Chapter 14
Mining Videos for Features that Drive Attention

Farhan Baluch and Laurent Itti

Abstract Certain features of a video capture human attention and this can be measured by recording eye movements of a viewer. Using this technique combined with extraction of various types of features from video frames, one can begin to understand what features of a video may drive attention. In this chapter we define and assess different types of feature channels that can be computed from video frames, and compare the output of these channels to human eye movements. This provides us with a measure of how well a particular feature of a video can drive attention. We then examine several types of channel combinations and learn a set of weightings of features that can best explain human eye movements. A linear combination of features with high weighting on motion and color channels was most predictive of eye movements on a public dataset.

14.1 Background

Videos are made up of a stream of running frames each of which has a unique set of spatial and textural features that evolve over time. Each video therefore presents a viewer with a large amount of information to process. The human visual system has limited capacity and evolution has incorporated several mechanisms into the visual processing systems of animals and humans to allow only the most important and behaviorally relevant information to be passed on for further processing. The first stage is the limited amount of high resolution area in the eye, i.e., the fovea. When we want to focus on a different spatial region of a video we make an eye movement, also

F. Baluch (✉)
Research and Development Group, Opera Solutions,
12230 El Camino Real, San Diego, CA 92130, USA
e-mail: farhanbaluch@gmail.com

L. Itti
Department of Computer Science, Psychology & Neuroscience Graduate Program,
University of Southern California,
3641 Watt Way, HNB 10, Los Angeles, CA 90089, USA
e-mail: itti@usc.edu

A.K. Baughman et al. (eds.), *Multimedia Data Mining and Analytics*,
DOI 10.1007/978-3-319-14998-1_14

known as a saccade, to bring the area of interest into alignment with the fovea. Within the fovea too, attention can focus our perception on features that either interest us or that are intrinsically visually conspicuous or *salient*. The former is termed top-down attention and the latter bottom-up attention [2].

In this chapter we discuss in detail how a video may be decomposed into a set of features that coarsely map to features computed in the human brain. Using this neuromorphic approach to predicting eye movements, we can begin to understand what features of a video attract human attention. This understanding is not only essential for answering the scientific question of how attention works in the brain, but, in addition, this understanding can also help us build better computer vision systems and furthermore has other applications. A model that can successfully predict where humans allocate attention can be used to enhance marketing displays [5], provide a means to intelligently compress videos [21], speed up object recognition [32], and also improve video based security systems [37].

14.1.1 Human Attention and Eye Movements

The study of eye movements as a measure of human attention dates back to over 100 years ago; however, it was Yarbus [35] who first reported the manner in which eye movements may reveal the goal of an observer and those items in the scene that are determined to be interesting. Often objects that are of interest functionally (or cognitively) also inherently possess the visual attributes that attract attention, i.e., these objects are considered salient both visually and psychologically [7]. Therefore, studying the eye movements of human subjects while they view static or dynamic scenes can reveal a lot about the cognitive processes underlying human visual perception. Despite the proliferation of tools now available to study the brain, eye movements provide a simple, quick and non-invasive method to probing human attention using experimental means. Eye movements are monitored using infra-red or high definition cameras that can detect and continually track the pupil of the eye. By calibrating the change in position of the pupil with certain calibration points on a screen, a mapping or transformation can be used to translate the movement detected by the eye tracker to screen coordinates. Using this method, an observer's eye movement traces can be overlaid on the image or video being presented, thereby providing a means of locating the observer's attentional allocation to the scene.

Features of the visual world attract human attention and gaze. There is necessarily a relationship between the visual world and a particular human or animal's behavioral goals that results in eye movements and shifting of attention. The study of the origin of an eye movement has been the subject of numerous studies and is a very actively debated and active area of research [2, 22]. Broadly, however, the cause of an eye movement or attention shift is categorized as *bottom-up* if it is a result of the visual appeal of the sensory input (e.g., orientation of attention towards a light that suddenly

blinks brightly), and *top-down* if it is a result of the behavioral goal of the human or animal in question (e.g., a volitional shift of gaze to the left when nothing in the visual environment has changed). While this distinction helps us model and understand attention, the separation of the two purported sources (i.e., top-down and bottom-up) is a very challenging question in neuroscience. Since the onset of visual experience, a human or animal begins to form a subjective percept which, depending on experience, may force certain stimuli to appear a certain way that may be different from another individual's percept. Subjective experience and perception therefore can challenge the existence of a "normative" visual percept, and, therefore, make it very difficult to separate bottom-up and top-down influences on attention [2, 6, 9].

When modeling the human processes of attention, eye movements serve as the empirical evidence used to validate and quantify the quality of model. Any model of visual attention serves to indicate with faithfulness the likelihood of a human observer allocating attention to certain salient parts of the scene, i.e., a model generates a saliency map. Similar to the manner in which eye movements can be overlaid on an image, these eye movement traces can also be overlaid on a saliency map generated by a model. In this manner, we can find models that have an output that closely corresponds with human eye movements. Furthermore, we can use the deviation between the model output and the human eye movements to construct a cost function that can be optimized to fit parameters of new models developed.

14.1.2 Models of Human Attention

The development of saliency models lies at the interface of understanding human perception and developing visual intelligence in machines. In several domains, engineers have built systems that mimic or are inspired by biological systems. Biologically-inspired computer vision has a similar goal. In particular, modeling of attention has been an area of interest with numerous publications dedicated to the topic over the years [4]. Modeling attention is equivalent to computing the most conspicuous or salient regions of an image, that are likely to drive attention and, as a result, elicit an orientation of gaze towards the location in the image. Two approaches can be taken to building a model that can best explain a human viewers' eye movements. In the first approach, the functioning of the human visual system can be studied and modeled to arrive at a computational model of human attention, several models take this approach [14]. The second approach is to examine the patches of an image that are fixated by human observers and understand their features to build a dictionary of patches that are likely to elicit eye movements [17, 26]. In this chapter, we describe in detail a model that follows the first approach and attempts to arrive at a model based on the functioning and anatomy of the visual systems in biological organisms.

The Itti and Koch [16] model of salience has been widely successful [14] in accounting for overt allocation of attention in natural images and has become a benchmark for comparing other models. The model builds on previous theories [18,

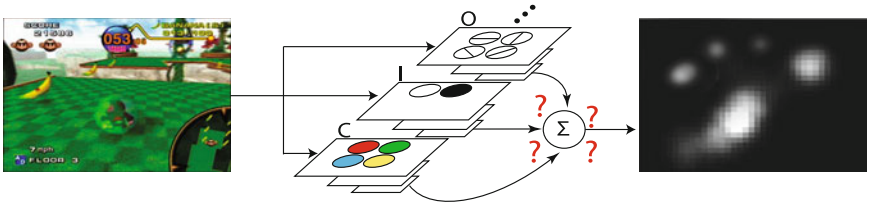

Fig. 14.1 Computation of features and saliency maps from video frames. Multi-scale features are computed from each video frame e.g. color, intensity, orientation etc. The feature maps are then combined generally using a simple linear combination to form a conspicuity maps also known as a saliency map. This example shows a video frame from a recording of a video game and the resulting saliency map for this frame after linear combination of features

27] of how attention is allocated by computing individual feature maps and then combining these maps into a single conspicuity map, where salient locations in the map attract attention. The general framework of this model consists of a decomposition of an image into various feature maps, by computing filter based features. Each computed feature forms a channel in the model that can be added or removed from the final saliency computation. Examples of these features include intensity contrast, motion energy, color opponent contrast, etc. Numerous such feature channels can be computed, and, since the development of the original model, a large number of channels have been added based on neuroscience discoveries of mechanisms of vision in the brain as well as useful features based on computer vision. Figure 14.1 illustrates the manner in which an image is decomposed into a set of features computed at multiple scales and then finally combined to form a saliency map. The saliency map can be viewed as an attention probability map that assigns high probability to regions of the image that are inherently interesting or likely to elicit human attention. The figure shows the color (C), intensity (I) and orientation (O) channels [16]. In a similar manner, several other channels can be computed and these have been listed in Table 14.1.

Each channel from this large set may contribute toward the salience of a location in the image and, therefore, the potential to elicit a gaze shift from a human. Each channel outputs a feature map that consists of pixels corresponding to the image. Each pixel in the feature map indicates the energy that the feature in question contributes at that location. In the standard implementation, feature maps output from all channels are linearly summed to form a final saliency map. This saliency map, after some normalization, serves as a probability map that consists of the same number of pixels as the input image and the value at each pixel indicates the likelihood of that pixel eliciting an attention orientation towards it by a human viewer. There are several strategies to combining the features maps into a final saliency map [15] and this continues to be an active area of research. In this chapter we will also focus on methods to combine feature maps and build a saliency map that maximizes the probability of predicting human gaze.

Table 14.1 List of feature channels

Channel name	Abbrev	Refs	Description
Color	C	[16]	Double-opponent color center-surround, for red-green and blue-yellow contrasts
Flicker	F	[12]	Flicker center-surround channel based on frame by frame differences
Multi-color band	G	[24]	Multi-color band channel with N Gaussian bands spanning the hue axis
H2SV	H	[23]	A variant over the HSV color space
Intensity	I	[16]	Intensity center-surround channel
DKL Color	J	[36]	A biologically-inspired color model
Skin hue	K	[31]	Skin hue detector
L-junction	L	[20]	Channel tuned to L-shaped corner edges
Motion	M	[12, 31]	Motion channel based on frame by frame differences
Intensity-band	N	[24]	Intensity channel with N Gaussian bands spanning the intensity axis
Orientation	O	[16]	Gabor-based orientation channel with N orientations
CIELab Color	Q	[10]	Color channel using the CIE L*a*b* color model
Pedestrian	R	[31]	Pedestrian channel based on simple template matching for humans
Single-opp. color	S	[24]	Composite of single-opponent color center-surround computed separately in the red, green, blue and yellow color bands
T-junction	T	[20]	Detector tuned to T-shaped edge junctions
Foreground	U	[13]	Foreground/background detection channel
Contour	W	[23, 25]	Elongated contour detection channel
X-junction	X	[20]	Detector tuned to X-shaped crossings of edges

In addition to the specific references listed below, papers [12, 16, 20, 25, 28] provide summary descriptions of collections of channels, and [13] provides reference source code implementation

14.2 Experimental Study of Attention

To evaluate a model of attention we need to obtain evidence of correspondence between the output of the model, i.e., its prediction of attention allocation within a scene, and human attention allocation. As discussed above, one means of measuring human attention allocation is by examining human eye movements using an eye tracker. Typically in experiments a specific set of stimuli is chosen and displayed on the screen. Study participants are given instructions on how to observe the scenes. Instructions can make a large difference on eye movements, in particular different types of instructions can emphasize either bottom-up or top-down aspects of the scene. For example, asking subjects to look for a yellow road sign in scenes may influence their eye movements spatially towards expected locations of road signs

(spatial bias) and also may influence them to fixate on items that are yellow (feature bias). On the other hand, providing minimal instructions and asking subjects to watch and enjoy the scenes may emphasize bottom-up aspects of attention allocation by recording eye movements based on scene changes. While efforts can be made to emphasize bottom-up aspects of a scene there is no way to completely eliminate the influence of top-down aspects such as the viewers' personal bias and preferences.

14.2.1 Methods

We will discuss a study where three females and five males aged 23–32 with normal or corrected-to-normal vision were recruited. This data set, including both the videos as well as the recorded eye movement traces, are available openly to the public through the CRCNS program [11] for exploration. All subjects were USC students or staff members. Subjects gave written consent under a protocol approved by the Institutional Review Board and were paid for participating in the study. The stimuli for this study consisted of 50 video clips between 6 and 90 s each shown at 30 fps. A total of 46,000 video frames and 25 min of total video time. The videos contain a mix of indoor and outdoor scenes including park scenes, crowds, rooftop bars, TV news, sports, commercials, and video game footage. Figure 14.2 shows an example of these stimuli. The stimuli were presented on a monitor at 640×480 resolution running at 60 Hz. An ISCAN RK-464 eye tracker was used to track the subjects' eyes at 240 Hz. A nine point calibration was performed every five clips.

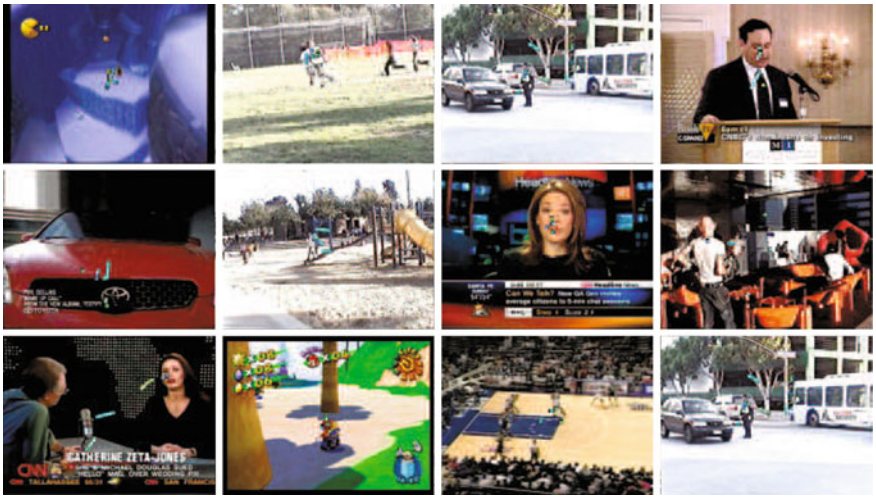

Fig. 14.2 Sample frames from video stimuli consisting of videos of different scenes including video game, TV adverts, outdoor and indoor scenes

Subjects were seated in a comfortable chair and asked to view the clips while their eyes were tracked. The instructions to the subjects were: "Follow the main actors and actions, try to understand overall what happens in each clip. We will ask you questions about the main contents. Do not worry about details". This instruction aimed to emphasize the bottom up component of the visual input being presented to the subjects. If they were asked to look for anything specific this would introduce a heavy top-down component and subjects' eye movements would reflect their own search strategies more than the inherent ability of the stimulus to draw attention. The goal of our modeling effort is to model bottom up or purely sensory components of the environment that can explain attentional shifts and allocation. Therefore subjects are instructed to focus more on the general scene rather than any specific targets.

As described earlier, the eye movements can be overlaid on the images being displayed in a post-processing step and in this manner we can observe the viewer's location of gaze on the scene and thus infer attentional allocation. The eye traces recorded during the viewing of the stimuli by the subjects were parsed into saccades based on a threshold of velocity as described before [1]. A total of 11,430 saccades were extracted and analyzed. Using the saliency model, we were able to extract feature maps and saliency maps using different combination rules. We then sampled these feature and saliency maps at the saccade endpoints to look for correlations between gaze location and saliency/feature values.

14.2.2 The Inter-observer Model and AUC Metric

To set an upper bound for the performance of models we built an inter-observer model. To build this model we grouped together the eye movements of all the subjects at each video frame and added them into a map consisting of all zeros and ones at locations of eye movement end points as shown in Fig. 14.3. A Gaussian centered at the location of each saccade endpoint or eye movement was defined with radius 5 pixels and applied to the map. This generated smooth "salience" maps defining the output of an inter-observer model. Since we know that each human observer will be different and we do not expect all to be the same we build this map as an average location of where we expect humans to fixate in a scene. The expectation is that a group of humans should predict the eye movements of a new observer who was not in the set of observers used to build the inter-observer model. To assess the quality of the model, however, we need a metric.

To quantify the performance of a particular model in predicting gaze, we use an ROC (receiver operating characteristics) like measure called area under the curve (AUC). This measure is computed by plotting the values generated in the models map at saccade end points against the values at 100 random locations on the map [3, 29]. Once these samples are drawn we can slide a threshold of salience and ask what percentage of human versus random locations were selected by the model at this threshold. A good model would result in a larger number of human fixated locations containing high values of and few random locations. The plot serves as an ROC curve

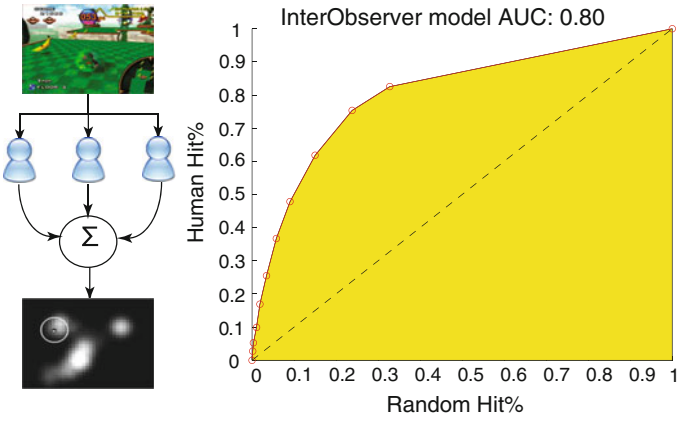

Fig. 14.3 Inter-observer model. *Left* shows a schematic of how the inter-observer model saliency maps were generated by pooling together eye movements from all subjects and then applying a gaussian at the saccade end points. *Right* shows the ROC curve by predicting saccades based on the inter-observer model. This *curve* was computed by computing inter-observer maps from seven subjects and then predicting the saccades of the one left out subject

and the area under this curve gives a measure of the quality of the model in question. A value of 1 indicates a model that completely accounts for saccade allocation while a value of 0.5 indicates a model that is no better than chance at predicting the location of gaze. All models we discuss in later sections will be gauged using this metric. Note that the theoretical maximum AUC of 1 is not achievable with a generic model that is not tailored to each particular individual, because all humans do not always agree, hence a single model cannot perfectly capture attention allocation of every single human.

Calculating the AUC metric for the inter-observer model, we obtain a very high AUC score of 0.80, indicating high (though not perfect) inter-observer agreement. This AUC score was significantly higher than all individual channels computed as well as various trained and untrained models as we examine in later sections. The inter-observer model, therefore sets the upper bound on the performance of the models. Intuitively, we do not expect a computational model of attention to be any better (or as good) at predicting human attention than a model constructed from the eye movements of a group of human observers.

14.3 Analysis of Feature Contributions

To understand the manner in which features interact to guide attention we decomposed each video frame into a set of feature maps that when combined would provide a saliency map [14] as discussed in earlier sections. Each so called channel provided a single feature map for each video frame. The channels computed were color, intensity, orientation, flicker, motion, and several others including complex junction channels

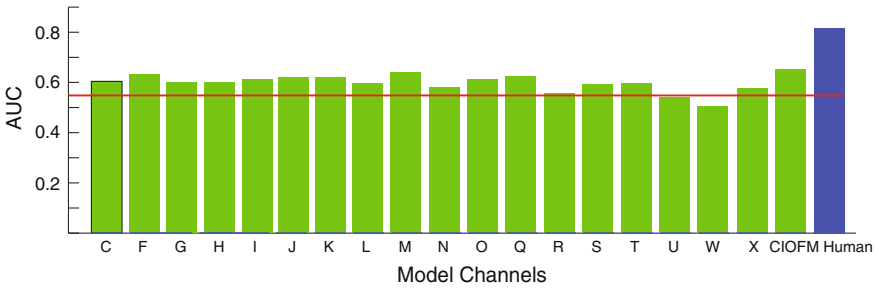

Fig. 14.4 Individual channel AUC. Each bar represents the performance of models built form individual channels in predicting gaze. See Table 14.1 for channel descriptions. CIOFM represents a linear combination of the C, I, O, F and M channels without any weighting. The bar labeled *Human* represents the inter-observer model. The *red line* indicates chance level i.e. AUC = 0.5

as listed in Table 14.1. We first analyzed the performance of each of these channels individually at predicting gaze. This is done by computing a feature map or channel on each video frame and then applying the AUC metric to test for performance. The lower bound for AUC is 0.5 i.e. a model is at chance at predicting whether a location will receive human attention or not, while the upper bound is set by the inter-observer model.

Figure 14.4 plots the AUC scores for individual channels as well as the inter-observer model. It is clear that individual channels fall short of the AUC score obtained from the inter-observer model. As expected humans are able to predict the attention allocation of other humans better than models of attention. A simple linear combination of the color (C), intensity (I), orientation (O), flicker (F) and motion (M) channels results in a model that performs reasonably well at predicting human attention allocation in scenes. In the rest of our discussion we focus on methods of finding highly predictive combinations of features using linear and non-linear combinations. We focus on the C, I, O, F, M channels because those have been historically prevalent, and we examine how different combinations of these would provide differences in gaze prediction.

When comparing features, we found that the motion channel was most predictive among the five analyzed channels (C, I, O, F and M). Figure 14.5 plots a histogram of the number of locations versus saliency value assigned by the model. One set of bars indicate the number of random locations that were assigned a certain saliency value while another set indicates the number of locations that were targets of human attention/saccades assigned that same level of saliency. The inset ROC curve in Fig. 14.5 plots is used to compute the AUC score which for the motion channel is 0.64, a reasonably good score.

Since the videos contain a significant amount of motion and the instruction to the subjects was to follow the main actors and actions it is intuitive that the motion channel is predictive of locations where subjects make saccades. While the inter-observer map sets the absolute upper bound on models, the motion channel with its

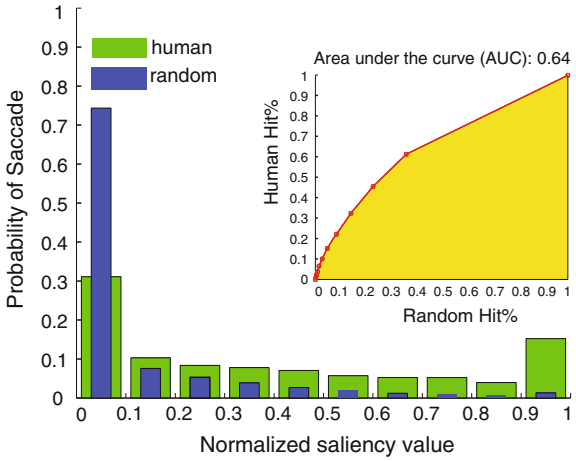

Fig. 14.5 Performance of the motion channel. Histogram shows probability of saccade (*green*) or random location hit (*blue*) towards locations of different saliency values from 0 to 1. The histogram is for saliency of the motion channel. Inset is an ROC curve computed by sliding a threshold along saliency values. The *yellow area* shows the area over which the AUC score is computed

high AUC score also sets a benchmark for other channels and models that combine channels.

14.4 Results

14.4.1 Linear Model with Trained Weights

To test the prediction that is driven by a weighted linear combination of features we trained a linear model to predict saccades by optimizing an objective function defined by the AUC cost. We used a genetic algorithm to find the optimal combination of weights for five features C, I, O, F and M. A genetic algorithm approach was used to enable comparison of this model with the larger optimization of a model with 20 features as discussed below. We started with random weights and enforced a constraint of allowing a weight to vary between 0 and 1. A population of 100 candidates was used and each individual consisted of five values of weights for each of the features. At each iteration each individual provided five weights for features which were used to build a saliency model and output a final saliency map. This map was used to compute the AUC score and determine how well the model did at predicting human gaze. Each individual's AUC score was computed and this was used as the fitness value for this candidate. Standard mutation and cross-over operators were used to breed new individuals. The results of the optimization with five features are shown in Fig. 14.6.

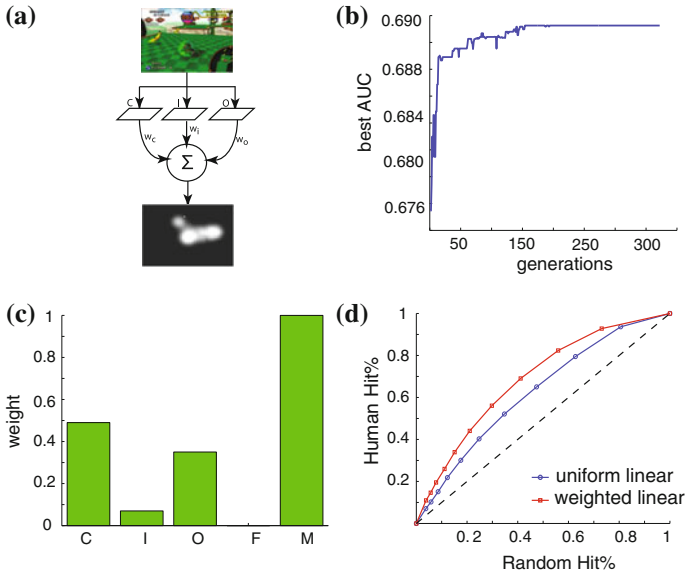

Fig. 14.6 Linear combination with learned weights. **a** Multi-scale features are computed from each video frame e.g. color, intensity, orientation etc. The feature maps are then combined generally using a simple linear combination to form a conspicuity maps also known as a saliency map. **b** The evolution of the best AUC of the population of individuals used in the genetic algorithm optimization. **c** The final weights learned by the genetic algorithm. It can be seen that motion has the highest weight at the end of learning. **d** A comparison of ROC curves between a linear model with uniform weights versus a linear model with learned weights

The genetic algorithm converged on a solution in about 200 generations. The fittest individuals had an average AUC value of 0.69. As predicted, the weight for the motion channel was the highest. The color and orientation channels also show a significant contribution in predicting salience. The weight for the flicker channel was very low, probably because this channel is highly correlated with the motion channel [36], and hence the optimization algorithm discarded it as redundant. Our results are in line with studies that have found that color and motion are among the top features that attract gaze [1, 34]. The weighted linear model with first order features performed significantly better than the uniform linear model that combined the features with uniform weights. As discussed below this turns out to be the most predictive model among the ones analyzed in this study.

14.4.2 Second Order Feature Interaction Model

To study the effect of non-linear interactions we generated second order features by point-to-point multiplication of each of the five features studied (CIOFM). This

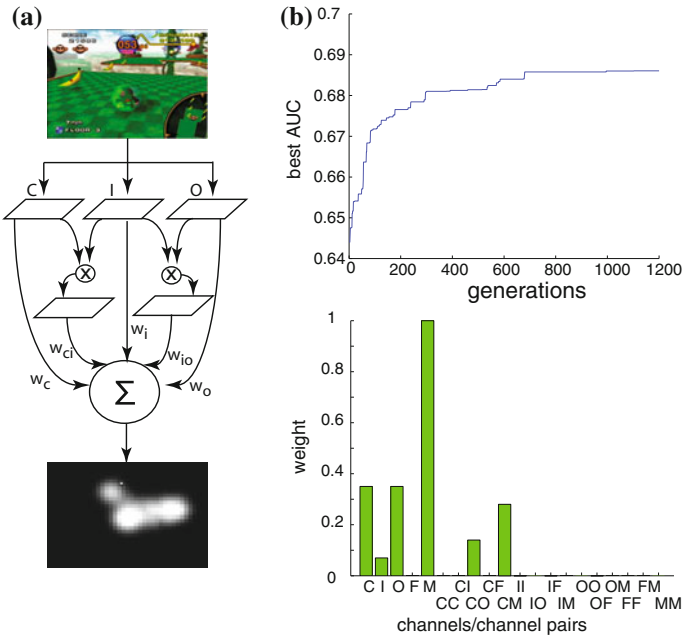

Fig. 14.7 Inter-observer model. **a** The manner in which features were combined to form second order terms and were then linearly combined. 20 weights were learned by the genetic algorithm. **b** *Top* shows the evolution of AUC scores as a function of generation of the population, the score converges around 800 generations. *Bottom* shows the weights for each term

generated a total of 20 features including combinations such as CO (color orientation combination), CI (color intensity interaction) etc. Once again we used the genetic algorithm approach to search for the parameters for this model. In this case we had 20 features to learn and therefore this was a much larger optimization problem which took longer to converge on a solution. Figure 14.7 shows the manner in which the genetic algorithm learned the best weights and converged to a solution after about 800 generations. This is significantly longer than the linear weighted model with five terms as would be expected since a much higher dimensional space (20) was explored in this experiment.

The results from this experiment are surprising in that second-order complex features do not help boost performance and the genetic algorithm converges to a solution that is similar to the linear case with only first order terms. Motion again is the strongest feature. The best AUC score 0.69 is similar to the model with only first order term. Second order features therefore did not improve the score and while one would expect this additional interaction information to help predict eye movements better, the principle of parsimony compels us to consider the first order model the better one. This is somewhat consistent with evidence from the physiology of the visual system in that there is a very small number of cells that might be tuned to second order combinations of features. While there is evidence of hierarchy of

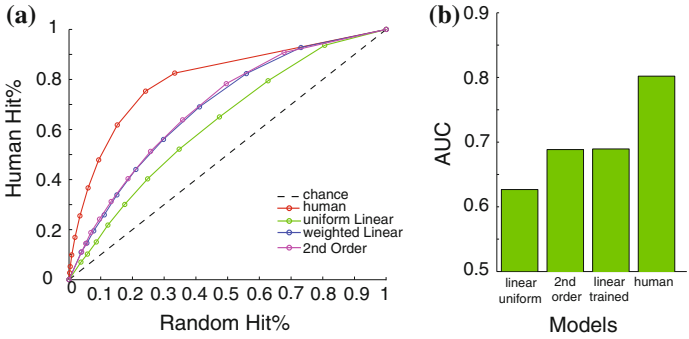

Fig. 14.8 Model comparison. **a** The ROC *curves* for the different models computed. Chance is represented by the *dashed black line* the *red line* labeled *human* in the legend represents the inter-observer model which performs the best. The uniform linear model represents the model using CIOFM features in a simple linear combination with uniform weights. The weighted linear model is the one that was trained using a genetic algorithm but consists only of linear terms. The 2nd order model uses both first order and 2nd order terms to define feature combinations. **b** The AUC computed from the *curves* in **a**. As can be seen the second order model does not perform any better than the linear model

features building up to a single percept, neurons tuned to combinations of features within our set are limited to color-orientation and color-motion cells [19, 30]. Even in lower brain areas like the superior colliculus, a key structure that enables the mechanisms of attention, there exist cells that are responsive to motion and even color [8, 33].

14.4.3 Model Comparison

We compared the performance of all the models studied (Fig. 14.8), i.e., linear model with uniform weights, linear model with learned weights, trained model with 2nd order terms and the human inter-observer model. The linear model with trained weights performs the best, but the inter-observer model has the highest performance in predicting human attention. This shows that a linear model with no higher level knowledge of the semantic content of the scene performs the best when compared to other models.

It is important and interesting to note that the model that included both linear and second-order terms performed no better than a model that consisted of linear terms but these terms were weighted. The weights were learned by using a genetic algorithm that maximized the AUC score. These results suggest that a biologically inspired model of attention that combines features in a linear manner to form a saliency map is likely to be closely related to mechanisms of attention in the brain that give rise to the observed eye movements.

14.5 Conclusions

In this chapter we processed and decomposed videos into several features and then searched for a good combination of features that can predict human attention allocation. Human attention consists of a volitional top-down component and an image driven bottom-up component. We presented a study that focused on the bottom-up aspects of attention. By recording of human observers as they watched natural videos we established a means to validate various models explored.

A linear combination of features was sufficient to provide prediction of human gaze, and second-order interactions of these features did not help performance. Therefore salience of a region in an image is determined through a linear combination of features and we can account for almost 70 % of the variance through a weighted linear model. Top-down attention then may act by providing the weights that were learned by our genetic algorithm.

While the linear combination model did reasonably well in predicting human gaze, an inter-observer model built from the eye movements of several observers outperformed the linear model. Humans are therefore better at predicting the eye movements of each other when compared to such a model of saliency. There is much further research to be done to both elucidate the mechanisms of attention in humans as well as build models that can mine videos for features that drive attention.

Acknowledgments This work was supported by the National Science Foundation (grant numbers CCF-1317433), the Office of Naval Research (N00014-13-1-0563), and the Army Research Office (W911NF-12-1-0433). The authors affirm that the views expressed herein are solely their own, and do not represent the views of the United States government or any agency thereof.

References

1. Baluch F, Itti L (2010) Training top-down attention improves performance on a triple-conjunction search task. PLoS One 5(2):e9127
2. Baluch F, Itti L (2011) Mechanisms of top-down attention. Trends Neurosci 34(4):210–240
3. Berg DJ, Boehnke SE, Marino RA, Munoz DP, Itti L (2009) Free viewing of dynamic stimuli by humans and monkeys. J Vis 9 5(19):1–15
4. Borji A, Itti L (2013) State-of-the-art in visual attention modeling. IEEE Trans Pattern Anal Mach Intell 35(1):185–207
5. Chiang A-YD, Berg D, Itti L (2011) Saliency, memory, and attention capture in marketing. J Vis 11(11):493–493
6. Connor CE, Egeth HE, Yantis S (2004) Visual attention: bottom-up versus top-down. Curr Biol 14(19):R850–R852
7. Elazary L, Itti L (2008) Interesting objects are visually salient. J Vis 8(3):3
8. Fecteau J, Bell A, Munoz D (2004) Neural correlates of the automatic and goal-driven biases in orienting spatial attention. J Neurophysiol 92(3):1728–1737
9. Gilbert C, Sigman M (2007) Brain states: top-down influences in sensory processing. Neuron 54(5):677–696

10. Grant WS, Itti L (2012) Saliency mapping enhanced by symmetry from local phase. In: Proceedings of IEEE international conference on image processing (ICIP). Florida, pp 653–656
11. Itti L (2008) Crcns data sharing: eye movements during free-viewing of natural videos. In: Collaborative research in computational neuroscience annual meeting. California
12. Itti L, Dhavale N, Pighin F (2003) Realistic avatar eye and head animation using a neurobiological model of visual attention. In: Bosacchi B, Fogel DB, Bezdek JC (eds) Proceedings of SPIE 48th annual international symposium on optical science and technology, vol 5200. SPIE Press, Bellingham, pp 64–78
13. Itti L et al (1998) The ilab neuromorphic vision C++ toolkit (INVT), http://ilab.usc.edu/toolkit
14. Itti L, Koch C (2001) Computational modelling of visual attention. Nat Rev Neurosci 2(3): 194–203
15. Itti L, Koch C (2001) Feature combination strategies for saliency-based visual attention systems. J Electron Imaging 10(1):161–169
16. Itti L, Koch C, Niebur E (1998) A model of saliency-based visual attention for rapid scene analysis. IEEE Trans Pattern Anal Mach Intell 20(11):1254–1259
17. Kienzle W, Franz MO, Schölkopf B, Wichmann FA (2009) Center-surround patterns emerge as optimal predictors for human saccade targets. J Vis 9(5):7
18. Koch C, Ullman S (1985) Shifts in selective visual attention: towards the underlying neural circuitry. Hum Neurobiol 4(4):219–227
19. Koene AR, Zhaoping L (2007) Feature-specific interactions in salience from combined feature contrasts: evidence for a bottom-up saliency map in v1. J Vis 7(7):6
20. Li Z, Itti L (2011) Saliency and gist features for target detection in satellite images. IEEE Trans Image Process 20(7):2017–2029
21. Li Z, Qin S, Itti L (2011) Visual attention guided bit allocation in video compression. Image Vis Comput 29(1):1–14
22. Moore T (2006) The neurobiology of visual attention: finding sources. Curr Opin Neurobiol 16(2):159–165
23. Mundhenk TN, Itti L (2005) Computational modeling and exploration of contour integration for visual saliency. Biolo Cybern 93(3):188–212
24. Navalpakkam V, Itti L (2006) An integrated model of top-down and bottom-up attention for optimal object detection. In: Proceedings of IEEE conference on computer vision and pattern recognition (CVPR). New York, pp 2049–2056
25. Peters RJ, Iyer A, Itti L, Koch C (2005) Components of bottom-up gaze allocation in natural images. Vis Res 45(8):2397–2416
26. Rajashekar U, Bovik AC, Cormack LK (2006) Visual search in noise: revealing the influence of structural cues by gaze-contingent classification image analysis. J Vis 6(4):7
27. Treisman A, Gelade G (1980) A feature-integration theory of attention. Cogn Psychol 12(1): 97–136
28. Tseng P, Cameron IGM, Pari G, Reynolds JN, Munoz DP, Itti L (2013) High-throughput classification of clinical populations from natural viewing eye movements. J Neurol 260: 275–284
29. Tseng P, Carmi R, Cameron IGM, Munoz D, Itti L (2009) Quantifying center bias of observers in free viewing of dynamic natural scenes. J Vis 9 7(4):1–16
30. Ts'o D, Gilbert CD (1988) The organization of chromatic and spatial interactions in the primate striate cortex. J Neurosci 8(5):1712–1727
31. Walther D (2006) Interactions of visual attention and object recognition: Computational modeling, algorithms, and psychophysics. PhD thesis, California Institute of Technology
32. Walther D, Itti L, Riesenhuber M, Poggio T, Koch C (2002) Attentional selection for object recognition—a gentle way. In: Biologically motivated computer vision. Springer, pp 472–479
33. White BJ, Boehnke SE, Marino RA, Itti L, Munoz DP (2009) Color-related signals in the primate superior colliculus. J Neurosci 29(39):12159–12166

34. Wolfe JM, Horowitz TS (2004) What attributes guide the deployment of visual attention and how do they do it? Nat Rev Neurosci 5(6):495–501
35. Yarbus AL, Haigh B, Rigss LA (1967) Eye Mov Vis. Plenum Press, New York
36. Yoshida M, Itti L, Berg DJ, Ikeda T, Kato R, Takaura K, White BJ, Munoz DP, Isa T (2012) Residual attention guidance in blindsight monkeys watching complex natural scenes. Curr Biol 22(15):1429–1434
37. Yubing T, Cheikh FA, Guraya FFE, Konik H, Trémeau A (2011) A spatiotemporal saliency model for video surveillance. Cogn Comput 3(1):241–263