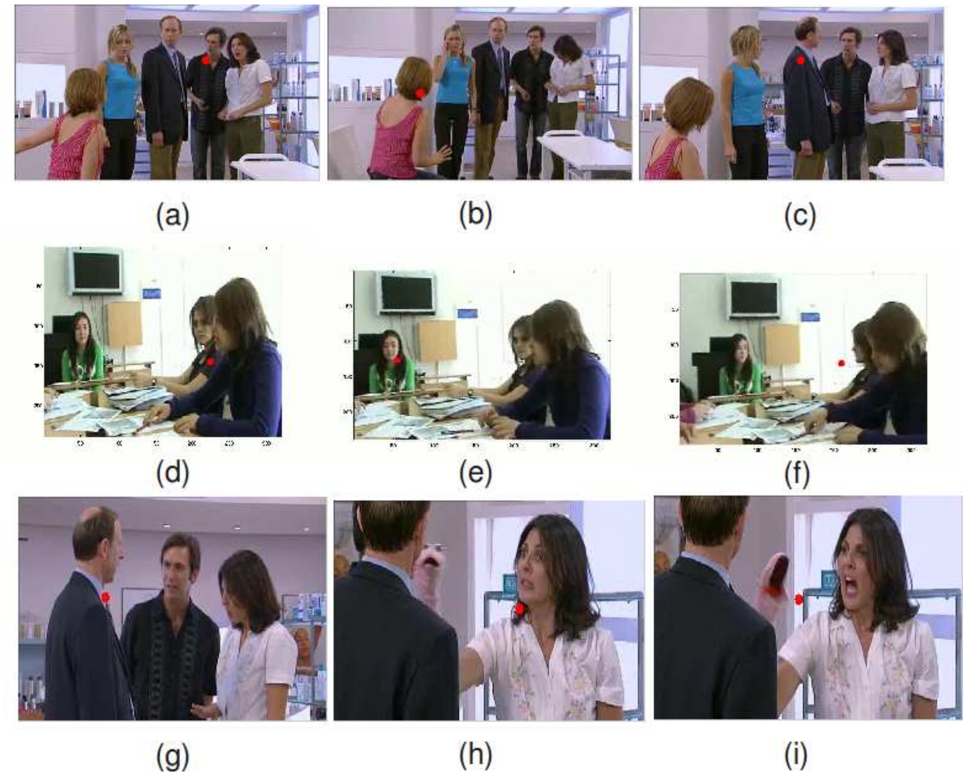


Figure 47: (a),(b) and (c) Illustrate shift in *HVA* shown by the *red dot*, as the prominent speaker changes in a video sequence. (d),(e) and (f) show the same in a different video sequence. An interesting event is depicted in (g),(h),(i), where the *HVA* shifts from prominent speaker in (h) to the *talking-puppet* in (i), which is actually more *meaningful* and *compelling* in the scene.

significant objects and their interaction in the recent past also play a key role in the immediate future video frames. This is a reasonable assumption for the duration of a short video scene or event spanning few seconds or more. In such scenes, main elements are likely to remain same and the the video frame rate is sufficient to capture their dynamics. It is important to note here that video editing can bring about abtrupt changes at shot boundaries, this might include change in foreground and background content. Our current strategy is to detect such shot boundaries and reset the ROI position to the center or bottom of the screen. Our evaulation shows this to be a satisfac-

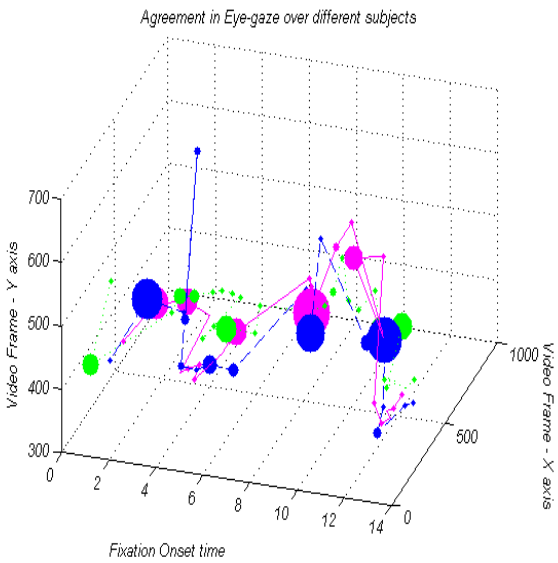


Figure 48: The graph illustrates spatio-temporal eye-gaze data, it indicates good agreement of human eye-movements over 3 different subjects while viewing a video clip, clip height and width form two axes and a third one is formed by the video frame display time. Eye fixations are aligned according to the onset time and each subject is depicted using distinct colors. The colored blobs depict the eye-fixation duration on a ROI in the video stimulus.

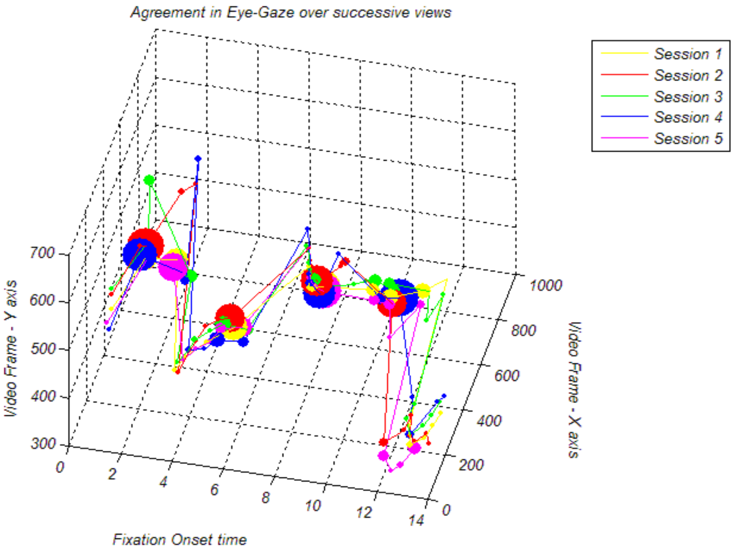


Figure 49: Good agreement of human eye-movements over successive views of a video clip, clip height and width form two axes and a third one is formed by the video frame display time. Eye fixations are aligned according to the onset time and each viewing session is depicted using distinct colors. The colored blobs depict the eye-fixation duration on a ROI in the video stimulus.

tory strategy for video themes such as dramas and operas and sports commentaries.

In a video scene a few key objects compete for user attention, one of which is selected and brought to the center of our visual field. Thus the ROI is one amongst few candidates in the dynamic scene. We extend the basic *binning* method [46] to work over a moving window of time and discover ROIs. More formally,

For a buffer of b gaze points, let the buffer of gaze points at time t be G_t, forming the framework's short term memory,

$$G_t = \{g_{t-b}, ..., g_{t-k}, ..., g_t\} \tag{15}$$

For our current setup with the ERICA©eye-tracker, the gaze sampling rate matches the video frame rate. Hence, we have one gaze point for every video frame being processed.

Now, due to camera motion as or object motion, the position of key objects changes between successive video frames. This in turn means that gaze points $g_{t-b}, ..., g_{t-1}$ may have drifted away from the current location of the objects of interest. A content based cue that can indicate the extent by the visual concepts underlying the buffered eye-gaze points shift, would be very useful at this juncture. Such a cue can help in aligning past eye-gaze points in the buffer, back onto the corresponding visual concept in the current frame. We use *bottom-up* motion vector cues from video to propagate eye-gaze points through $g_{t-b}, ..., g_{t-1}$ prior to running *video-binning*. Motion vector information is ubiquitous in compressed video and readily available at the video

decoder. Averaged motion vector based shifts mv_{t-k} are computed for each $g_{t-k} \in G_t$, over a neighborhood equal to the size of the average ROI in the current video stream.
The pre-computed motion vector information MV exists in the video clips,

$$mv_{t-k} \in MV \tag{16}$$

A point to note here is that mv_k is a recursively computed value over motion vector shifts in frames $t-k, ..., t-b$. As motion vectors are generated using macroblock similarity and do not have any semantic connotation large motion vector shifts are unreliable and we drop such gaze points. The updated eye-gaze buffer G_t is then passed as input to the binning method. This is also a novel notion of grouping eye gaze points together, unlike conventional definition of eye fixations, which are groups of gaze points falling within 1^o visual angle over a duration of 100 milliseconds or more and are termed as an *eye fixation*. Our motion vector based update is described in Algorithm 4.4.

Let $ROIs_t$ be the output ROIs from binning, these are not always co-incident on the object of interest, which in our case are faces. This could be due to a combination of factors including calibration errors in the eye-tracker, personal biases on attending to faces, ambient noise in eye-gaze signal. We re-align eye-gaze based ROIs by using face detection and tracking as a *top-down* cue. Frontal and profile faces are detected using OpenCV based routines. For the current framework, face tracks are maintained by preserving face positions the the location of detection for the duration of the time window giving the set of face tracks,

$$f_t \in F_t \tag{17}$$

This simple notion of a track holds across the moving, time window buffer. The detection routines give many false positives, as can be seen in Figure 66 (d). This is compensated by the specificity of eye gaze based ROIs, human subjects seldom look away from face regions in social videos, this makes $ROIs$ a very good prior to identify the current face track. We construct a face saliency map for each frame as shown in 66 (c) and search for face tracks in the vicinity of each eye-gaze based ROI. The neighborhood is a dynamic parameter and is taken to be $1/3$ rd the average distance between eye-gaze based ROIs (ROI_{mean}) corresponding to past few seconds of video. This in effect is assuming that key objects such as faces and people can be up to $1/3$rd the inter-object distance. The face saliency based update is described in Algorithm 4.5.

Algorithm 4.3: VIDEOBINNING(G_t)

```
$ROIs \leftarrow$ [NULL]
comment: Motion vectors based gaze point update
 $UpdateMotionVectors(G_t, MV)$
comment: Call binning method
 $ROIs \leftarrow binning(G_t)$
comment: Align ROIs to closest face track
 $ROIs \leftarrow UpdateFaceTracks(ROIs, F_t)$
return $(ROIs)$
```

Algorithm 4.4: UPDATEMOTIONVECTORS(G_t, MV)

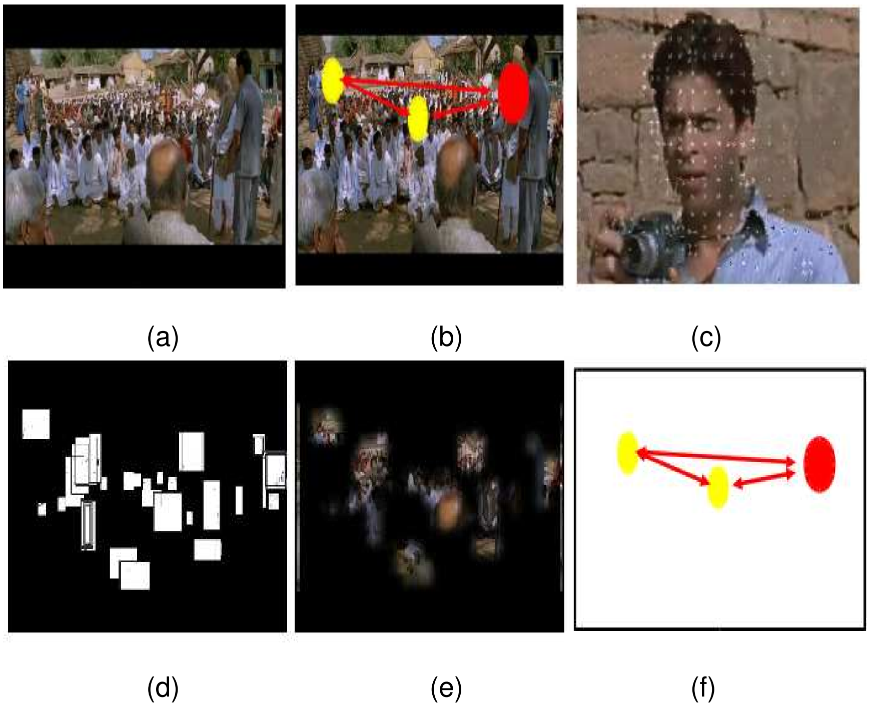
```
for each $g_{t-k}$ $in$ $G_t$
  do {
    $totalShift \leftarrow 0$
    for $i \leftarrow 1$ to $k$
      do {
        comment: Retrieve MV information for $g_{t-i}$
        $(mv_{t-i}.x, mv_{t-i}.y) \leftarrow$ motion vector for $g_{t-i}$
        comment: Add motion vector shift to $g_{t-i}$
        $g_{t-i}.x \leftarrow g_{t-i}.x + mv_{t-i}.x$
        $g_{t-i}.y \leftarrow g_{t-i}.y + mv_{t-i}.y$
        $totalShift \leftarrow (g_{t-i}.x + g_{t-i}.y)/2$
        comment: Reject $g_{t-i}$ if large MV based update
        if $totalShift > ROI_{mean}$
          then
          $reject(g_{t-i})$
return $(G_t)$
```

Algorithm 4.5: UPDATEFACETRACKS($ROIs_t, F_t$)

```
for each $roi \in ROIs_t$
  do {
    $trackDist \leftarrow \infty$
    $roi.face \leftarrow NULL$
    for each $f_t \in F_t$
      do {
        comment: Align roi to closest face track
        comment: falling in thresh 1/3rd inter-ROI dist.
        if $(trackDist <|| f_t, roi.face ||$
        $\& || f_t, roi.face ||< neighbourhood)$

          then  $trackDist \leftarrow|| f_t, trackDist ||$
          $roi.face \leftarrow f_t$
          $\{$
      }
  }
return $(ROIs_t)$
```

An important point to note here is that, when *video-binning* is applied on G_t, motion vector update would already are applied on g_t. This is the estimation step that anticipates the position of the next gaze point g_{t+1} that will come next. Applications such as dynamic captioning, which use the updated G_t will thus place the content on the estimated position of eye-gaze for the next frame. The *video-binning* method is summarized here in the following pseudocode,



(a) (b) (c)

(d) (e) (f)

Figure 50: Panels in the top row illustrate important stages in the interactive framework. (a) A frame from the video stream. (b) Eye-gaze based ROIs discovered using the *video binning* method [46]. The red circle shows current location being attended and yellow circles show past ROIs, the arrows show dominant eye movement trajectories. (c) An example image region overlayed with motion saliency computed using motion vector information in the encoded video stream. (d) Face saliency map constructed by detecting and tracking frontal and side profile faces. Panels in the middle row visualize stages in the dialogue captioning framework. (e) Regions likely to contain faces are combined with ROI and likely eye movement paths shown in (f) to compute likely concepts of human interest. Video frame taken from the movie *swades*©UTV Motion Pictures.

4.5 Summary

This chapter describes the statistical analysis and algorithms that can be used to infer scene semantics. The analysis and computational model for detection of interesting images is described first, followed by description of statistical and algorithmic analysis of eye-movement data. The framework developed in this chapter is put to use in the applications presented in subsequent chapters.

5 Applications to Image and Video understanding

This chapter presents a variety of applications that harness a combination of eye-gaze, content analysis and meta-data for image and video understanding. The applications are highlighted in the overall schema as shown in Figure 5.

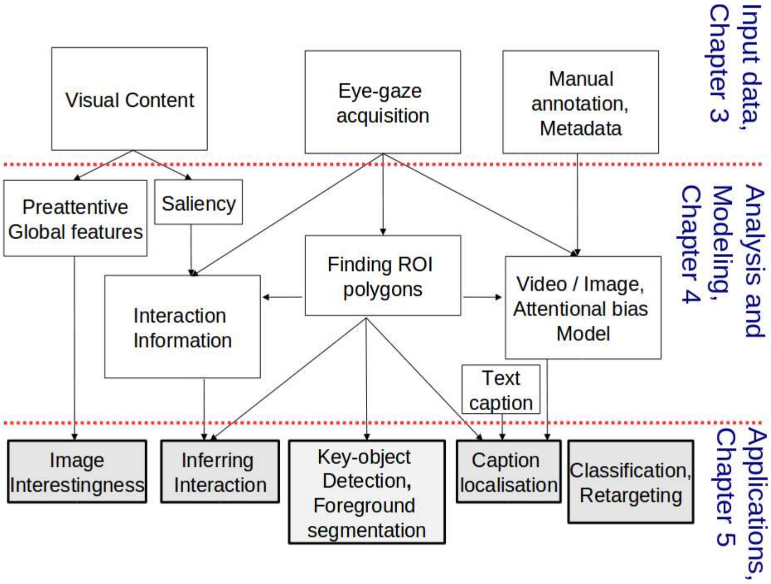


Figure 51: The figure highlights the components described in this chapter. The current chapter deals with image and video understanding applications using the framework developed in chapter 4.

The first application is based on insights from pre-attentive image classification 5.1 and does not employ eye-movements. Subsequent applications use statistical models built from eye-movements or algorithms to process eye-movement data.

5.1 Automatically predicting pre-attentive image interestingness

The insights gained from behavioural experiments for interestingness discrimination in Flickr®images is now put to use to build a model for interestingness prediction, it is important to note that the pre-attentive stage does not involve eye-fixations. The first approach to build a model of the user's perception of aesthetics in images has been done using EXIF information accompanying digital images. This novel approach has also been tried out in context to scene classification in [81]. Global image features hence turn out to be important attributes in basic tasks such as scene categorization and aesthetics discrimination. This thesis shows results relating to aesthetics discrimination [45], other groups have shown its role in scene-classification [101],[64]. Frequently occurring EXIF fields in our dataset were extracted (highlighted in Figure 52), some of these encode global image information directly or indirectly. These attributes were used to train two types of classifier models, the first type models personalized preference, ie; an individual user's notion of an aesthetic image. The second type models community-preference which is the notion of an entire community, like the community of Flickr users. The former model is termed as a *personal-agent* and the latter as a *community-agent*

Training data for the models consists of positive examples made up of highly interesting images and negative examples from semantically relevant, less interesting images. Both types of images are selected such that they are accompanied with EXIF information. Positive examples are taken from the top of the interesting list and the negative examples are taken from images which are low in the interesting list, but have good relevance scores. This is done to ensure that the negative examples are still semantically relevant. The training scheme is visualized in Figure 53.

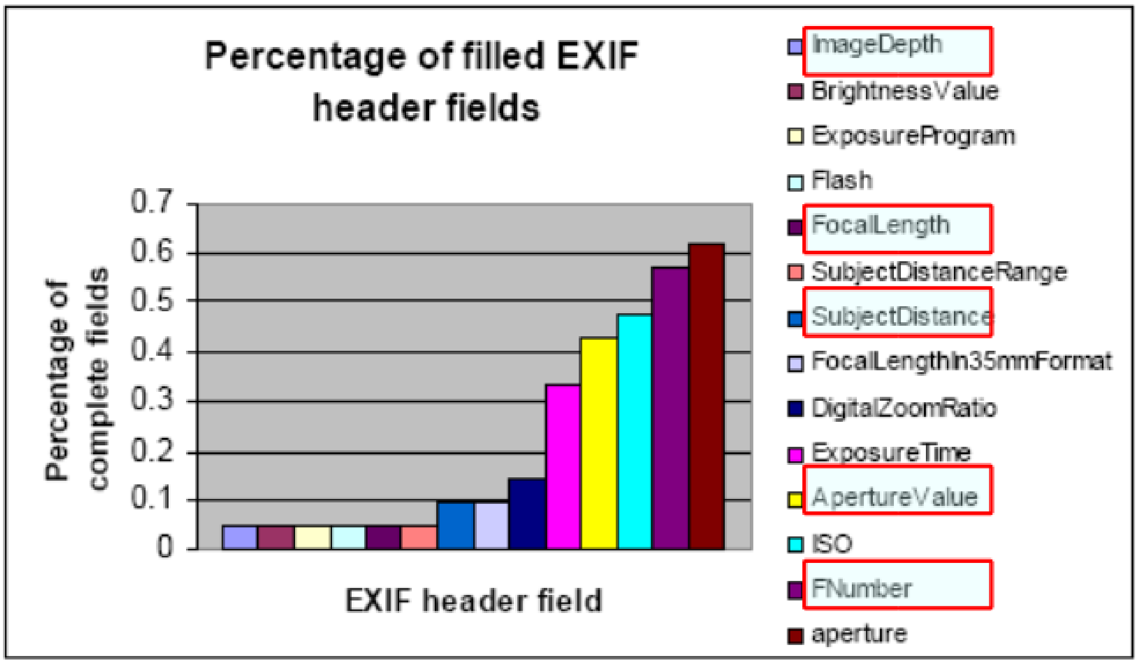


Figure 52: Normalized frequency of occurrence of different EXIF attributes in our database. Important EXIF attributes that encode global image information directly or indirectly are highlighted(boxed) in red.

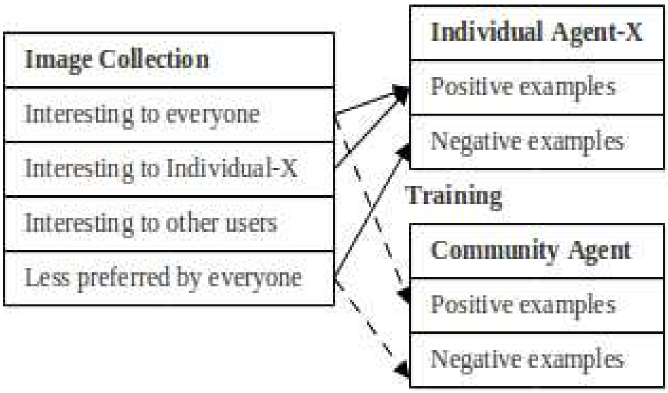


Figure 53: Appropriate subsets of the dataset can be chosen as positive and negative samples to trian individual preferences and community preferences.

The effectiveness of the agents is verified by performing SVM regression in the Weka environment (SVMreg, polynomial kernel, exponent=1) with $\frac{10}{1}$ cross-validation and $\frac{1}{3}$ split with over 2100 images containing EXIF information.The community agent yields an accuracy of 65% on classification between image with High and Low aesthetics scores. This is statistically significant, as a correct decision places the image in the set of positive examples, which typically represent a small fraction $\leq 10\%$ of the total images for that semantic theme. Thus random selection would yield success rates close to just 10%. Training the community agent with images from different users avoids the excessive influence from any particular user. The personal agent is trained from portions of the community training data belonging to the same user. This ensures that the positive and negative samples represent the preferences of a single user. Results for three user agents trained for Flickr members who have significant contribution in Flickr for the concepts chosen in our database, are shown in Figure 7.

Flickr User Id	**#Images**	**Accuracy achieved**
25056484@N00	130	60%
37985559@N00	123	53%
78779687@N00	157	55%

Table 7: The accuracy achieved by a personalized model trained for individual user's aesthetics preference.

The correlation value and accuracy obtained is limited compared to earlier work like [14] where authors use extensive content based features. The motivation of these experiments is to demonstrate that EXIF based classification adds value to any classification scheme. Though the accuracy values appear close to chance (50%) for a 2-class problem, in reality the number of classes is very large as we are modeling the individual's notion of aesthetics amongst a thousands of users in the system.

The experiments and analysis in this section establish the following facts,

- Humans can rapidly discriminate whether a briefly shown image will be interesting over a longer span of viewing. From Flickr®data, the visual properties of such images co-incide well with simple notions of aesthetics.
- Since eye-movements do not set in within pre-attentive time spans, the interestingness discrimination is done largely without understing detailed image semantics.
- The minimum presentation time for such discrimination is about 50 milliseconds.

Subsequent sections describes applications that exploit either statistical models constructed from eye-movement data, or algorithms that can analyze eye-movements over images and videos.

5.2 Applications of Attentional bias to image classification

The insights from attentional bias w_i is now used for a sequence of 2-class classification problems, namely *normal* vs *expressive* face, *portrait* vs *nude* image categories. The interaction measure $Int_{(l,m)I}$ is used for *action* vs *non-action* image classification. For classifying *face* and *person* images, automated detectors are necessary to infer the respective ROIs. However, for *action* images, where the interacting entities are spatially separated, concept detectors are redundant. This is owing to the fact that while concept detectors can only identify that there is a 'Man' and 'Book' for Fig.34(a), the presence or absence of inter-entity interactions has to be purely determined using gaze information.

The methodologies for determining ROIs automatically in *face* and *person* images are the same as described earlier. For images in which a face is detected and eyes have been localized, attention-bias weights for *eyes* and *nose+mouth* are computed using Eq.(3). Similarly, upper-body and face detectors are used to identify *face* and *body* ROIs for person images.

For classification, we perform leave-one-out cross-validation, *i.e.*, all but as training data, while the chosen one is used as test data. The training data is then used to learn representative $w_i/Int_{(l,m)I}$ for the classes involved ($Int_{(l,m)I}$ is employed for *action* images only) This process is repeated until all images are chosen for the test data. Table.8 presents the classification results for the *face* and *person* images. An overall accuracy of 69.6% and 60.2% are obtained for the *face* and *person* classes respectively. The classification is done based on thresholds for $w_i/Int_{(l,m)I}$ values obtained in earlier analysis.

Category	**Instances**	**Correctly Classified**	**Accuracy**
Normal Face	**37**	**28**	**0.76**
Expressive Face	**25**	**15**	**0.6**
Nude	**32**	**18**	**0.57**
Person	**36**	**23**	**0.63**

Table 8: Combining concept detectors and fixations to classify *face* and *person* images.

Results obtained for 70 *action* images are presented in Tab. 9. Overall, correct classification is achieved for 62.5% of the images.

Category	Instances	Correctly Classified	Accuracy
Action	**34**	**21**	**0.62**
No Action	**36**	**23**	**0.63**

Table 9: Using eye-gaze information to classify for *Action* and *No Action* social scenes.

5.3 Application to localization of key concepts images

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* [74]. Difficulty in determining image objects (concepts) from visual content has necessitated image retrieval algorithms [28] 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) [10], 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 [86] to establish the *importance* of some scene objects, even in simple scenes. Also, observations made from eye-gaze statistics in [18] suggest that humans are attentive to *interesting* objects in semantically rich photographs.

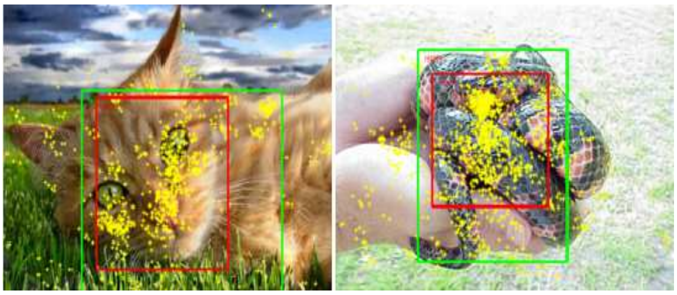
This problems is similar in spirit to [73], where caption text and image segments are combined to localize the subject of a natural image. On the other hand, we focus on localizing attention grabbing and emotion evoking concepts in images. Contrary to the notion that human subjectivity influences the choice of interesting scene objects, we observe that consistently fixated upon by a majority of subjects. These concepts may correspond to individual objects or interactions between two objects (actions). The **attentional bias** model for world encodes world ontology as a tree, whose vertex weights denote concept importance. This helps localize the most important and affective concepts corresponding to the caption of an unlabeled image.

5.3.1 Steps followed

The proposed method 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.55). 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* [21]. 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 ROIs 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 [16] enables refinement of fixation clusters, which are noisy.



(a) (b)

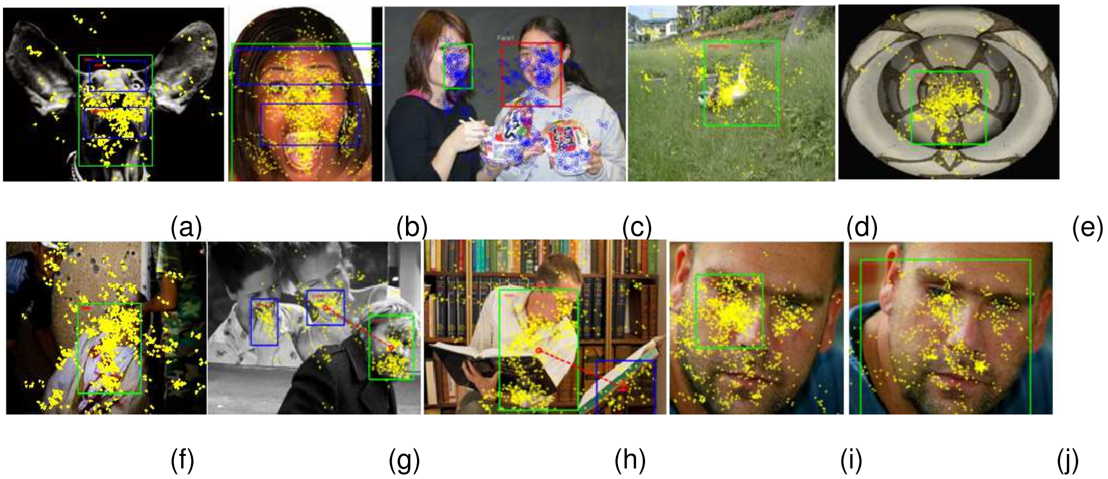
Figure 54: Color-homogeneous cluster (red) obtained from original fixation cluster (green) on (a) *cat face* and (b) *reptile*. Fixation points are shown in yellow.

A more accurate localization of the ROIs is possible by retaining only those cluster points that correspond to homogeneous color segments (Fig.54). For some concepts like *face*, ROI 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.

- *ROI-based post-processing for action localization*- Upon localization of ROIs corresponding to affective objects, actions can be inferred from extensive fixation transitions between interacting objects, as described in Section 2.2.

5.3.2 Results

Localization of *italicized* objects and actions from textual image captions is demonstrated in Fig.55. Blue rectangles in (Fig.55(a),(b)) correspond to *face* sub-concepts localized through recursive fixation clustering. For action images (Figs.55(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 ROIs localized in Fig.55(h),



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

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

Figure 55: Affective object/action localization results for images with captions (a) A dog's *face* .$aoi's : eyes, nose + mouth, face$ (b) Her *surprised face* said it all! $aoi's : eyes, nose + mouth, face$(c) *Two girls* posing for a photo. $aoi's : face1, face2)(d)$ *Birds* in the park. $aoi's : bird$ (e) *Lizard* on a plate. $aoi's : reptile$ (f) *Blood-stained* war victim rescued by soldiers. aoi's:blood (g) *Two ladies looking* and laughing at an *old man*. $aoi's : face_1,\ face_2,\ face_3$ (h) *Man reading* a *book*. $aoi's : human, book$ (i) Man with a *damaged* eye. $aoi's : damage$ (h) Fixation patterns and face localization when the damaged eye is restored. $aoi's : face$

$CP_{2,1_I} = 0.351$ and $CP_{1,2_I} = 0.071$, which enables assignment of labels to AOI_1, AOI_2 as *man* and *book* respectively. The *look* direction in Fig.55(g) is inferred similarly ($CP_{p,q_I} = 0.361$). While labels assigned in multiple object and action images may not always be correct, the accuracy of gaze-based labeling can improve tremendously when used along with object recognition algorithms. For a representative set of 50 unlabeled images, correct labeling of affective concepts from image caption text is achieved with 80% accuracy using the attentional bias model. Localization to a wrong ROI is considered as a failure, the method works best for *face* images.

5.4 Applications of interaction discovery to image classification

Use of the binning method to identify important visual elements and possible relationships between them has been described earlier in Section 4.3.3. The binning method 4.3 can be used to compute scores that indicate the presence of strong visual elements. The accumulation of scores is described earlier in Eqn. 14 in the previous chapter 4. 100 images were chosen at random covering the themes *action*, *aesthetic*, *affective* and neutral. The images were chosen from NUSEF [71].

Lower values are observed on average for images having strong visual elements and interactions. Fig. 56 shows good separation of images with strong visual elements and interactions from the themes *action, aesthetic, affect* plotted in red, green and blue respectively, in contrast to the high values observed for images with weakly interacting elements as shown in the magenta plot. Separating out different themes like affect and action from the former type is part of future work.

5.5 Application of ROIs discovered using the binning method, Image segmentation

ROIs discovered using the binning method can be used to segment out foreground objects of human interest. An algorithm for automatically segmenting the image region containing a fixation point is described in [60]. 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

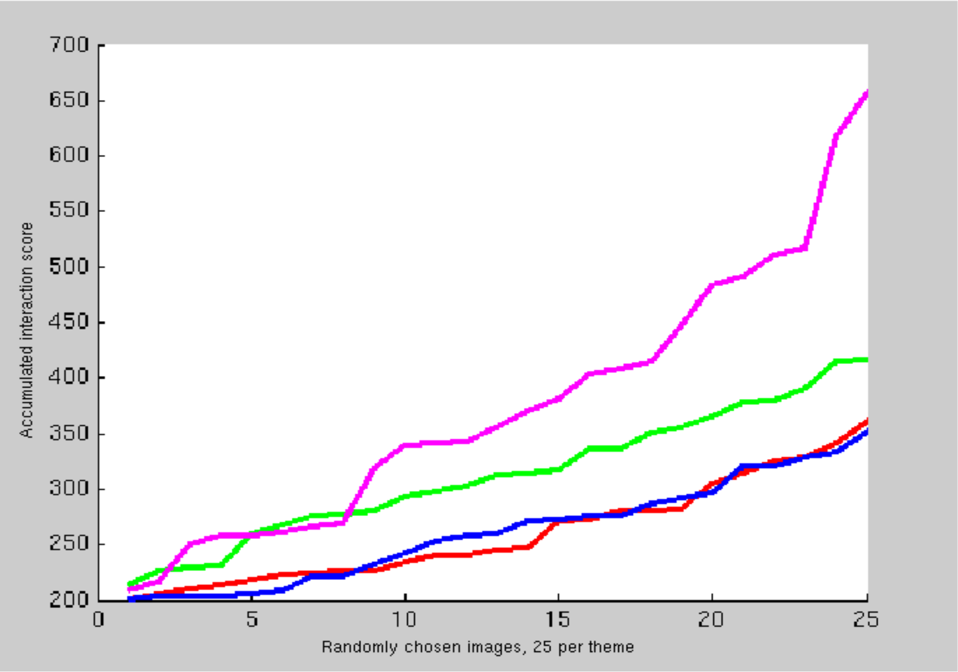


Figure 56: Discrimination obtained by the cluster profiling method, the vertical axis plots accumulated scores for different images measured using equation 14. Distinct images have grouped under each of the 4 themes, the plot represents values over more than 100 images. The method separates out images with strong visual elements and interactions *affective*-red,*aesthetic*-green and *action*-blue from those which have low interaction or weak visual elements (magenta). *action* and *affect* images are grouped together by the measure described earlier in 14, this needs to be investigated further.

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.

(a) To determine whether the segmentation performance of [60] 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 fmeasure for the 20 segments obtained from the random seeds. The fmeasure, which is used as a measure of the segmentation performance accuracy, is defined similar to Equation 9.

(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.57 and 58, the centroid of the largest fixation cluster generally lies within the *salient* object, and therefore, the segmentation output with the centripetal fixation seed should be comparable to that obtained in (a). As seen from Fig.57 (rows 2 and 3), using the centroid 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 Euclidean 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 incorporating this information in the energy minimization process. In particular, we reinitialize 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.57, 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.58). 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.

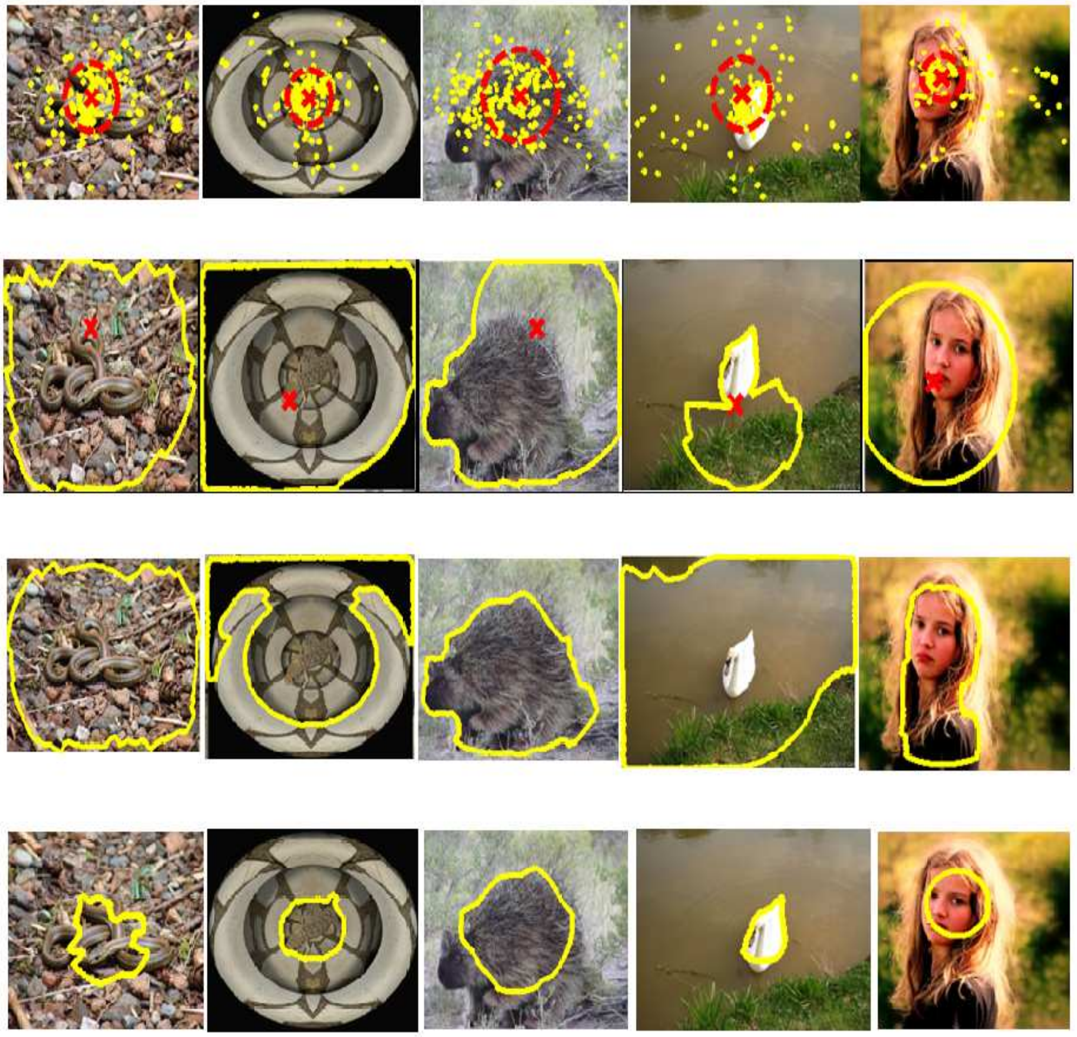
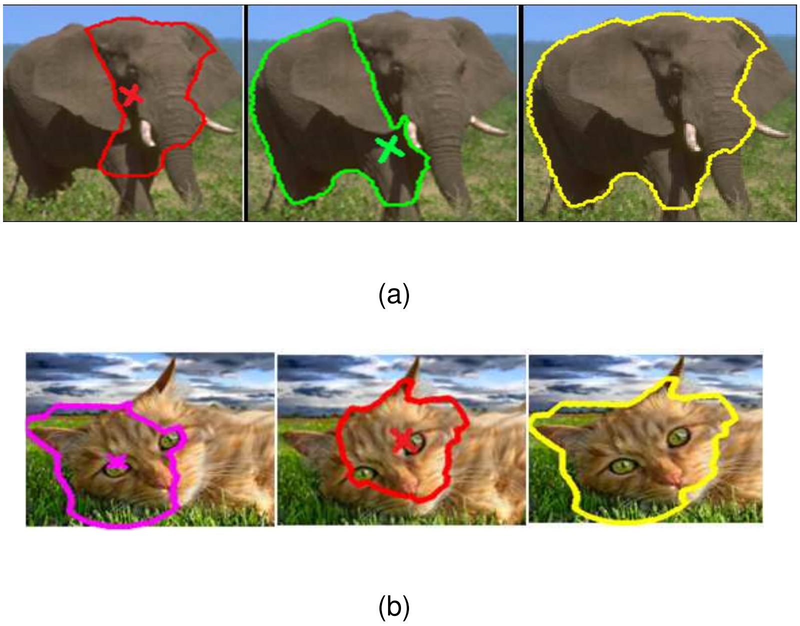


Figure 57: Enhanced segmentation with multiple fixations. The first row shows the normalized fixation points (yellow). The red 'X' denotes centroid of the fixation cluster around the *salient* object, while the circle represents the mean radius of the cluster. Second row shows segmentation achieved with a random fixation seed inside the object of interest[60]. 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)

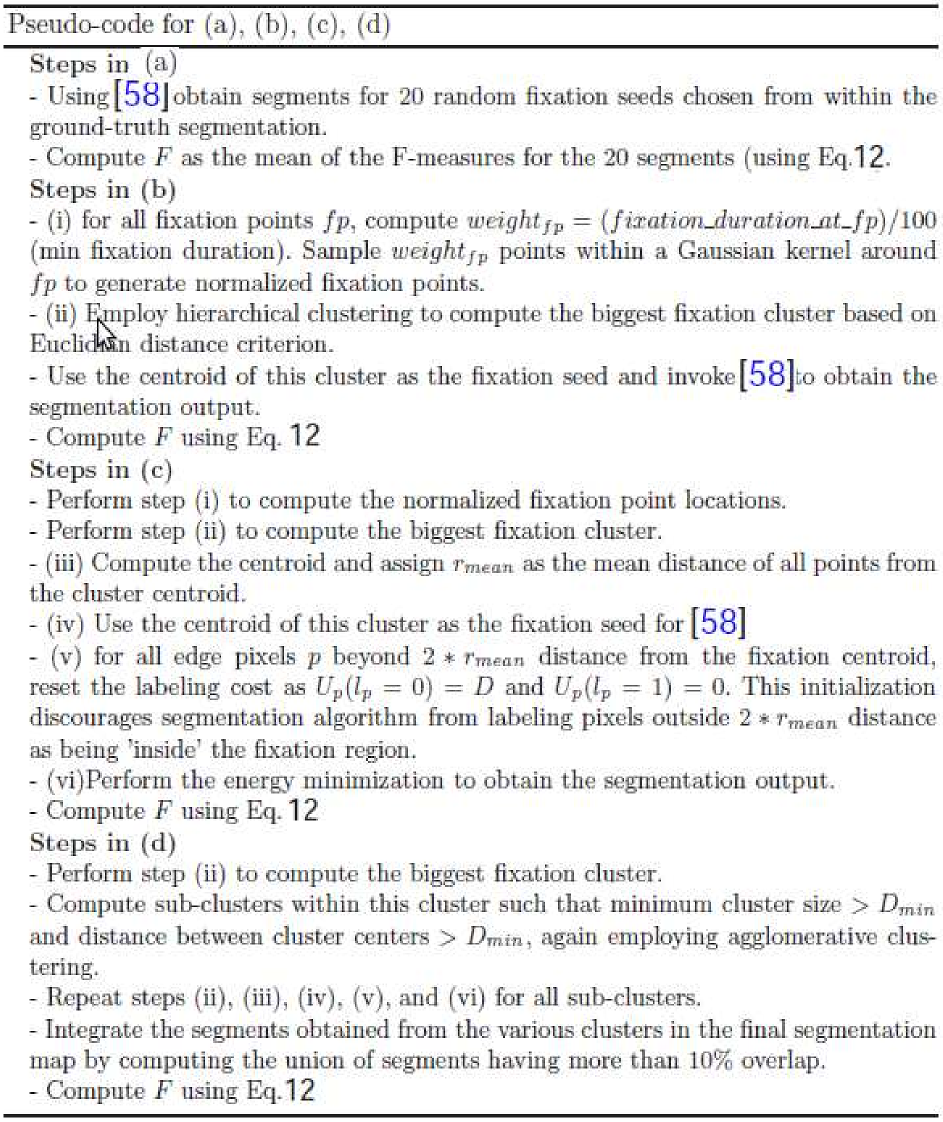
(b)

Figure 58: 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.15.

The pseudo-code summarizing the steps involved in (a), (b), (c) and (d) is provided in the pseudocode described in panel 59.

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 fmeasure is used for evaluating segmentation accuracy. For the baseline method, the mean fmeasure 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 fmeasure is computed. The fmeasure



```
Pseudo-code for (a), (b), (c), (d)
Steps in (a)
- Using [58] 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.12.
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 Euclidean distance criterion.
- Use the centroid of this cluster as the fixation seed and invoke [58] to obtain the segmentation output.
- Compute $F$ using Eq. 12
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 [58]
- (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. 12
Steps in (d)
- Perform step (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.12
```

Figure 59: The pseudocode describes details of steps (a) to (d).

scores for segmentation procedures (a), (b), (c) and (d) are tabulated in Table 10.

Table 10: Performance evaluation for segmentation outputs from (a), (b), (c) and (d).

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

The fmeasure 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 method can isolate the foreground more accurately for difficult cases (shown in Fig.57), 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.60 presents the fmeasure 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.

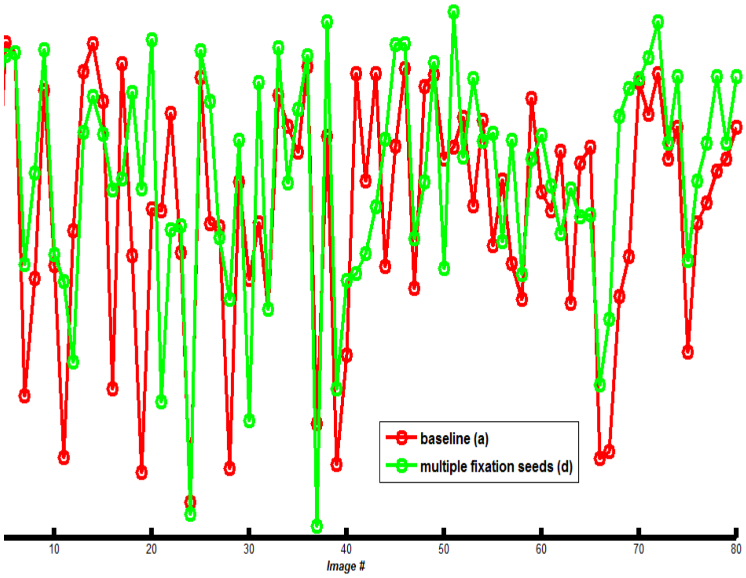


Figure 60: F measure plot for 80 images showing the improvement brought about by using multiple fixation seeds for segmentation (d) in comparison to the baseline (a) using equal number of random locations within the object as segmentation seeds. The legend is as follows - *red* baseline and *green* - Integration of segments obtained from multiple sub-clusters.

5.6 Application of ROIs discovered using the binning method, guiding object detectors

Object detection in natural images remains an open challenge in computer vision, despite some success in detecting faces, people and other important generic concepts in recent PASCAL VOC challenges [23]. Performance of state-of-art detectors is far behind the practical needs of image-understanding systems that attempt to index, tag or summarize natural images. We present a novel way to demonstrate the effectiveness of human eye-gaze and using it to guide a state-of-art object detector. The general schema of a sliding window based detector and guidance using eye-gaze information is shown in Fig. 61

Sliding window based object detectors are essentially image classi-

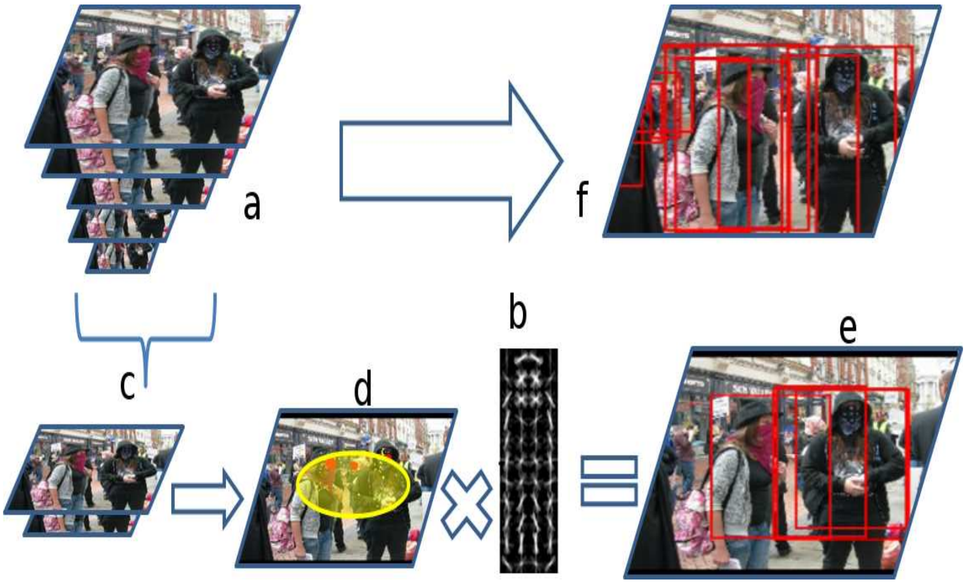


Figure 61: The schema for guiding sliding window based object detectors using visual attention information. Image pyramid (a) is obtained by successively resizing the input image *I* over L levels. Features corresponding to areas covered by sliding, rectangular windows at each level l_i are combined with a template based filter (b) to generate scores indicating presence of the object. These are combined over all levels that indicate the presence of the object. Eye-gaze information is used to extract Regions of attention (ROIs) (d), which then restrict the image region for object search. The number of scales (c) are restricted to a small fraction of possible levels, using scale information from ROIs (e) is the output from our method and (f) from a state-of-art detector [23].

fiers. The classifier is used to exhaustively inspect rectangular regions over successively scaled down versions of the input image. Classification performed over each window generates detection scores, these scores are then combined across multiple scales to give image regions with maximum likelihood for presence of the object. Lacking prior knowledge of the location or size of key objects in the image, object detectors search exhaustively through an exponentially large search space of windows. For example, a 1024×768 image consumes 15-20 seconds to be searched in totality by the [23] detector on a standard PC *(Pentium Core 2 Duo, 2 Ghz, 2 Giga bytes RAM)*.

Similar time durations are taken by other detectors such as [94]. We demonstrate the effectiveness of visual attention information captured by eye-gaze, to guide a state-of-art detector [23]. We also show an improvement in precision and recall within an interactive time-span of a few seconds.

5.6.1 Using eye-gaze based ROIs

ROIs generated from eye-gaze data are passed as input to the object detector.The first level of reduction of the search space is achieved by limiting the search within *ROIs* obtained from eye-gaze information. The detector is then limited to operate on a fraction of the scales in the image pyramid that is illustrated in Fig. 61. The scales chosen are the ones where the sliding window size is close to the size of ROIs. This, in turn, is close to the size of the object itself.

More formally, trained model is a filter *F* of size $w \times h$. Let *H* be the feature pyramid extracted from the successively resized images as illustrated in Fig. 61 and *pos(x,y,l)* be a position *(x,y)* in the *l th* level of the pyramid. The score of *F* at *pos(x,y,l)* then is

$$F^{*}\psi(H, pos, w, h), \tag{18}$$

where,

$$\psi(H, pos(x, y, l), w, h) \tag{19}$$

is the vector obtained by concatenating feature vectors in the $w \times h$ sub-window at level l. The final likelihood is then obtained by combining scores so obtained across different levels. The exact score generation and combination strategy varies across specific detector implementations and is not governed by our method. Feature extraction, pyramid construction and likelihood generation follows the implementation in [23].

In the following sections, we describe our approach to solve some challenging problems encountered while using eye-gaze information to guide object detectors. We begin by describing our approach for good ROI size estimation, this is followed by the case where ROIs capture key object parts and object size needs to be inferred by a *sum-of-parts* approach. Finally we address the cases in which state-of-art detection can give positive detection for multiple categories over the same ROI.

5.6.2 ROI size estimation to reduce false positives and low relevance detections

A template based object detector operating at multiple scales is prone to lot of false positives. Some examples can be seen in Fig. 65 *(e)* and *(f)* where many false detections of size different from that of the object are returned by [23]. Such false positives can arise due to non-object image regions responding to template based matching as classification progresses through different levels of the image pyramid. It then becomes important to make a correct choice of levels in the pyramid.

For every ROI that is passed to the object detector, we enforce scale selection by choosing levels *l* such that *l* $\subseteq$ *L* and area of resized *ROI* is close to the sliding window area at these levels,

$$\frac{\textit{area(l)}}{w \times h} \approx \textit{1} \tag{20}$$

The effectiveness of this approach is demonstrated in Fig. 65 *(a)* & *(b)* where most detections are very close to actual size of object of interest. An example of the significant improvement that is made possible by the right combination of *ROI* size and scale selection is illustrated in Fig. 65 *(b)* where both objects of interest are correctly identified by our method. This is in contrast to Fig. 65 *(f)* where the detector from [23] fails to detect the objects, but also throws up many false positives. Selecting a few levels in the image pyramid starting from a *ROI* is akin to creating a partial pyramid with finer grained resizing scales. This enables detection of objects at sizes that would be missed out in the larger image pyramid. A badly estimated *ROI* however inevitably leads to negative results as shown in Fig. 65 *(d)*.

We address the problem posed by less relevant instances of objects that are not central to the meaning of the image. For example in Fig. 65 *(a) & (e)*, one can expect that the *woman* is the key object in the image and people in the background do not contribute equally for image-understanding. Another example is Fig. 64 *(a)* where *man & woman* appear to be key objects. This is confirmed from eye-gaze clusters across many subjects as shown in Fig. 64 *(c)*. This is further reinforced by manual ground truth annotation boxes for *person* by human annotators asked to identify interesting and key objects. By localizing object search to eye-gaze based *ROIs*, we are able to improve the quality of detection as illustrated in Fig. 64 *(g)* against that by [23] in Fig. 64 *(h)*.

5.6.3 Multiple ROIs as parts of an object

The relatively uncontrolled nature of our dataset presents objects and their parts at different scales. Though good estimation of ROI sizes as described in the previous section helps restrict detector search space and reduce false positives in detection results, it doesn't address the problem of object-scale estimation. For this purpose, we assume each ROI to either be one key part of an object or the entire object itself. Thus each object O_j can be represented by a combination one or more ROI_k detected from eye-gaze data. We generate a set of candidate *meta-ROIs* from combinations of one or more ROIs. The object detector is then run over each of these *meta-ROIs*, successful detections with high scores on these *meta-ROIs* help infer the presence and correct scale of an object. Good results were obtained using this strategy for objects with large sizes, such as *person* where different parts were identified as an ROI and the detector is able to identify the whole as opposed to the parts.

5.6.4 Experimental results and Discussion

We demonstrate the applicability of our method using the generic object detector in [23], which has been the top performing system in the recent PASCAL VOC 2009 challenge [23]. A diverse set of images capturing a variety of object at different scales, have been chosen from IAPS, Flickr, Photo.net and Google search results. These represent affective, consumer, amateur photography and web images. Eye-gaze data has been obtained from human free-viewing experiments over 2000 such images over a variety of themes. A subset of

200 images is chosen for evaluation from *person*, *dog*, *cat* and *bird* concepts.

Fig.62 illustrates the computation times taken by our method (red) as compared to the baseline $(green)$ to detect the concept *person* in close to 150 natural images. The images are social scenes with varying number people in many poses, at different depths-of-field. We obtain a significant improvement of 80% over the baseline as illustrated in Fig. 62.

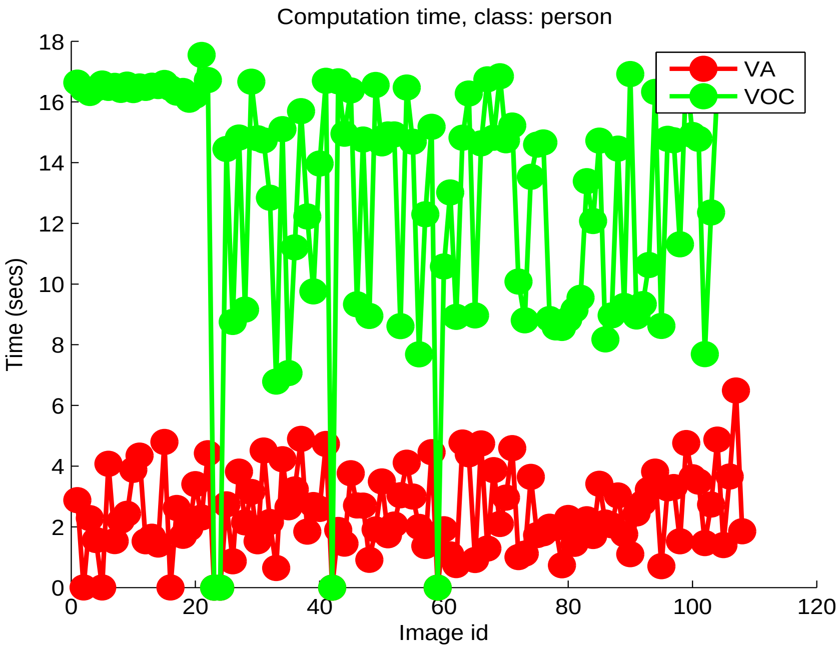


Figure 62: Illustration of the significant reduction in computation time achieved by constraining the state-of-art object classifier in [23] using eye-gaze information.

We combine $precision$ and $recall$ against the baseline using an $fmeasure$ score computed over detection boxes (*bbox*) with respect to human annotated ground-truth (*gtruth*) boxes using definitions in Equations 10,11 and 9 respectively.

The evaluation over 150 images from the concept *person* yields a 18% improvement in $fmeasure$ as illustrated in Fig. 63.

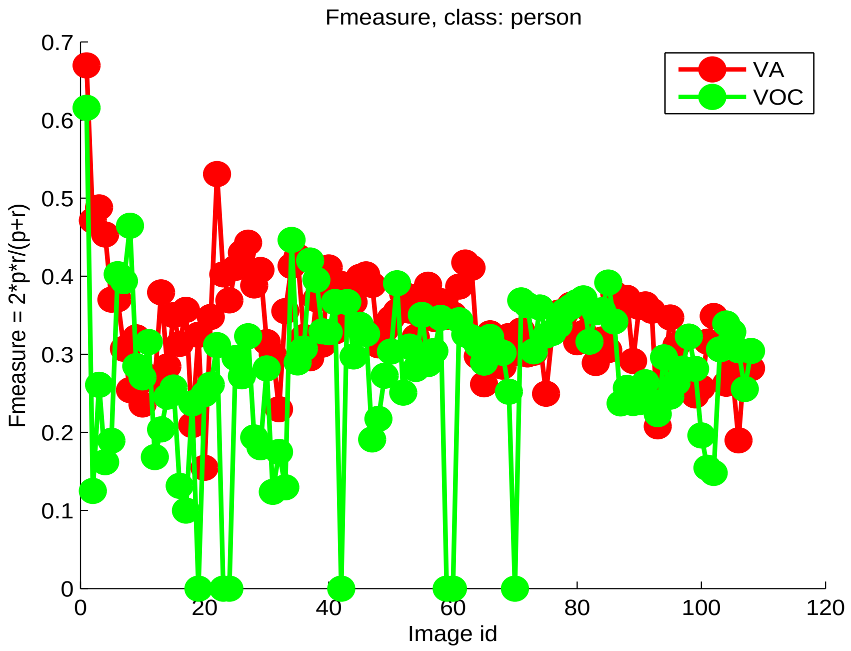
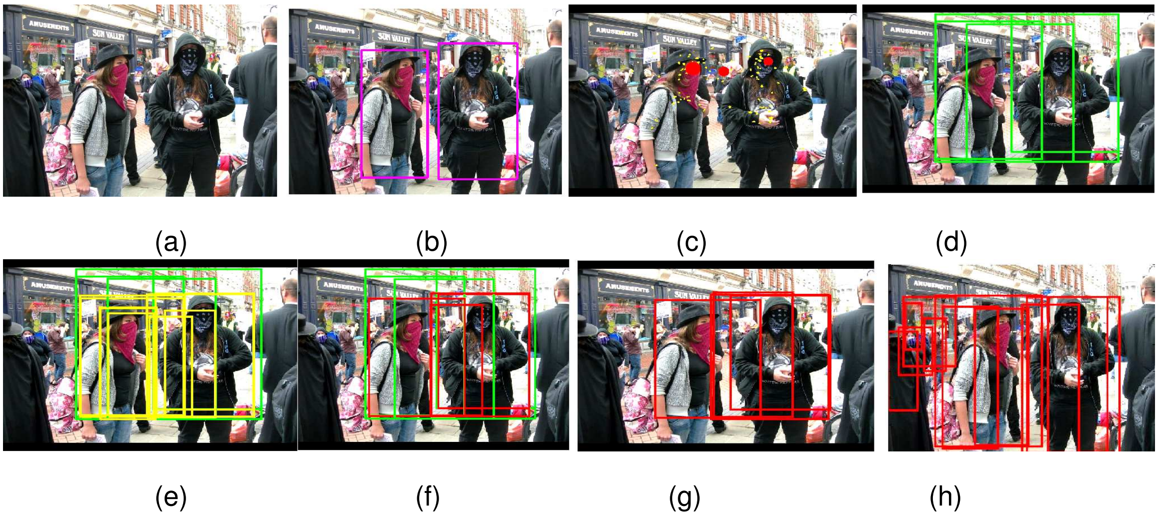


Figure 63: Illustration of the improvement in $fmeasure$ of over 18 % achieved by constraining the object classifier in [23] using eye-gaze information. $fmeasures$ are recorded from our method *VA* and that of [23] attempting to find the concept *person* over 150 images. The images were chosen to capture diversity in number of instances, size, activity and overall scene complexity.

Our method is independent of the application that the ROI boxes are put to and in this case the specific object-detector employed within the ROI boxes. We demonstrate this by running the method over the classes *person*, *dog/cat*, *bird* and use the ROI boxes generated from eye-gaze data as detections by a hypothetical detector. The $fmeasure$ is evaluated and yields interesting and satisfactory scores as illustrated in Table.11.

5.7 Applying *video binning* to *Interactive* and *online* dynamic dialogue localisation onto video frames

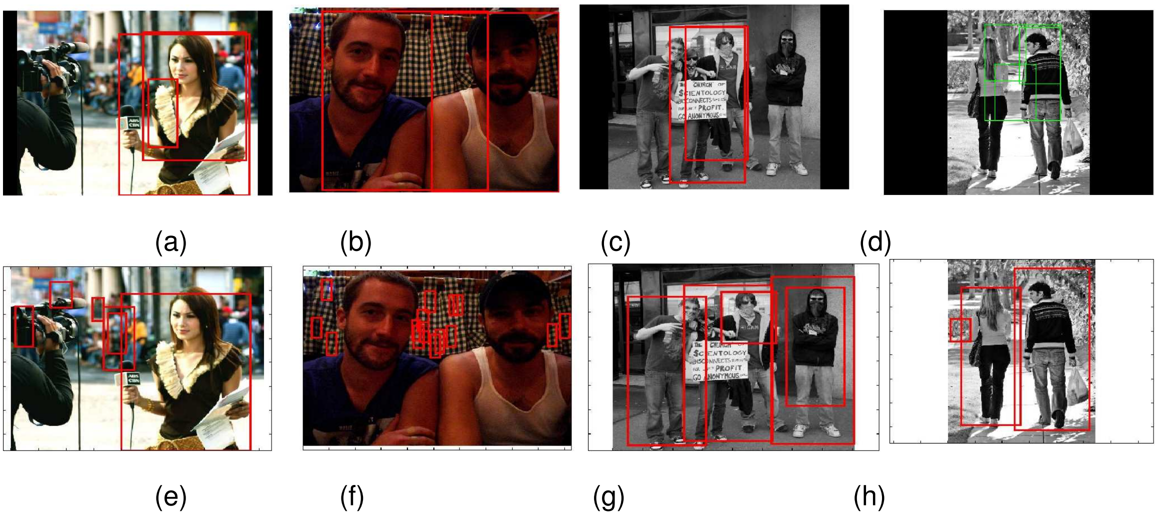
Dialogues are obtained from the public domain Opensubtitles.org, the dialogues are in English and are time-stamped. After ROI discov-



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

(e) (f) (g) (h)

Figure 64: The panel illustrates outputs at every stage of our attention driven method. (a) Original image of a crowded street-scene. (b) Manual ground truth annotation boxes for key objects. (c) Clusters identified from eye-gaze information, centroids marked by red circles. (d) ROIs generated based on cluster information. (e) Detected instances of *person* class within ROIs, using detector from [23] marked by yellow boxes. (f) Finally result detections after filtering for ROI size, marked by red boxes. (g) Results for the same image from the baseline detector.



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

(e) (f) (g) (h)

Figure 65: (a),(b) Cases where visual attention greatly enhances performance of the detection system, (e),(f) are the corresponding results for (a),(b) from the multi-scale, sliding window method in [23]. (c) A case where attention directs ROIs away from non-central, but seemingly important *person*s. This problem is not faced by the baseline as seen in (g). (d) Generated ROIs are not good enough to permit detection, the baseline outperforms our method in this case as seen in (h).

Category	Fmeasure-VA	Fmeasure-VOC
Person	**0.3**	**0.34**
Bird	**0.41**	**NA**
Cat or Dog	**0.43**	**NA**

Table 11: Evaluation of the visual attention guided *ROIs* against human annotated ground truth. The object detector is not run in *ROIs* and instead the entire ROI is considered to be a detection, this experiment illustrates the meaningfulness of the ROIs generated by our method against human annotated ground truth boxes for different concepts in our database. This is especially significant in cases like *bird and cat/dog*, where the baseline detector fails completely.

ery and propagation using *video-binning* as described in Algorithm 4.3, the next step is to identify suitable locations to render dialogue text. Eye-gaze is a noisy signal, so are the bottom-up motion vectors and top-down face tracks, this in turn brings instability to $ROIs$ positions detected over successive frames. Resultant jittery dialogue placement can be extremely annoying. Hence, the dynamic captioning system maintains a history of past positions where a particular dialogue is rendered and averages these values to generate stable dialogue positions.

Dialogue position initialization is very important in dynamic captioning, as wrongly positioned dialogues not only confuse the viewer, but also distract badly as they *drift* into the correct position eventually by virtue of the dialogue position buffering just described. We exploit inter object relationships or interactions that are exposed during *video-binning*, an example has been visualized in 66 (f). We choose dialogue positions near ROIs, in a manner that the dialogue will not intersect eye-movement trajectories seen in the recent past. This is done because meaningful stimuli such as text will interfere in normal exploratory eye movements occurring between ROIs. This is quite different from the approach in [31], where low level saliency is used to position text to the left, right or top of a detected face. We are in the process of evaluating the impact of this strategy. After computation of optimal display location, dialogue content is overlayed on the video frame as a dialogue *blurb*. The video frame is then displayed to the subject and eye-gaze data tracked. The system is implemented in the Visual C++®environment using the ERICA®development toolkit.

5.7.1 Data collection

Video data has been chosen from Youtube® to represent a variety of themes based on three parameters, extenuate of activity involved, spoken language in the video and whether the video has been shot indoors or outdoors. The original video clips are of 5 minutes duration. The clips were normalized and re-encoded to have either a 640 pixel height or 480 pixel width, without altering the aspect ratio. Video clips were of sufficient quality for subjects to identify details such as lip movements. All clips are social scenes.

5.7.2 Experiment design

The experiment was designed in two stages. In the first stage, the subjects watch one of 3 variants of a video clip,

- Original video clip with audio track *original*.
- Video with static caption at the bottom of the video frame *s-caption*.

Video clip	Theme	Location	Language	Activity level
mmtat	*Animated clip*	*Outdoor*	*English*	*medium to high*
jbdy	*Comedy scene*	*Indoor*	*Hindi*	*high*
dotrc	*Social scenes*	*Indoor and Outdoor*	*Mandarin*	*low to high*
village	*Social scenes*	*Outdoor*	*Hindi*	*medium to high*
hockey	*Sports, Hockey*	*Outdoor*	*Hindi*	*medium to high*
goal	*Sports, Soccer*	*Outdoor*	*English*	*medium to high*

Table 12: Description of the video clips chosen for evaluation of the online framework and applications. The clips were obtained from the public domain and normalized to a 5 minute duration. The clips were chosen from amongst social scenes, to have variety in the theme, indoor and outdoor locations, spoken language and extent of activity in the video clip.

- Video with interactively rendered dynamic captions using online eye-gaze information *d-caption*.

Subjects are shown 9 to 10 clips of 5 minute duration over an hour long session, a short break is given between successive clips. The clips are ordered in random fashion to avoid systematic biases. After watching each clip, the users are asked to rate each clip on their overall understanding of the clip and in case of captioned clips, the utility of text captions in their understanding of the clip in case of a captioned clip. At the end of this phase of the experiment, the users also give their preference of the captioning strategy. Subjects also give their feedback on the overall interestingness of the clips and any drawbacks or advantages seen in the captioning methods. Video clips are shown randomly in one of the following modes described in Table. 13

Mode	Number of views	**Manipulation**	**Motivation**
original	60	**Both audio & video, no caption**	**comparison baseline**
s-caption, blurb	25	**s-caption, blurb**	**comparison baseline**
d-caption, blurb	29	**d-caption, blurb**	**evaluation**
s-caption, no blurb	41	**s-caption, text only**	**comparison baseline**
d-caption, no blurb	40	**d-aption, text only**	**evaluation**

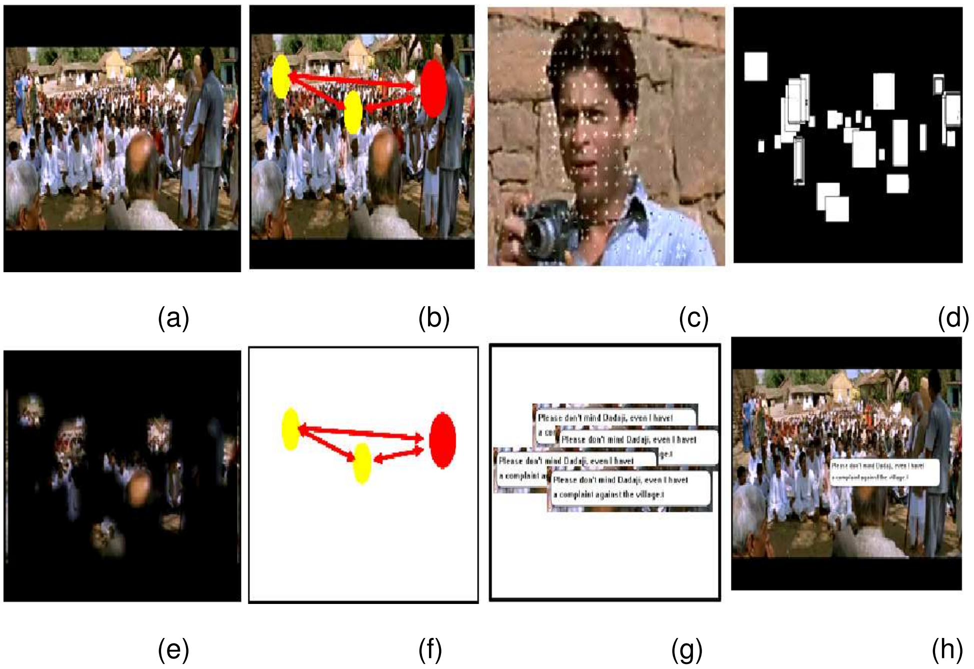
Table 13: Different modes in which video clips were shown to subjects during evaluation of the online, interactive captioning framework. The first 8 participants were shown captions with opaque or semi-transparent blurbs, subsequent participants saw text-only dialogue captions.

5.7.3 Evaluation of user attention in captioning

For objective evaluation, we measure two parameters to assess changes in the subjects behavior while watching the baseline closed captioned video and that in the eye-gaze driven dynamic captioning. The first measures overlap between eye-gaze ROIs in the two modes, the second measures proportion of eye-fixations allotted to ROIs and caption content respectively. In the case where the subject knows the language being spoken in the video clips, the second parameter can be treated as a measure of distraction due to the caption.

Subjects are asked the following subjective questions are to assess their comprehension of the video clip and effectiveness of the captioning strategies,

- Is eye-gaze driven placement easier and more natural to follow and understand ? (5 more natural, 1 not as good as original)
- Were captions presented at appropriate locations in video (5 helpful, 1 annoying)


Please don't mind Dadaji, even I havet
a complaint against the village.

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

(e) (f) (g) (h)

Figure 66: Panels in the top row illustrate important stages in the interactive framework. (a) A frame from the video stream. (b) Eye-gaze based ROIs discovered using the *video binning* method [46]. The red circle shows current location being attended and yellow circles show past ROIs, the arrows show dominant eye movement trajectories. (c) An example image region overlayed with motion saliency computed using motion vector information in the encoded video stream. (d) Face saliency map constructed by detecting and tracking frontal and side profile faces. Panels in the middle row visualize stages in the dialogue captioning framework. (e) Regions likely to contain faces are combined with ROI and likely eye movement paths shown in (f) to compute likely locations to place the dialogue currently in progress. (g) Video sequences are dynamic and object motion as well as camera motion cause change in position of the dominant objects over successive video frames, this combined with noisy eye-gaze ROIs in turn gives rise noticeable and annoying jitter. (h) A history of dialogue placement locations is maintained and smoothed over to obtain smooth movements of overlayed dialogue boxes across the screen. Video frame taken from the movie *swades*©UTV Motion Pictures.

5.7.4 Results and discussion

We explore some important results and insights in this section. The online framework is taken up first, followed by the captioning application. We then describe the results from evaluation of video summaries and story boards.

5.7.5 The online framework

The online framework has low computational cost, making it possible to performs dynamic captioning on-the-fly at 25 frames per second, as the user watches the video. We are currently detecting frontal and side profile faces beforehand using OpenCV®, this makes the overhead for face track maintenance and Algorithm 4.4.1 to within 10 milliseconds per frame. Overheads on dialogue placement and rendering are negligible, we plan to incorporate online face detection using ROIs discovered from *video-binning* as priors for location. The entire framework and dialogue captioning was implemented in C++ for efficiency and Windows®GDI api's were used for rendering. The moving window buffer of gaze points and dialogue position buffers are stored internally as run time data structures for speed. We evaluated dialogues with and without blurbs (surrounding boxes). Both static (*s-caption*) and dynamic (*d-caption*) captions were rendered with opaque, semi-transparent (black text on white blurb) and text-only (gray text) modes and sans serif fonts were used for easy readability.

5.7.6 Lessons from dynamic captioning

Blurbs in captioning interfere with the subjects understanding of the visual scene. This was reported in [31], but severity is was more for subjects capable of hearing and having conventional exposure to video content.We collected subject feedback about the overall experiment from users and found a majority objected to the presence of a blurb box.

More importantly, we find that dynamic captioning can introduce a considerable cognitive overload of tracking two important targets, one visual in nature and another having text content. Contrary to the idea that a dialogue rendered near the speaker would aid in comprehension [31], we find that a better strategy for dynamic captioning is to initialize the caption at a good location and keep it there, instead of letting it float across the screen with the speaker. Though counter-intuitive, this can also be seen on comparing comprehension scores of 3.2 for dynamic caption with only initialization, as compared to 1.1 for dynamic caption with movement of blurb in the Table. 14. An additional set of clips were generated by asking subjects to indicate ideal dialogue placement using a mouse as proxy for eye-gaze. This input was then used to initialize the position of dialogues for subsequent viewing by other subjects. Five participants were chosen for this task and were part of the overall subject pool. Clip comprehension and overall subjective feedback improved as the dialogues were constrained to the location of initialization and furthermore as the initialization locations were previously input manually.

We also explored the notion of manually initialized dynamic caption positions by getting subjects to indicate caption placement locations using the mouse as proxy for eye-gaze. It turns out that replaying these positions for other subjects gives an improved score of 3.8 for dynamic captioning that are closer to 4.4 for static captioning as seen in Table14. This is a significant improvement from a score of 1.1 for dynamic captioning with dialogue movement. Hence, the main challenge in dynamic captioning might not really be to follow key speakers, but reduce interference with the user's cognitive process. This might be a potentially important aspect to consider for any application aiming to introduce dynamic visual content into videos, such as advertisement logo insertion [105], etc. The overall comprehension was reported to be best for static captions, but on doing further analysis we discovered the possible influence of long-term habituation to static captions. This is explored next.

Mode	***Number of clips***	***Avg. Clip comprehension***
d-caption, initialize and track, no-blurb	***21***	***1.1 out of 5***
d-caption, initialize only, no-blurb	***19***	***3.2 out of 5***
d-caption, manual placement, no-blurb	***10***	***3.8 out of 5***
s-caption, no-blurb	***41***	***4.4 out of 5***

Table 14: Changes in clip comprehension when text caption placement in constrained in different ways. A clip is counted only once for the mode in which it is shown for the first time in a subject's viewing list. Floating dialogue captions were found to be very annoying and subjects also report that it hindered their comprehension. This is also visible in the *comprehension* value for the first row. The clip comprehension and overall subject feedback improved as the dialogue captions were restrained to the initialization locations. An additional *manual placement* mode was generated by using the mouse as a proxy for eye-movements and this improved the user feedback and comprehension slightly.

5.7.7 Effect of captioning on eye movements

On analyzing eye gaze patterns for *s-caption* and *d-caption* modes, we obtained interesting insights as to why subjects experience high cognitive load and interference in dynamic captioning. This analysis was done on no-blurb captioning to avoid additional influence of the

blurb itself on eye movements. Our initial impression was that the dynamic caption movement across screens might cause visual artifacts such as flicker. On the surface, eye movements in dynamic captioning appear to be more natural and closer to that, while viewing the original video clip. On the other hand, *s-caption* causes frequent eye movements to the bottom of the video frame where static captions are rendered. We discover that floating dynamic attract user attention very significantly, possibly due to the importance of the text content. Text based visual saliency has also been pointed out earlier in [11].
We compute the fraction of eye movements *attracted* by dialogue text in contrast to eye movements between ROIs. In effect we measure the proportion eye-movements allotted to tracking and understanding dialogue content. On analyzing 75 random clips, 25 from *s-caption*, 25 from *d-caption, initialize and track* and 25 from *d-caption, initialize only* modes respectively. The results show that 39.7 % of gaze points are drawn to the dialogue box in *d-caption, initialize and track* as compared to only 12 % for static caption as seen in Table. 15. In our implementation, we also avoided the influence of such eye-movements by rejecting those classified as ROI to dialogue or vice versa and relying only on inter ROI eye movements for clustering.

5.7.8 Influence of habituation on subject experience

Another big challenge for captioning is the long term habituation to static captions and this is difficult to overcome even using the *manual placement*. Interestingly, we find two distinct behaviors amongst our user group as illustrated in Figure 67. The first group (Group A,

Mode	***Mean fraction value***	***Std deviation***
d-caption, initialize and track, no-blurb	***0.397***	***0.124***
d-caption, manual placement, no-blurb	***0.31***	***0.16***
s-caption, no-blurb	***0.12***	***0.04***

Table 15: Dynamic caption strategies draw significant amounts of user attention as can be seen from the fraction of eye movements spent on exploring dialogue boxes. This can be seen in the high fraction of gaze points taken up by dialogue boxes in column 2.

14 subjects) always prefers the static captioning. The second group (Group B, 5 subjects) on the other hand changes its preference as the dynamic captioning is changed from online initialization of dialogue positions (*d-caption manual*), to those indicated previously by other users using a mouse *d-caption manual* . in our subjects and the preferences of these two groups is illustrated in the chart shown in the following Figure 67.

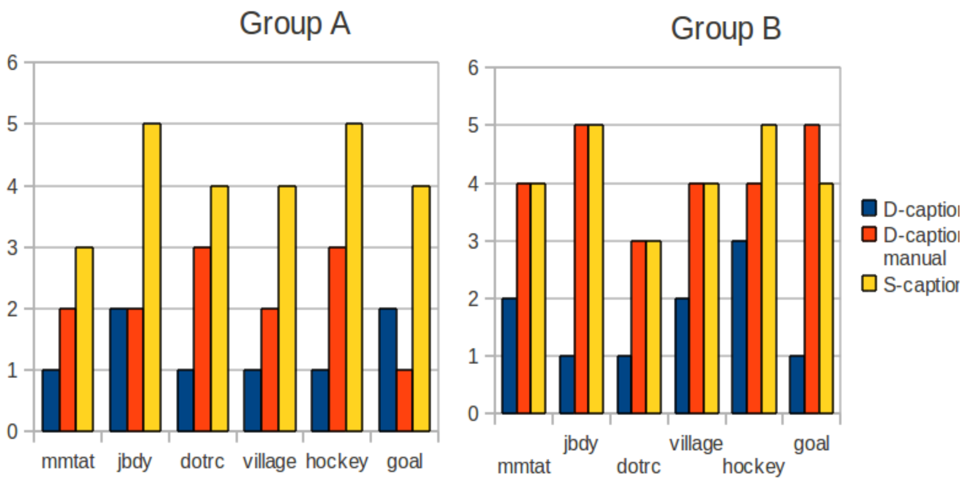


Figure 67: Group A part of the subject pool, changes its decision as the dialogues are restricted to the locations where they are initialized. On the other hand, subjects in the larger pool, Group B, do not their change their preference and consistently report better comprehension and viewing comfort with static captions. One reason for such response could be the familiarity and habituation to static captions through long exposure to current captioning.

6 Discussion and Future work

Important insights and results and possible directions of future work arising from this thesis, are discussed in this section.

6.1 Discussion of important results

Rapid and global image perception is shown to be possible for interestingness discrimination in an image [45]. This is an intriguing result because it means humans can very quickly infer whether an image would be interesting for viewing. It is also useful for systems that are prone to have a high recall for a semantic query, but low precision or agreement with the user's notion of interestingness. The role of intensity and color information at local and global scales has been quantified. Our experimental results also bound the minimum display rate of about 2-5 Hz Figure 22 at which such discrimination can be done by humans [45]. Modeling these properties as image features can be evaluate or even modify the interestingness or aesthetic quality of an image. Some effort towards this has been done in [104], but this work expects users to manually input ROIs and their relative priority in image. This is a promising area of work and richer image analysis combined with aesthetics model can give useful applications to photography and image content based retrieval.

The close relationship between eye-movements and abstract semantics in visual content has been shown in this thesis via the *attentional bias* model [70], also visualized in Fig. 32. Top down influence of affective content has also been quantified and represented in the attentional bias model. This can allow non-intrusive, interactive generation meaningful tags for visual content [70]. Eye-fixations have also been shown to be consistent across subjects and can be seen in the applications presented in [70] and [71], this data has also been distributed along with visualization routines in [71]. The consistency of attentional bias across important image themes has been demonstrated via a sequence of classification results on important categories such as faces, people and actions as shown in Tables 8,9 respectively, in Chapter 5.

For useful applications like dialogue localisation, it is important to find ROIs in visual content in a robust and time-efficient manner. A *binning* algorithm has been presented for this in Chapter 4. Our method identifies important visual components along with its location and rough estimate of size, it also gives an indication of the attentional bias in the scene. The novel algorithm is biologically motivated and implements the concept of inhibition-of-return in eye-movements [48]. This gives far superior clustering results than simple euclidean distance based clustering on eye-gaze data. This algorithm has also been extended to identify dominant and stable visual elements in an image, this is a multi scale analysis of eye-gaze information and does make use of any prior knowledge of scale of objects, depth of scene, etc. This has been shown to give promising results for classification of image into themes related to *interaction* and *aesthetic composition* in Fig. 56 in chapter 5.

Eye movements have been used to improve the performance of state-of-art foreground segmentation[71] and object detection[46] algorithms.

These applications also show how human perception and the allocation of attention, can influence groundtruth for such applications and in-turn change the interpretation of precision and recall. Online and interactive scenarios for dynamic caption localisation and low cost eye-tracking setups have been explored successfully in chapter 5 and also set the stage for a whole new host of scenarios where visual attention can be utilized.

It is important to note a few important limitations and assumptions underlying the work in this thesis. Flickr®is assumed to contain representative images of natural and man-made scenes encountered in daily life and this in turn enables the results on pre-attentive discrimination of interestingness to generalize natural and urban scenes. The attentional bias model assumes scenes that are representative of normal events in the world and the statistical results do not apply for unusual objects such as a very conspicuous inanimate object (eg; building) near a human. The attentional bias model, results on interaction modeling and localization of verbs such as *look*,etc, apply to the NUSEF dataset[71]. Freewiewing experiments assume that the user does not have fatigue and uncommon biases for the content shown.

6.2 Future work

Though the work on pre-attentive discrimination of interesting images establishes the minimum time threshold, it doesn't explain the influence of different manipulations on changes in observed human response times. There is scope for extension of the computational model using global and local image information to predict human decisions on interestingness. The NUSEF dataset is a valuable resource for exploring visual saliency in social and affective scenes, I want to develop a novel saliency model mimic human attentional bias and also address the more challenging problem of predicting the timing and sequence of eye-movements in such images. Our ongoing research has already delved into the influence of depth information on eye movements and this has the potential to shed more light on how we understand scenes in the real world. I am also exploring new applications for gaze based interactive scenarios eg; interactive advertising as part of my ongoing work. The role of eye-movements in key-object detection and foreground segmentation has been explored in separate problems in this thesis, an interesting direction is to combine these two stages in a single hybrid framework. Other tangential directions of research arising out of this thesis are to analyze Pupillary dilation, which is an accompanying behavioural signal that comes out of eye-tracking experiments. My ongoing work has showed its usefulness for affective video analysis [47] and this has potential for further research.

From a more basic scientific perspective, eye-movements have the potential to helps us understand mechanisms involved in visual object categorization (segmentation, recognition and context analysis) and attention. Eye-tracking offers a valuable non-invasive source of information to understand how local and global image information play a role in these visual tasks and I am exploring these as part of my current and future research directions.

References

[1] A. E. Abdel-Hakim and A. A. Farag. Csift: A sift descriptor with color invariant characteristics. In *CVPR '06: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1978–1983, Washington, DC, USA, 2006. IEEE Computer Society.

[2] D. Alexandra and S. Santini. Context based semantics for multimedia retrieval. *Proceedings of the SPIE, Multimedia Content Access: Algorithms and Systems III*, 7255, 2009.

[3] P. Baldi and L. Itti. Of bits and wows: A bayesian theory of surprise with applications to attention. *Neural Netw.*, 23(5):649–666, 2010.

[4] D. H. Ballard. Generalizing the hough transform to detect arbitrary shapes. In *Readings in computer vision: issues, problems, principles, and paradigms*, pages 714–725, San Francisco, CA, USA, 1987. Morgan Kaufmann Publishers Inc.

[5] Y. Boykov and V. Kolmogorov. An expeirmental comparison of min-cut/max-flow algorithms for energy minimisatin in vision. In *Transactions in Pattern Analysis and Machine Intelligence*, volume 26, page 359–374, 2004.

[6] N. Bruce and J. Tsotsos. Saliency based on information maximization. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 155–162, Cambridge, MA, 2006. MIT Press.

[7] D. S. Butterfield, C. Fake, C. J. Henderson-Begg, and S. Mourachov. Interestingness ranking of media objects, 2006.

[8] M. G. Calvo and P. J. Lang. Gaze patterns when looking at emotional pictures: Motivationally biased attention. *Motivation and Emotion*, 28(3):221–243, 2004.

[9] R. Carmi and L. Itti. Visual causes versus correlates of attentional selection in dynamic scenes. *Vision Research*, 46(26):4333–4345, Dec 2006.

[10] 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.

[11] M. Cerf, J. Harel, W. Einhäuser, and C. Koch. Predicting human gaze using low-level saliency combined with face detection. In John C. Platt, Daphne Koller, Yoram Singer, and Sam T. Roweis, editors, *Neural Information Processing Systems*. MIT Press, 2007.

[12] N. Craswell and M. Szummer. Random walks on the click graph. *SIGIR '08: ACM Special Interest Group On Information Retrieval*, page 239–246, 2008.

[13] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR '05: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 886–893, 2005.

[14] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Studying aesthetics in photographic images using a computational approach. In *The 9th European Conference on Computer Vision*, pages 288–301, 2006.

[15] J. Deng, W. Dong, R. Socher, J.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR '09: Proceedings of the 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2009.

[16] Y. Deng, B.S. Manjunath, and S. Hyundoo. Color image segmentation. *CVPR '99: Proceedings of the 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1999.

[17] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: A benchmark. In *CVPR*, June 2009.

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

[19] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge 2009 (voc2009) results, 2010.

[20] G. T. Fechner. *Elemente der Psychophysik (Elements of Psychophics)*. Breitkopf, 1860.

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

[22] P. F. Felzenszwalb, R. B. Girshick, D. A. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(9):1627–1645, 2010.

[23] Pedro F. Felzenszwalb, Ross B. Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 99(PrePrints), 2009.

[24] K. Fukunaga and L. Hostetler. The estimation of the gradient of a density function, with applications in pattern recognition. *Information Theory, IEEE Transactions on*, 21(1):32 – 40, jan 1975.

[25] J. Gall and V. Lempitsky. Class-specific hough forests for object detection. In *In Proceedings IEEE Conference Computer Vision and Pattern Recognition*, 2009.

[26] B. Gong and R. Jain. Segmenting photo streams in events based on optical metadata. In *ICSC 2007. International Conference on Semantic Computing*, pages 71–78, Irvine, CA, USA, 2007. IEEE.

[27] G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. Technical Report 7694, California Institute of Technology, 2007.

[28] L. Haojie, T. Jinhui, L. Guangda, and T. S. Chua. Word2image: towards visual interpreting of words. In *ACM Multimedia*, pages

813–816, 2008.

[29] C. Harris and M. Stephens. A combined corner and edge detection. In *Proceedings of The Fourth Alvey Vision Conference*, pages 147–151, 1988.

[30] D. Holman, R. Vertegaal, C. Sohn, and D. Cheng. Attentive display: paintings as attentive user interfaces. In *CHI '04: CHI '04 extended abstracts on Human factors in computing systems*, pages 1127–1130, 2004.

[31] Richang Hong, Meng Wang, Mengdi Xu, Shuicheng Yan, and Tat-Seng Chua. Dynamic captioning: video accessibility enhancement for hearing impairment. In *ACM Multimedia*, pages 421–430, 2010.

[32] A. R. Hunt and A. Kingstone. Covert and overt voluntary attention: linked or independent ? 18:102–105.

[33] Amelia R. Hunt and P. Cavangah. Looking ahead: The perceived direction of gaze shifts before the eyes move. *Journal of vision*, 9(9), 2009.

[34] Phillip Isola, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. What makes an image memorable? In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 145–152, 2011.

[35] L. Itti. Quantitative modeling of perceptual salience at human eye position. *Visual Cognition*, 14(4-8):959–984, Aug-Dec 2006.

[36] L. Itti and P. Baldi. Bayesian surprise attracts human attention. *Vision research*, 49(10):1295–306, 2009.

[37] L. Itti and C. Koch. Computational modelling of visual attention. *Nature reviews. Neuroscience*, 2(3):194–203, March 2001.

[38] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(11):1254–1259, 1998.

[39] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(11):1254–1259, 1998.

[40] W. Y. K. Wai Y. H. Lai N. Davis J. K. Tsostos, S. M. Culhane and F. Nuflo. Modeling visual attention via selective tuning. *Artificial Intelligence*, 78:505–545, 1995.

[41] R. J. K. Jacob. The use of eye movements in human-computer interaction techniques: what you look at is what you get. *ACM Trans. Inf. Syst.*, 9(2):152–169, 1991.

[42] T. Judd, F. Durand, and A. Torralba. Fixations on low-resolution images. *J Vis*, 11(4), 2011.

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

[44] G. L. Malcolm K. Rayner, T. J. Smith and J. M. Henderson. Eye movements and visual encoding during scene perception. *Psychological Science*, 20(1):6–10, 2009.

[45] H. Katti, Y. B. Kwok, T. S. Chua, and M. S. Kankanhalli. Pre-attentive discrimination of interestingness in images. In *ICME*, pages 1433–1436, 2008.

[46] H. Katti, S. Ramanathan, M. S. Kankanhalli, N. Sebe, T. S. Chua, and K. R. Ramakrishnan. Making computers look the way we look: exploiting visual attention for image understanding. In *Proceedings of the international conference on Multimedia*, MM '10, pages 667–670, New York, NY, USA, 2010. ACM.

[47] Harish Katti, Karthik Yadati, Mohan Kankanhalli, and Tat-Seng Chua. Affective video summarization and story board generation using pupillary dilation and eye gaze. In *Proceedings of the International Symposium on Multimedia*, ISM '11. IEEE, 2011.

[48] R. M. Klein. Inhibition of return. *Trends in Cognitive Sciences*, 4(4):138 – 147, 2000.

[49] S. B. Kotsiantis. Supervised machine learning: A review of classification techniques. *Informatica (Slovenia)*, 31(3):249–268, 2007.

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

[51] C. Li, A. C. Loui, and T. Chen. Towards aesthetics: a photo quality assessment and photo selection system. In *Proceedings of the international conference on Multimedia*, MM '10, pages 827–830, New York, NY, USA, 2010. ACM.

[52] F.F. Li, R. VanRullen, and C. Kochand P. Perona. Rapid natural scene categorization in the near absence of attention. *Proceedings of the National Academy of Sciences*, 99(14):9596, 2002.

[53] M. Lipps and J. B. Pelz. Yarbus revisited: task-dependent oculomotor behavior. *4th Annual Meeting of the Vision Sciences Society*, 2004.

[54] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[55] Yijuan Lu, Lei Zhang, Qi Tian, and Wei-Ying Ma. What are the high-level concepts with small semantic gaps? In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8, 2008.

[56] Michael Lyons, Shigeru Akamatsu, Miyuki Kamachi, and Jiro Gyoba. Coding facial expressions with gabor wavelets. pages 200–205, 1998.

[57] Bar M. Visual objects in context. In *Nature Reviews in Neuroscience*, volume 5, pages 617–29, 2004.

[58] M. Yumiko M. Masaya, T. Hiroyuki and K. Ryoji. Access concentration detection in click logs to improve mobile web-ir. *Information Sciences*, 179:1859–1869, 2009.

[59] G. A. Miller. Wordnet: a lexical database for english. *Commun. ACM*, 38(11):39–41, 1995.

[60] A. Mishra, Y. Aloimonos, and C. Loong Fah. Active segmentation with fixation. In *International Conference on Computer Vision*, 2009.

[61] A. G. Money and H. Agius. Video summarisation: A conceptual framework and survey of the state of the art. *J. Vis. Comun. Image Represent.*, 19(2):121–143, 2008.

[62] H. Müller, T. Pun, and D. Squire. Learning from user behavior in image retrieval: Application of market basket analysis. *Int. J. Comput. Vision*, 56(1-2):65–77, 2004.

[63] V. Navalpakkam and L. Itti. An integrated model of top-down and bottom-up attention for optimal object detection. In *CVPR '06: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2049–2056, 2006.

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

[65] A. Oliva and A. Torralba. The role of context in object recognition. *Review TRENDS in Cognitive Sciences*, 11(12):145–175, 2007.

[66] P. Over, G. Awad, R. T. Rose, J. G. Fiscus, W. Kraaij, and A. F. Smeaton, editors. *TRECVID 2008 workshop participants notebook papers, Gaithersburg, MD, USA, November 2008*. National Institute of Standards and Technology (NIST), 2008.

[67] C. Peters, M. Agosti, and M. de Rijke. Image retrieval in clef-imageclef, September 2010.

[68] M. I. Posner. Orienting of attention. *The Quarterly Journal of Experimental Psychology*, 32(1):3–25, 1980.

[69] L. Jia J. Z. Wang R. Datta, D. Joshi. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Survey*, 40(2), 2008.

[70] S. Ramanathan*, H. Katti*, R. Huang, T. S. Chua, and M. S. Kankanhalli. Automated localization of affective objects and actions in images via caption text-cum-eye gaze analysis. In *ACM Multimedia 2009*, pages 729–732, 2009. (* indicates equal contribution).

[71] S. Ramanathan*, H. Katti*, M. S. Kankanhalli, T. S. Chua, and N. Sebe. An eye fixation database for saliency detection in images. *ECCV 2010*, 2010. (* indicates equal contribution).

[72] L. W. Renninger, J. M. Coughlan, P. Verghese, and J. Malik. An information maximization model of eye movements. pages 1121–1128, 2005.

[73] 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.

[74] 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.

[75] L. Zelnik-Manor S. Goferman and A. Tal. Context-aware saliency detection. *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

[76] A. A. Salah, T. Gevers, and N. Sebe. *Communication and Automatic Interpretation of Affect from Facial Expressions*, chapter X, page X. IGI Global, 2010.

[77] Anthony Santella and Doug DeCarlo. Robust clustering of eye movement recordings for quantification of visual interest. In *Proceedings of the 2004 symposium on Eye tracking research & applications*, ETRA '04, pages 27–34, New York, NY, USA, 2004. ACM.

[78] V. Setlur, T. Lechner, M. Nienhaus, and B. Gooch. Retargeting images and video for preserving information saliency. *IEEE Computer Graphics and Applications*, 27:80–88, 2007.

[79] A. Shamir and S. Avidan. Seam carving for media retargeting. *Commun. ACM*, 52(1):77–85, 2009.

[80] A. Shamir and O. Sorkine. Visual media retargeting. In *SIGGRAPH ASIA '09: ACM SIGGRAPH ASIA 2009 Courses*, pages 1–13, New York, NY, USA, 2009. ACM.

[81] P. Sinha and R. Jain. Classification and annotation of digital photos using optical context data. In *Proceedings of the 2008 international conference on Content-based image and video retrieval*, CIVR '08, pages 309–318, New York, NY, USA, 2008. ACM.

[82] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and trecvid. In *MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, pages 321–330, New York, NY, USA, 2006. ACM Press.

[83] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(12):1349–1380, 2000.

[84] C. Snoek. Semantic image and video indexing in broad domains. In *IEEE Transactions on Multimedia*, volume 9, New York, NY, USA, 2007. IEEE.

[85] C. G. M. Snoek, M. Worring, O. de Rooij, K. E. A van de Sande, R. Yan, and A. G. Hauptmann. Videolympics: Real-time evaluation of multimedia retrieval systems. *IEEE MultiMedia*, 15:86–91, January 2008.

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

[87] Y. Sugano and Y. Matsushitaand Y. Sato. Calibration-free gaze sensing using saliency maps. *CVPR '2010: Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2667–2674, 2010.

[88] A. Torralba and A. Oliva. Statistics of natural image categories. *Network: Computation in Neural Systems*, pages 391–412, Au-

gust 2003.

[89] 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, October 2006.

[90] A. B. Torralba, R. Fergus, and W. T. Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(11):1958–1970, 2008.

[91] A. B. Torralba, K. P. Murphy, and W. T. Freeman. Using the forest to see the trees: exploiting context for visual object detection and localization. *Commun. ACM*, 53(3):107–114, 2010.

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

[93] M. Varma and D. Ray. Learning the discriminative power-invariance trade-off. In *Proceedings of the IEEE International Conference on Computer Vision, Rio de Janeiro, Brazil*, October 2007.

[94] A. Vedaldi, V. Gulshan, M. Varma, and A. Zisserman. Multiple kernels for object detection. In *Proceedings of the International Conference on Computer Vision*, September 2009.

[95] R. Vertegaal. A fitts law comparison of eye tracking and manual input in the selection of visual targets. In *IMCI '08: Proceedings*

of the 10th international conference on Multimodal interfaces, pages 241–248, 2008.

[96] R. Vertegaal, A. Mamuji, C. Sohn, and D. Cheng. Media eyepliances: using eye tracking for remote control focus selection of appliances. In *CHI '05: CHI '05 extended abstracts on Human factors in computing systems*, pages 1861–1864, 2005.

[97] R. Vertegaal and S. J. Shell. Attentive user interfaces: the surveillance and sousveillance of gaze-aware objects. *Social Science Information*, 47(3):275–298, September 2008.

[98] O. Villon and C. L. Lisetti. Toward building adaptive user's psycho-physiological maps of emotions using bio-sensors. In *1rst Workshop on Emotion and Computing at KI 2006, 29th Annual Conference on Artificial Intelligence, June, 14-19, 2006, Bremen, Germany*, 06 2006.

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

[100] J. Vogel, A. Schwaninger, and C. Wallraven. Categorization of natural scenes: Local versus global information and the role of color. In *ACM Transactions in Applied Perception*, 2007.

[101] J. Vogel, A. Schwaninger, C. Wallraven, and H. H. Bülthoff. Categorization of natural scenes: Local versus global information and the role of color. *TAP*, 4(3), 2007.

[102] U. Vural and Y. S. Akgul. Eye-gaze based real-time surveillance video synopsis. *Pattern Recogn. Lett.*, 30(12):1151–1159, 2009.

[103] J. Wang, E. Pohlmeyer, B. Hanna, Y. G. Jiang, P. Sajda, and S. F. Chang. Brain state decoding for rapid image retrieval. In *Proceeding of the ACM international conference on Multimedia (ACM MM)*, October 2009.

[104] L. K. Wong and L. K. Lim. Saliency retargeting: An approach to enhance image aesthetics. IEEE, 2011.

[105] Di Xu and P. Nasiopoulos. Logo insertion transcoding for h.264/avc compressed video. In *16th IEEE International Conference on Image Processing (ICIP)*, pages 3693 – 3696. IEEE, 2009.

[106] M. Ortega Y. Rui, T. S. Huang and S. Mehrotra. Relevance feedback: A power tool in interactive content-based image retrieval. *IEEE Trans. on Circuits and Systems for Video Technology , Special Issue on Segmentation, Description, and Retrieval of Video Content*, 8(5):644–655, 1998.

[107] Q. Zhao and C. Koch. Learning a saliency map using fixated locations in natural scenes. *J Vis*, 11(3), 2011.

I want to express my gratitude to the writers and poets of today and yesteryears for allowing me the liberty to explore their thoughts and emotions. To the wonderful solitude of Wang Wei, the turmoil of Du Fu, the unbashed joy of Li Bai and to Vikram Seth who introduced me to them.

South and north of my house lies springtime water,
And only flocks of gulls come every day,
The flower path's unswept: no guests.
The gate is open: you are the first one to come this way,
The market's far: my food is nothing special,
The wine, because we are poor, is an old brew,
Across the fence to drink it with us two.
*- Du Fu, The visitor, 300 Tang poems (quán táng shī)**

I will not ever see my friend again.
Day after day Han waters eastward flow.
Even if I asked of the old man, the hills
and rivers would seem empty in Caizhou
*- Wang Wei, Grieving for Meng Haoran, 300 Tang poems (quán táng shī)**

A pot of wine among the flowers,
I drink alone, no friend with me.
I raise my cup to invite the moon.
He an my shadow and I make three.

The moon does not know how to drink;
My shadow mimes my capering;
But I'll make merry with them both-
And soon enough it will be Spring.

I sing – the moon moves to and fro,
I dance – my shadow leaps and sways.
Still sober, we exchange our joys.
Drunk – and we'll go our seperate ways.

Let's pledge – beyond human ties – to be friends,
And meet where the Silver River ends.
*- Li Bai, Drinking alone with the Moon, 300 Tang poems (quán táng shī)**

** Reproduced from Vikram Seth's translation of these classic Tang poems* *in his book “Three Chinese Poets”, Faber and Faber, London, 1992*

to Kobayashi Issa, for seeing profoundness in the commonplace, and seeing beauty in the immense pain thrown on him by life,

Do not kill
the housefly,
it rubs its hand it rubs it feet in prayer.
- Kobayashi Issa, About the housefly.

Outliving them,
outliving them all,
ah,
the cold!
– Kobayashi Issa, Grieving for the loss of his wife and children

to Robert Frost for giving meaning to devotion,

The heart can think of no devotion,
Greater than being shore to ocean.
Holding the curve of one position,
Counting an endless repetition.
- Robert Frost, On devotion.

to Rumi, who understood love far greater than I do,

Your task is not to seek for love,
but merely to seek and find,
all the barriers within yourself,
that you have built against it.
– Maulana Jalaluddin Rumi, On finding love

and finally to writers whose work and insights are far greater than what I can quote or describe, I just feel lucky to have crossed paths with your work through these years.