Factors underlying Inter-observer agreement in gaze patterns: Predictive modelling and analysis

Shafin Rahman
Neil D.B. Bruce*
University of Manitoba

Abstract

In viewing an image or real-world scene, different observers may exhibit different viewing patterns. This is evidently due to a variety of different factors, involving both bottom-up and top-down processing. In the literature addressing prediction of visual saliency, agreement in gaze patterns across observers is often quantified according to a measure of inter-observer congruency (IOC). Intuitively, common viewership patterns may be expected to diagnose certain image qualities including the capacity for an image to draw attention, or perceptual qualities of an image relevant to applications in human computer interaction, visual design and other domains. Moreover, there is value in determining the extent to which different factors contribute to inter-observer variability, and corresponding dependence on the type of content being viewed. In this paper, we assess the extent to which different types of features contribute to variability in viewing patterns across observers. This is accomplished in considering correlation between image derived features and IOC values, and based on the capacity for more complex feature sets to predict IOC based on a regression model. Experimental results demonstrate the value of different feature types for predicting IOC. These results also establish the relative importance of top-down and bottom-up information in driving gaze and provide new insight into predictive analysis for gaze behavior associated with perceptual characteristics of images.

Keywords: gaze patterns, complexity, objects, modelling, prediction

Concepts: •**Computing methodologies** → *Scene understanding; Perception;*

1 Introduction

In past decade, numerous models have been proposed towards predicting interesting (salient) locations within an image, video or scene. The goal of such models is typically to predict human viewing behavior, with eye tracking data used to quantify algorithm performance [Borji et al. 2013]. One metric that forms part of this analysis is termed inter-observer congruency (IOC). This measures the extent to which eye movement data from a subset of observers is able to predict the fixations of other observers. This has been used in saliency research to provide an upper bound on the capacity for visual saliency models to successfully predict gaze patterns. IOC itself has received much less attention outside of its role in supporting models of visual saliency. IOC prediction presents an important adjunct determination at the *whole image* level. That is, what factors contribute to commonality in viewing patterns, and to what extent can one predict whether an image will elicit common viewing patterns across observers?

One means of addressing this question, is comparing different image derived statistics with measured IOC scores on a per image basis. In this paper, we therefore examine correlation between IOC scores and various image-derived statistics related to complexity of image structure. In prior research addressing perceptual characteristics of images, predictive models are considered that employ machine learning to predict image qualities including memorability [Isola et al. 2011], aesthetic appeal [Murray et al. 2012] and other *image level* categories. The predictions are based on holistic characterizations of entire scenes (e.g. features diagnostic of scene category) [Xiao et al. 2010] or other image derived features (e.g. based on visual saliency). In a similar vein, we also examine the extent to which commonality in viewing patterns across observers may be predicted based on combinations of different types of features. This allows for analysis that is richer than the more basic correlation based approach, and also presents the ability to gauge the relative importance of different factors from low-level image properties, to higher-level concepts (e.g. objects).

The IOC value of an image may be calculated directly from the eye tracking data of many users through a signal detection theoretic approach. One prior effort by LeMeur et al. [Le Meur et al. 2011] has examined a number of different factors in their ability to predict IOC scores for individual images. In this prior work, a regression based learning approach is applied to predict IOC values. Features considered by LeMeur et al. [Le Meur et al. 2011] include a Haar cascade based *face detector* [Viola and Jones 2004], *color harmony* which measures dispersion in hue and saturation within the image, *depth of field* estimated by differences in sensitivity to blur kernels of varying size, and *scene complexity* based on Lab space channel entropy, granularity of mean shift segmentation, and average edge energy.

In this paper, we explore in detail the capacity of different types of features to predict IOC values. This is done across 3 separate data sets for a more complete picture of the relative value of features. This includes correlation based analysis for some single features including local entropy, visual clutter, and JPEG image size. The rationale behind these measurements, is in the relationship between these quantities and image complexity including irregular structure, variations in shape, texture, clutter or noise. These factors presumably affect the viewing patterns of observers and might imply more variability. Establishing a base value for correlation associated with these individual features provides a sense of how different measures of image complexity interact with IOC. As discussed, this is accompanied by regression based predictive modeling using other rich feature sets with the goal of predicting IOC scores. These additional types of features are divided into *Bottom-Up* and *Top-Down* feature sets, reflecting the extent to which they are pattern driven, or involve prior knowledge and experience.

*e-mail:bruce@cs.umanitoba.ca

ETRA 2016, March 14 - 17, 2016, Charleston, SC, USA
ISBN: 978-1-4503-4125-7/16/03
DOI: http://dx.doi.org/10.1145/2857491.2857495

In predicting IOC, intuitively existing algorithms that predict gaze patterns may be of value. Saliency algorithms are therefore considered as one type of feature within our analysis of IOC. Moreover, different saliency produce different predictions and therefore both individual model predictions, and differences among models may have value. We therefore propose a histogram based measure named Histogram of Predicted Salience (HoPS) for prediction of image level IOC based on summary statistics of visual saliency output across various (high performing [Borji et al. 2013]) algorithms. HoPS therefore presents a feature vector containing concatenated histograms derived from saliency images based on 12 well established algorithms [Borji et al. 2013]. A support vector regression (SVR) model is applied to predicting IOC using a RBF kernel within the regression model. Results demonstrate that HoPS and other holistic representations predict IOC scores better than some of the simpler diagnostic features, with HoPS also outperforming traditional structural measures of image content such as HoG [Felzenszwalb et al. 2010] and Gist [Oliva and Torralba 2001] features. We also establish that the value of different types of features (*Bottom-up* vs. *Top-Down*) in making this prediction depends on the type of content within the images. For simpler images with a weaker presence of behaviorally relevant patterns such as faces, simpler measures based on image structure such as HoPS are most predictive of agreement in gaze patterns. However, when there is a greater presence of behaviorally relevant patterns, high-level object detectors carry much more value. The combined strength of these 2 disparate feature sets is also shown to be especially effective across different types of data.

In the Discussion section of this paper, we discuss the broader implications of these results and also the relationship between scene complexity, compressibility of images, neural encoding and IOC.

2 Inter-observer congruency (IOC)

There are several factors which may be determine image IOC, or in particular, be cause for a low IOC score. Firstly, IOC can be low if the image has no *salient* content i.e. nothing is prominent or especially significant to attract attention. This is a scenario that is not uncommon for naturalistic images for which algorithm performance and upper bounds dictated by IOC are especially low [Rahman et al. 2014]. Secondly, IOC can be low if the image has many salient locations, coupled with a shorter viewing time. Such scenarios may arise when the image contains many objects or outliers with potential to attract attention. It is also commonly observed that the first few fixation points show significant congruency among observers with inter-observer agreement progressively dropping off with time [Le Meur and Baccino 2013]. On a slightly longer time course, task dependence and contextual guidance exert greater influence on search behaviour which also interacts with congruency. Given a wide array of salient targets in a single scene, fixations are divided differently among observers across targets. Therefore, there exists an upper bound on performance in predicting the viewing patterns of observers quantified by the Inter-observer Congruency score. In Figure 1, we show a ranking of images corresponding to the 5 highest and 5 lowest IOC values for the three datasets considered in the current work. There are a few observations that may be made immediately at an anecdotal level concerning these images. Those that exhibit high IOC scores tend to be marked by one of: (i) A small number of discrete, or localized regions of high contrast (ii) the presence of people, or faces in the scene. (iii) the presence of text within the scene. On the low end, images tend to consist of a broad distribution of structure, cluttered or busy scenes, and landscapes. The low IOC images also appear somewhat more generic in what they capture.

IOC is typically quantified based the average IOC across all images

Figure 1: *Images are sorted based on their actual IOC score from (a) Bruce (b) Judd (c) Memorability datasets. Top rows corresponding to each dataset depict those images having low IOC and bottom rows depict images having high IOC. Specific properties common among higher or lower IOC cores at a semantic level are discussed in detail throughout the paper.*

within a dataset. In the current study we consider the IOC score on a per image basis, for each individual image. For 3 individual and distinctive datasets, we have computed individual IOC scores for each image. This is accomplished through the *leave one out* method for IOC computation [Torralba et al. 2006]. Each dataset includes eye tracking data from n observers. For any $i \in n$, a predictive map based on $(n-1)$ observers is used to predict fixation data for the ith observer through ROC analysis. The AUC for the ROC curve is computed for all possible leave-one-out combinations. For AUC computation, we have used the shuffled AUC process employed in [Borji et al. 2013; Kanan et al. 2009]. The IOC of each image is given by averaging the AUC of all $i \in n$. In creating the predictive output from $(n-1)$ observers, the raw fixations of $(n-1)$ observers are convolved with a 2-D Gaussian distribution. The optimal size of Gaussian window was determined independently for each data set, in considering the maximal mean IOC score across the entire data set. The optimal σ varies in a fashion related to the resolution/scale of the input images. This provides a measure of the degree of agreement in viewing patterns across observers for each image within each dataset.

3 Experimental Methods and Results

To better understand the problem of IOC prediction, we have chosen three well-known and distinct datasets: The Bruce/Toronto dataset [Bruce and Tsotsos 2009], the Mancas/LeMeur dataset [Mancas and Le Meur 2013] and the Judd/MIT dataset [Judd et al. 2009]. The Bruce/Toronto dataset contains 120 color images of 681×511 resolution with 20 human observers. Images from the Mancas and Le Meur [Mancas and Le Meur 2013] dataset include

fixations for 17 observers, an image resolution of 384×384 pixels, and 135 total images. We have also considered 464 images (Landscape orientation) with resolution 1024×768 from the Judd/MIT dataset. The landscape orientation subset was selected to avoid diminished performance from including both landscape and portrait orientations. In particular given Gist features as one category for testing, the *squashing* of images to a square shape may present a disadvantage if differences in initial image sizes are present. Images across all databases include a wide variety of scene categories and objects, and include natural scenes, indoor, and outdoor scenes. There are some notable differences across the datasets, with one important difference being the relatively fewer instances of faces or text in the Bruce/Toronto dataset (This is evident in Fig. 1).

3.1 Simple diagnostic features

As a starting point, we consider a number of simple features related to image complexity, some of which overlap with measures in the study of LeMeur et al. [Le Meur et al. 2011]. In this subsection, we describe these features and the underlying reasons for assuming a possible relation to IOC.

Entropy: Entropy is a measure of randomness, or local heterogeneity. Images typically exhibit correlation in their pixel values locally, and are far from being completely random. High entropy at the image level implies a weaker local correlation structure expressed broadly across the image. This will tend to occur in instances where there is a wide variety of objects/items/structure in the image, or sources of noise. Local entropy is computed based on the distribution of local grey values with a bin count matching bit depth.

Visual clutter: Visual clutter is a quantitative measure of the level of ease to add a target that will draw attention within an existing image [Rosenholtz et al. 2007]. The clutter measure of Rosenholtz [Rosenholtz et al. 2007] is the most well established measure of visual clutter, and is therefore used to gauge the impact of image clutter on IOC. The visual clutter measure is centers around the notion of feature congestion, based on feature covariance expressed across multiple scales. This interacts with a statistical visual saliency metric also based on feature covariance to derive predictions concerning target detectability. This approach is validated in considering the match to psychophysical performance in target detection in the presence of clutter. As clutter increases, one would naturally expect the associated IOC to fall given the distraction from clutter may imply less potential for individual items in the image to elicit perceptual contrast.

JPEG image size: The effect of image compression has been related to image clutter [Rosenholtz et al. 2007] and is also intuitively related to scene complexity. Other interesting observations have also been made on the relationship between compressibility and visual saliency [Hossein Khatoonabadi et al. 2015]. Regularity in an image will tend to result in larger gains in compression. In this instance, we have employed the JPEG 2000 method due to its use of the Discrete Wavelet transform as a representation. Given a Wavelet decomposition as the basis for representation there is an inherent relationship to the spectral decomposition that one observes in early visual cortex, in the form of Gabor-like features with selectivity for angular and radial frequencies. An implication of this, is that one might expect the compressibility of an image as expressed by a spectral decomposition, carries diagnosticity concerning compressibility from a neural perspective. For a fixed resolution, if the JPEG image size is high, one might expect more irregular patterns, textures, or noise. JPEG quality is also relevant in the number of coefficients that are driven to zero and its interaction with scale.

Without any learning these single diagnostic features demonstrate a significant degree of inverse correlation with IOC scores. It is somewhat surprising that some simple features are inherently tied to a complex prediction such as IOC. That said, a common denominator across each of these is the relation to sparsity, compressibility of signal, redundancy and efficient representation in an information theoretic sense. There are numerous examples that suggest that these are some of the central tenets of cortical representation [Jazayeri and Movshon 2006; Wainwright 1999; Olshausen and Field 1997], which perhaps sheds light on the nature of this relationship. In Table 1, we present the Pearson correlation value (r) with confidence p_{val} for each single diagnostic feature. We observe that JPEG image size and visual clutter have significant negative correlation with IOC. These values are also useful as a relative reference point for considering the performance of IOC prediction subject to richer feature sets that follow. We have also tested the correlation in applying varying degrees of JPEG compression. When quality is very low, many wavelet coefficients are ignored and below a certain threshold one may fail to adequately represent the image even at a relatively coarse grained level. When quality is increased above a certain threshold, then correlation does not change much with increasing quality. One does tend to observe a maximal degree of correlation for a specific intermediate level of JPEG compression quality for any given dataset. These results suggest a relative lack of importance of content above some high-frequency limit. This observation is consistent with the nature of foveation, and that fixation patterns tend to be relatively stable as a function of image above a critical limit [Judd et al. 2011].

It is interesting to note for these relatively simple features, that there is a stronger degree of correlation observed with the Bruce/Toronto dataset as compared with the 2 other datasets. It is conceivable that a weaker presence of *high-level* patterns (e.g. faces) is involved in this difference. In a free-viewing paradigm, one observes behavior that is distinct from explicit judgements of visual salience [Koehler et al. 2014], and there is evidently the involvement of complex patterns in the range of behaviour observed.

3.2 More complex features

Having established that entropy, visual clutter and JPEG image size provide simple features with diagnostic value for predicting IOC, we now turn to the strategy of extracting richer feature vectors in combination with a learning algorithm to determine the value of scene level, structural, saliency based, or object based features in predicting IOC. The set of features used in predicting IOC through a learning strategy follow:

Bottom-Up and Top-Down Image Analysis

An important distinction that often appears within the visual attention literature, is the notion of bottom-up versus top-down processing. Bottom-up processing refers to the viewing behavior that is driven by the properties of the image content itself. A bright pattern, an unusual patch of color, or sudden movement tends to draw an observers gaze independent of any semantic information, or higher level concepts. Top-Down processing refers to the portion of viewing behavior that is driven by higher level cognition or prior knowledge. This may appear in the context of a certain task (e.g. looking for keys), or based on inherent reward seeking behaviour. For example, there is a strong tendency to look at faces or other socially relevant cues [Le Meur et al. 2011]. It is also observed that the overall holistic structure of a scene can be understood rapidly, and influence how a scene is examined [Oliva and Torralba 2001]. To test the relative contribution of these different processes to observed viewing patterns, we therefore include a number of features that span these categories.

Saliency algorithms by design seek image content that draws inter-

Table 1: *Correlation between actual IOC scores for simple diagnostic features determined to have significant correlation with IOC. Parentheses next to JPEG entries indicates quality factor ranging from very low (3) to lossless (100).*

Datasets	Bruce Dataset		Memorability Dataset		Judd Dataset	
Criteria	r	p_{val}	r	p_{val}	r	p_{val}
Entropy	-0.188	0.0396	-0.082	0.3428	-0.072	0.12227
Clutter FC	-0.345	0.0001	-0.171	0.0476	-0.187	0.00005
JPEG (3)	-0.293	0.0012	-0.132	0.1282	-0.203	0.00001
JPEG (5)	-0.321	0.0004	-0.164	0.0578	-0.208	0.00001
JPEG (20)	-0.341	0.0001	-0.179	0.0383	-0.200	0.00001
JPEG (45)	-0.336	0.0002	-0.179	0.0382	-0.200	0.00001
JPEG (75)	-0.327	0.0003	-0.181	0.0357	-0.197	$< 10^{-5}$
JPEG (100)	-0.332	0.0002	-0.188	0.0292	-0.193	0.00003

Table 2: *Correlation between actual IOC scores and predicted IOC scores using different* holistic *feature vectors for prediction.*

Datasets	Bruce Dataset		Memorability Dataset		Judd Dataset	
Criteria	r	p_{val}	r	p_{val}	r	p_{val}
HoG	0.370	0.00367	0.434	0.00027	0.340	0.2993
Gist	0.483	0.00009	0.438	0.00024	0.346	0.0154
HoPS	0.505	0.00004	0.470	0.00007	0.430	0.0156
DeepNet features (1L)	0.331	0.00990	0.370	0.00221	0.397	$< 10^{-5}$
DeepNet features (2L)	0.365	$< 10^{-5}$	0.473	0.004	0.444	$< 10^{-5}$
DeepNet (1L) + HoPS features	0.506	0.00004	0.519	0.00001	0.456	$< 10^{-5}$
LeMeur et al. [Le Meur et al. 2011]	N/A	N/A	N/A	N/A	0.340	<0.001

est from a stimulus driven perspective. Different models rely on different simple features, and measures of contrast with this common goal in mind. We therefore consider a rich bottom-up feature set derived from the combined output of a variety of popular saliency algorithms. The specifics of these features are described in more detail later in this section.

We also seek to include features that convey an overall structural representation of the scene. One popular method for representing the holistic scene envelope is through scene Gist [Oliva and Torralba 2001]. This provides a set of holistic receptive fields that distinguishes between different categories of scene that carry different holistic structure. We have also used the more standard histogram of oriented gradients (HoG) [Felzenszwalb et al. 2010] feature in a similar capacity. These representations are applied to the entire image to produce a feature vector that captures a coarse grained representation of scene structure.

For high-level features (top-down), we rely on object specific features that have shown significant success in large scale recognition tasks. These are features derived through deep learning using convolutional nets. High level features are derived from the BVLC Reference CaffeNet architecture [Jia et al. 2014] based on the AlexNet architecture [Krizhevsky et al. 2012] (and trained on the ILSVRC12 challenge data [Russakovsky et al. 2014]). This has been used in sampling outputs from the final layer only of responses corresponding to the highest layer of the deep convolutional net (referred to as 1L), and from the penultimate layer comprised of 4096 features (referred to as 2L). These both capture high-level concepts in the patterns they respond to, but vary in the relative specificity. The rationale for these features is to capture the presence of higher level concepts and patterns within the images, and to related these measurements to IOC. It is worth noting that subsequent efforts have revealed even greater capabilities for alternative network architectures (e.g. VGG-16 [Simonyan and Zisserman 2014]), and even greater performance from high-level semantically relevant features might be had in leveraging this and other alternatives. This is the subject of ongoing investigation.

Histogram of Predicted Salience (HoPS) features:

IOC depends of variability in viewing patterns of different observers. Each observer may have their own way of analyzing a scene, that guides the viewing patterns observed in fixation data. As different saliency algorithms may emphasize different types of feature contrast within an image, a feature level representation derived from a variety of saliency models is a natural option to consider. We have considered 12 different saliency algorithms in producing a feature vector to predict IOC. These algorithms include the Torralba [Torralba 2003], HouCVPR [Hou and Zhang 2007], HouNIPS [Hou and Zhang 2009], Itti-CIO2 [Itti et al. 1998], ImageSignatureLab [Hou et al. 2012], ImageSignatureRGB [Hou et al. 2012], SDSR [Seo and Milanfar 2009], AIM [Bruce and Tsotsos 2009], GBVS [Harel et al. 2007], AWS [Garcia-Diaz et al. 2012], Yan [Yan et al. 2010] and SOC [Rahman et al. 2014] models. For a detailed description of these algorithsm, the reader may refer to the original published work, or Borji et al. [Borji et al. 2013] for a summary. It is important to note that the specific choice of algorithms is not critical to the value of this feature. As new algorithms emerge, these may be used in place of in conjunction with those we have considered. The HoPS features are computed as follows:

For each of the aforementioned saliency algorithms, a histogram is created that provides a summary representation of the distribution of predicted visual saliency within a rectangular region of the image. Histograms are normalized and subsequently concatenated to produce a feature set for prediction. As these features are derived from saliency maps, we call these features Histograms of Predicted Salience (HoPS) in a manner similar to features based on histograms of normalized edge structure [Felzenszwalb et al. 2010; Lowe 2004]. Suppose, for the ith algorithm the normalized histogram is represented by $[f_1^i, f_2^i, f_3^i ....f_b^i]$ where $i \in 1, 2, ..., 12$ and b is number of bins in the histogram. The HoPS feature set is given by $f = [f_1^1, f_2^1, ...f_b^1, f_1^2, f_2^2, ...f_b^2, ...f_1^{12}, f_2^{12}, ...f_b^{12}]$

Where, any value f_j^i represents jth bin feature of the ith algorithm. If different algorithms agree with one other then this implies confidence across algorithms in a common region. Because of this char-

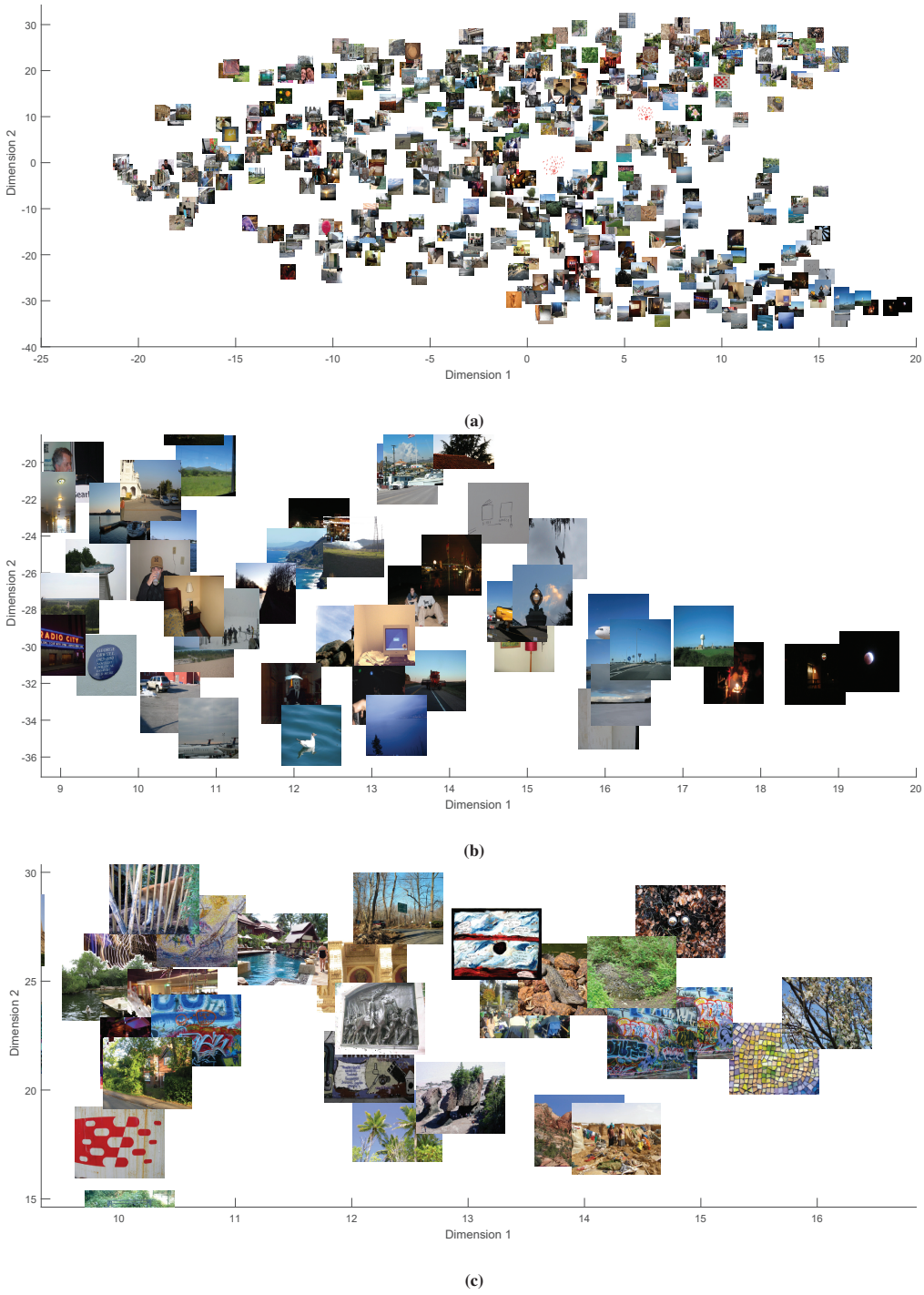

Figure 2: *A t-SNE embedding of images in 2-dimensions based on HoPS features. (a) All images from the Judd dataset. (b) A zoomed in subset revealing grouping of examples with simplistic structure. (c) A zoomed in subset revealing complex structure and patterns.*

acteristic, HoPS has very good capability to represent both variability in predicted salience, but also consistency and commonality across algorithms.

Learning: Extracted features are used to learn a regression model for IOC prediction. We have randomly selected half of the images from each dataset for training. Most of the images in any dataset exhibit a shuffled ROC score that falls within the range of 0.5 to 0.7. This may result in imbalance in the dataset in the relative representation of high and low IOC examples relative to average cases. To accommodate for bias in the distribution of IOC values, training is achieved in merging images of all three datasets. This results in a greater proportion of relatively high and low IOC values on a relative scale. As eye tracking data are derived from different numbers of observers for different datasets, we normalize the scale of IOC values via linear scaling for each dataset to a common range of 0 to 1. This helps to overcome the paucity of data inherent in perform-

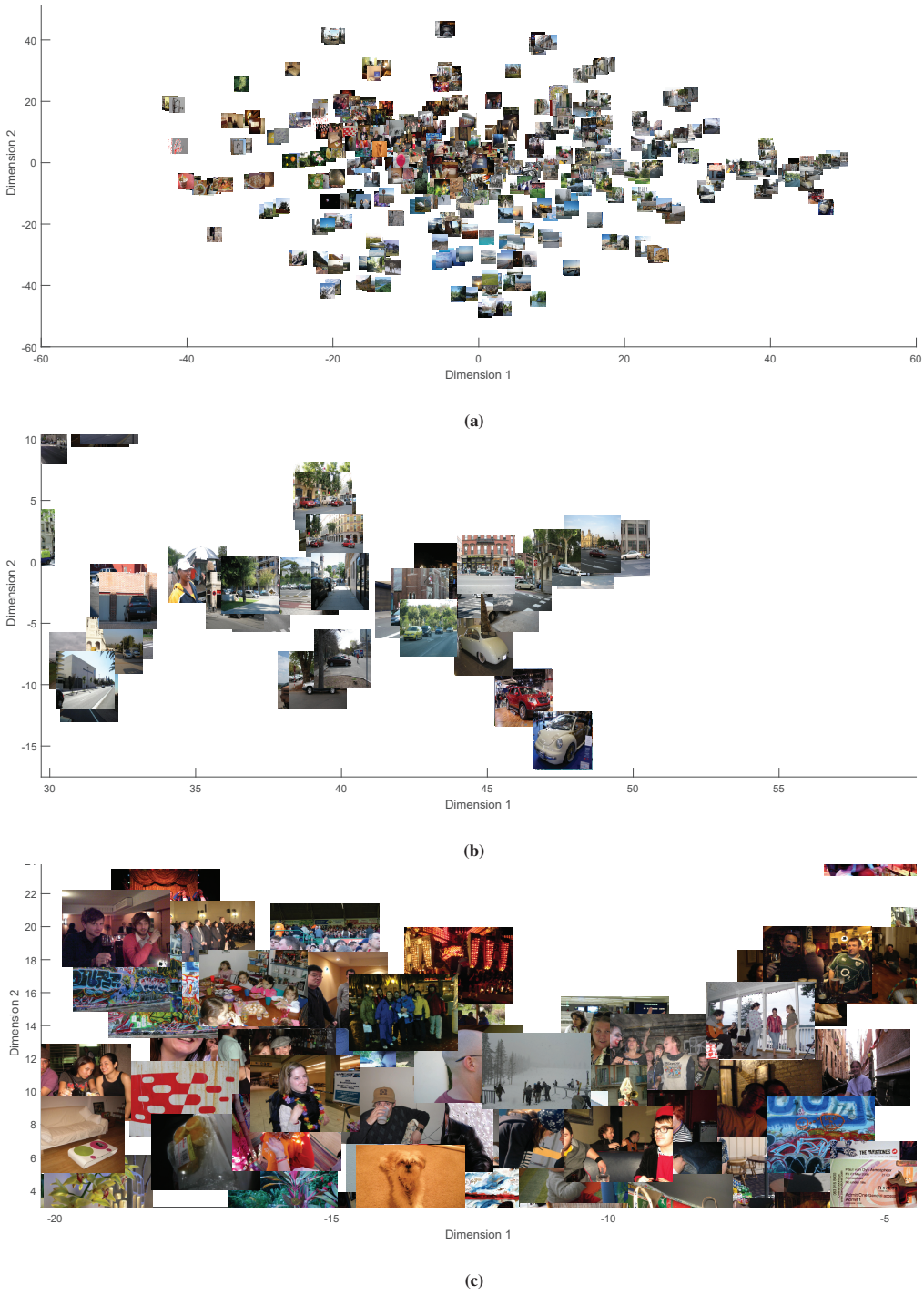

Figure 3: *A t-SNE embedding of images in 2-dimensions based on Deep Learning features. (a) All images from the Judd dataset. (b) A zoomed in subset revealing examples that include cars. (c) A zoomed in subset revealing examples containing mostly people.*

ing an image-level prediction (as opposed to pixel level where there are many more samples). The limited number of data samples is the reason for using a single dataset for both training and testing within the prior work of LeMeur et al. [Le Meur et al. 2011], Combining these data sets allows us to successfully overcome this limitation. Repeated selection of sub-samples of the training set allows for the selection of training samples with a more uniform range following the same strategy employed by Isola et al. [Isola et al. 2011].

Given training instances represented by feature vectors (HoG, GIST, HoPS, DeepNet) we apply ϵ-SVR [Chang and Lin 2011] for continuous value prediction. The dimensionality of HoPS features may be altered by varying the number of bins. Performance is relatively consistent across a range of bin values (5 to 100 bins were tested) with 20 bins used in test results. While either of HoPS or HoG may be applied subject to a grid or blockwise decomposition of the image, it was determined that there was no significant benefit

to such a decomposition when compared with features of the same variety derived over the complete image. This is somewhat surprising, however the degree of contrast in the saliency maps is quite diagnostic of the presence of highly salient regions, independent of where they are located. Parameters were also optimized for each of HoG and Gist to achieve the greatest correlation with IOC scores. These results correspond to HoG using a cell size of 128 and 8 orientations. Similarly, the best performance of GIST for orientation per scale was [2 2 2] (3 scales, 2 orientations per scale) and a subdivision using 4 blocks. As images from different datasets may have different resolution, these are resized to a fixed dimension for HoG and GIST which are 256×256 and 128×128 respectively. The results of these experiments are shown in Table 2. We observe that holistic structural feature extraction and summary statistics across saliency maps produce stronger correlation than that observed for the simpler features that were initially considered. Moreover, HoPS features outperform both HoG and GIST features suggesting that the distribution of saliency scores, and agreement across algorithms has strong diagnosticity for predicting IOC. It is also noted that out-of-bag analysis reveals that bins corresponding to the lowest levels of saliency are most diagnostic of IOC. Features considered in the LeMeur et al. study [Le Meur et al. 2011] present r values for Pearson correlation of 0.34 for the Judd Dataset, and 0.27 for an alternative dataset [Le Meur et al. 2006].

It is also interesting to note the contrast in the high level features derived from deep learning, and the HoPS features. While the HoPS features are of greatest value for the Bruce Dataset, they are of less value for the Judd dataset in particular in combination with high-level features derived from deep learning. The features derived from the deep neural network show the opposite trend, with these features having relatively less value for the Bruce dataset. This fits with the observation that more of the salient patterns in the Bruce dataset are a result of simpler feature configurations whereas the Judd dataset has a much higher incidence of people, faces and text. It is also noteworthy, that the best performance derives from combining the low-level stimulus driven features with the high-level object sensitive features. This underscores the importance of modeling both bottom-up and top-down facets of viewing behaviour.

It is useful to understand the characteristics that different types of features capture in relating these to IOC values. In Figures 2 and 3, a 2-dimensional embedding of the images from the Judd dataset are shown corresponding to the HoPS features, and DeepNet features respectively. The embedding is based on t-stochastic neighbor embedding [Van der Maaten and Hinton 2008], and affords a visualization of similarity of images according to the different high dimensional feature representations. Each example shows some image subsets close up to highlight specific characteristics of each image set. The HoPS features seem to carry a sense of image complexity, akin to the entropy and clutter based measures, albeit with additional nuances. The DeepNet features as expected, include groupings according to objects that are present. This provides an interesting orthogonal dimension of analysis, and one that is quite different than measurements typically considered in examining gaze data, or considering perceptual qualities of images. Analogous plots for the image memorability dataset are similar. For the Bruce dataset, the object groupings produced by the DeepNet features are weaker owing to a relative absence of the same semantically relevant features appearing in the Judd dataset. This visualization, combined with the regression coefficients is revealing with respect to both factors of importance in driving gaze behaviour, and also in bias accompanying the data under consideration.

4 Discussion

In this paper, we have explored the reasons underlying variability in gaze behaviour across observers. This has been examined according to the capacity to predict IOC for previously unseen images based on different types of image derived features. The extent to which this characterization may be made evidently depends in part on the nature of the images under consideration, including their contents and composition. IOC correlation associated with a number of low-level simplistic features hints at the importance of redundancy and neural coding in how a scene is parsed by a human observer. The success of saliency models built on principles drawn from information theory, compressive sensing and coding, (divisive) normalization and whitening evidently mirrors some of the characteristics that are relevant in characterizing gaze behavior in a holistic fashion, and the role of image complexity and clutter.

We have also introduced a novel image-level feature representation characterized by the distributions of values produced by different saliency models. This provides a feature with strong diagnostic value in predicting IOC. It is also expected that HoPS type features may have value in alternative analysis that considers gaze or perceptual characteristics of images.

The value of features we have considered in a regression model reveal some important characteristics of variability in gaze patterns, and also viewing behaviour in general. While local contrast (or saliency) measures carry significant value in this analysis there are also clearly many other elements beyond context or scene structure that factor into viewing behaviour, and that are also measurable. The value of more semantically grounded features (e.g. objects present) drawn from a deep learning paradigm underscores the value in marrying low-level image characteristics with higher level image understanding. This is important to the analysis of IOC considered in this paper, but also establishes the value of considering both low-level and high-level features in understanding gaze behaviour in general.

Acknowledgements

The authors acknowledge the financial support of the Natural Sciences and Engineering Council Canada Discovery Grants program, the University of Manitoba Graduate Fellowship program, and the University of Manitoba Graduate Enhancement of Tri-Council Stipends program.

References

BORJI, A., SIHITE, D. N., AND ITTI, L. 2013. Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study. *IEEE Transactions on Image Processing 22*, 1, 55–69.

BRUCE, N. D. B., AND TSOTSOS, J. K. 2009. Saliency, attention, and visual search: An information theoretic approach. *Journal of Vision 9*, 3.

CHANG, C. C., AND LIN, C. J. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology 2*, 27:1–27:27.

FELZENSZWALB, P. F., GIRSHICK, R. B., MCALLESTER, D., AND RAMANAN, D. 2010. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence 32*, 9, 1627–1645.

GARCIA-DIAZ, A., FDEZ-VIDAL, X. R., PARDO, X. M., AND DOSIL, R. 2012. Saliency from hierarchical adaptation through

decorrelation and variance normalization. *Image and Vision Computing 30*, 1, 51–64.

HAREL, J., KOCH, C., AND PERONA, P. 2007. Graph-based visual saliency. *Advances in neural information processing systems 19*, 545.

HOSSEIN KHATOONABADI, S., VASCONCELOS, N., BAJIC, I. V., AND SHAN, Y. 2015. How many bits does it take for a stimulus to be salient? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5501–5510.

HOU, X., AND ZHANG, L. 2007. Saliency detection: A spectral residual approach. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*.

HOU, X., AND ZHANG, L. 2009. Dynamic visual attention: Searching for coding length increments. 681–688.

HOU, X., HAREL, J., AND KOCH, C. 2012. Image signature: Highlighting sparse salient regions. *IEEE Transactions on Pattern Analysis and Machine Intelligence 34*, 1, 194–201.

ISOLA, P., XIAO, J., TORRALBA, A., AND OLIVA, A. 2011. What makes an image memorable? *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 145–152.

ITTI, L., KOCH, C., AND NIEBUR, E. 1998. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence 20*, 11, 1254–1259.

JAZAYERI, M., AND MOVSHON, J. A. 2006. Optimal representation of sensory information by neural populations. *Nature Neuroscience 9*, 5, 690–696.

JIA, Y., SHELHAMER, E., DONAHUE, J., KARAYEV, S., LONG, J., GIRSHICK, R., GUADARRAMA, S., AND DARRELL, T. 2014. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*.

JUDD, T., EHINGER, K., DURAND, F., AND TORRALBA, A. 2009. Learning to predict where humans look. In *Computer Vision, 2009 IEEE 12th international conference on*, IEEE, 2106–2113.

JUDD, T., DURAND, F., AND TORRALBA, A. 2011. Fixations on low-resolution images. *Journal of Vision 11*, 4, 1–20.

KANAN, C., TONG, M. H., ZHANG, L., AND COTTRELL, G. W. 2009. Sun: Top-down saliency using natural statistics. *Visual Cognition 17*, 6-7, 979–1003.

KOEHLER, K., GUO, F., ZHANG, S., AND ECKSTEIN, M. P. 2014. What do saliency models predict? *Journal of Vision 14*, 3.

KRIZHEVSKY, A., SUTSKEVER, I., AND HINTON, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, 1097–1105.

LE MEUR, O., AND BACCINO, T. 2013. Methods for comparing scanpaths and saliency maps: Strengths and weaknesses. *Behavior Research Methods 45*, 1, 251–266.

LE MEUR, O., LE CALLET, P., BARBA, D., AND THOREAU, D. 2006. A coherent computational approach to model bottom-up visual attention. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 28*, 5, 802–817.

LE MEUR, O., BACCINO, T., AND ROUMY, A. 2011. Prediction of the inter-observer visual congruency (iovc) and application to image ranking. In *Proceedings of the 19th ACM international conference on Multimedia*, ACM, 373–382.

LOWE, D. G. 2004. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision 60*, 2 (Nov.), 91–110.

MANCAS, M., AND LE MEUR, O. 2013. Memorability of natural scenes: The role of attention. *2013 IEEE International Conference on Image Processing, ICIP 2013 - Proceedings*, 196–200.

MURRAY, N., MARCHESOTTI, L., AND PERRONNIN, F. 2012. Ava: A large-scale database for aesthetic visual analysis. 2408–2415.

OLIVA, A., AND TORRALBA, A. 2001. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision 42*, 3, 145–175.

OLSHAUSEN, B. A., AND FIELD, D. J. 1997. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision Research 37*, 23, 3311 – 3325.

RAHMAN, S., ROCHAN, M., WANG, Y., AND BRUCE, N. D. 2014. Examining visual saliency prediction in naturalistic scenes. In *Image Processing (ICIP), 2014 IEEE International Conference on*, IEEE, 4082–4086.

ROSENHOLTZ, R., LI, Y., AND NAKANO, L. 2007. Measuring visual clutter. *Journal of Vision 7*, 2.

RUSSAKOVSKY, O., DENG, J., SU, H., KRAUSE, J., SATHEESH, S., MA, S., HUANG, Z., KARPATHY, A., KHOSLA, A., BERNSTEIN, M., BERG, A. C., AND FEI-FEI, L., 2014. ImageNet Large Scale Visual Recognition Challenge.

SEO, H. J., AND MILANFAR, P. 2009. Static and space-time visual saliency detection by self-resemblance. *Journal of Vision 9*, 12, 1–27.

SIMONYAN, K., AND ZISSERMAN, A. 2014. Very deep convolutional networks for large-scale image recognition. *CoRR abs/1409.1556*.

TORRALBA, A., OLIVA, A., CASTELHANO, M. S., AND HENDERSON, J. M. 2006. 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.

TORRALBA, A. 2003. Modeling global scene factors in attention. *Journal of the Optical Society of America A: Optics and Image Science, and Vision 20*, 7, 1407–1418.

VAN DER MAATEN, L., AND HINTON, G. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research 9*, 2579-2605, 85.

VIOLA, P., AND JONES, M. J. 2004. Robust real-time face detection. *International journal of computer vision 57*, 2, 137–154.

WAINWRIGHT, M. J. 1999. Visual adaptation as optimal information transmission. *Vision Research 39*, 23, 3960–3974. cited By (since 1996)114.

XIAO, J., HAYS, J., EHINGER, K., OLIVA, A., AND TORRALBA, A. 2010. Sun database: Large-scale scene recognition from abbey to zoo. 3485–3492.

YAN, J., LIU, J., LI, Y., NIU, Z., AND LIU, Y. 2010. Visual saliency detection via rank-sparsity decomposition. *Proceedings of International Conference on Image Processing, ICIP*, 1089–1092.