Predicting Saliency and Aesthetics in Images: A Bottom-up Perspective

A dissertation submitted by **Naila Murray** at Universitat Autonoma de Barcelona to fulfil the degree of **Doctor of Philosophy**.

Bellaterra, November 2012

Director	**Dr. Xavier Otazu** Dept. Ciencies de la Computació & Centre de Visió per Computador Universitat Autonoma de Barcelona
Co-director	**Dr. Maria Vanrell** Dept. Ciencies de la Computació & Centre de Visió per Computador Universitat Autonoma de Barcelona

This document was typeset by the author using LaTeX 2 .

The research described in this book was carried out at the Computer Vision Center, Universitat Autonoma de Barcelona.

ISBN:

Printed by

To my parents, Marlene and Anthony
To my brothers, Khari, Omari, Khafra and Lasana
And to Jose

Acknowledgments

The elaboration and completion of this dissertation would not have been possible without the guidance, support, and encouragement of many people. I am thankful to my adviser Xavier Otazu for his support and guidance during the last four years. I am deeply appreciative of the meticulousness of my second adviser Maria Vanrell, who impressed upon me the importance of clarity and firmness of ideas and expression.

I am also exceeding grateful to my supervisors and collaborators at Xerox Research Centre Europe with whom I worked on a substantial portion of the research presented in this dissertation. In particular, I am indebted to my project supervisor Luca Marchesotti, who helped me learn to look at problems from a broader perspective. I am also grateful to Florent Perronnin for giving generously of his time and his knowledge whenever asked.

The warmth and helpfulness of my fellow students were invaluable to me during my first days in Barcelona and after. I am especially grateful to Jaume Gibert, Pep Gongaus, Albert Gordo, Javier Marin and David Vazquez for their help in navigating a new culture and environment. I am also thankful for the camaraderie of Noha Elfiky, Wenjuan Gong and Hany SalahEldeen.

My wonderful colleagues in the Color in Context group of the Computer Vision Centre at the Universitat Aut0noma de Barcelona were also great sources of support and advice and they deserve my sincerest thanks. I shared many memorable moments in memorable places with Shida Beigpour, Fahad Khan, David Rojas, Eduard Vazquez and Javier Vazquez, who lightened my load with their humour, counsel, and generousity.

I feel extremely fortunate to have met Jose Carlos Rubio during this journey. His positive spirit and outlook and his unflagging and unlimited support means more to me than I can express. Lastly, I cherish and am deeply thankful for the love of my mother Marlene and father Anthony, and my brothers Khari, Omari, Khafra and Lasana. Their affection and encouragement sustained me throughout the past four years.

Abstract

This dissertation investigates two different aspects of how an observer experiences a natural image: (i) where we look, namely, where attention is guided, and (ii) what we like, i.e., whether or not the image is aesthetically pleasing. These two experiences are the subjects of increasing research efforts in computer vision. The ability to predict visual attention has wide applications, from object recognition to marketing. Aesthetic quality prediction is becoming increasingly important for organizing and navigating the ever-expanding volume of visual content available online and elsewhere. Both visual attention and visual aesthetics can be modeled as a consequence of multiple interacting mechanisms, some bottom-up or involuntary, and others top-down or task-driven. In this dissertation a bottom-up perspective is adopted, using low-level visual mechanisms and features, as it is here that the links between aesthetics and attention may be more obvious and/or easily studied.

In Part 1 of the dissertation, it is hypothesized that salient and non-salient image regions can be estimated to be the regions which are enhanced or assimilated in standard low-level color image representations. This hypothesis is proved by adapting a low-level model of color perception into a saliency estimation model. This model shares the three main steps found in many successful models for predicting attention in a scene: convolution with a set of filters, a center-surround mechanism and spatial pooling to construct a saliency map. For such models, integrating spatial information and justifying the choice of various parameter values remain open problems. The proposed saliency model inherits a principled selection of parameters as well as an innate spatial pooling mechanism from the perception model on which it is based. This pooling mechanism has been fitted using psychophysical data acquired in color-luminance setting experiments. The proposed model outperforms the state-of-the-art at the task of predicting eye-fixations from two datasets. After demonstrating the effectiveness of the basic saliency model, an improved image representation is introduced. The improved representation, based on geometrical grouplets, enhances complex low-level visual features such as corners and terminations, and suppresses relatively simpler features such as edges. With this improved image representation, the performance of the proposed saliency model in predicting eye-fixations increases for both datasets.

In Part 2 of the dissertation, the problem of aesthetic visual analysis is investigated. While a great deal of research has been conducted on hand-crafting image descriptors for aesthetics, little attention so far has been dedicated to the collection, annotation and distribution of ground truth data. Because image aesthetics is complex and subjective, existing datasets, which have few images and few annotations, have significant limitations. To address these limitations, a new large-scale database for conducting Aesthetic Visual Analysis is introduced, called AVA. AVA contains more than 250,000 images, along with a rich variety of annotations. Ways in which the wealth of data in AVA can be used to tackle the challenge of understanding and assessing visual aesthetics is investigated by looking into several problems relevant for aesthetic analysis. It is demonstrated that by leveraging the data in AVA, and using generic low-level features such as SIFT and color histograms, one can exceed state-of-the-art performance in aesthetic quality prediction tasks.

Finally, the hypothesis that low-level visual information in the proposed saliency model can also be used to predict visual aesthetics is entertained. This low-level information captures local image

characteristics such as feature contrast, grouping and isolation, characteristics thought to be related to universal aesthetic laws. The weighted center-surround responses that form the basis of the saliency model are used to create a feature vector that describes aesthetics. In addition, a novel color space for fine-grained color representation is introduced. It is then demonstrated that the resultant features achieve state-of-the-art performance on aesthetic quality classification.

As such, a promising contribution of this dissertation is to show that several vision experiences - low-level color perception, visual saliency and visual aesthetics estimation - may be successfully modeled using a unified framework. This suggests a similar architecture in area V1 for both color perception and saliency and adds evidence to the hypothesis that visual aesthetics appreciation is driven in part by low-level cues.

Resumen

Esta tesis investiga dos aspectos diferentes sobre cómo un observador percibe una imagen natural: (i) dónde miramos o, concretamente, qué nos atrae la atención, y (ii) qué nos gusta, e.g., si una imagen es estéticamente agradable, o no. Estas dos experiencias son objeto de crecientes esfuerzos de la investigación en visión por computador. La habilidad de predecir la atención visual tiene muchas aplicaciones, desde el reconocimiento de objetos a el márketing. La predicción de la calidad estética también ha visto aumentada su importancia, sobre todo para la organización y navegación del contenido visual online, cuyo volumen se encuentra constantemente en expansión.

Tanto la atención visual como la estética visual pueden ser modeladas como consecuencia de múltiples mecanismos en interacción, algunos bottom-up o involuntarios, y otros top-down o guiados por tareas. En este trabajo nos concentramos en una perspectiva bottom-up, usando mecanismos visuales y caracter´sticas de bajo nivel, ya que es aqu´ donde los v´nculos entre estética y atención son más evidentes, o fácilmente analizables.

En la Parte 1 de la tesis presentamos la hipótesis de que las regiones en una imagen que atraen o no la atención pueden ser estimadas usando representaciones estándar de bajo nivel de imágenes en color. Demostramos esta hipótesis usando un modelo de percepción de color de bajo nivel y adaptándolo a un modelo de estimación de la atención. Este modelo comparte los tres pasos principales encontrados en muchos de los modelos que han sido satisfactorios para predecir la atención en una escena: convolución de un conjunto de filtros, un mecanismo center-surround, y el spatial pooling para construir un mapa de la atención. Para estos modelos, integrar la información espacial y justificar el valor de varios parámetros son problemas que todav´a se mantienen abiertos. Nuestro modelo de atención hereda una selección de parámetros y un mecanismo de spatial pooling de los modelos de percepción en los que está basado. éste mecanismo de pooling ha sido ajustado usando datos psicof´sicos adquiridos a través de experimentos sobre color y luminancia. El modelo propuesto mejora el estado-del-arte en la tarea de predecir los puntos de atención en dos bases de datos. Tras demostrar la efectividad de nuestro modelo básico de atención, introducimos una representación de la imagen mejorada, basada en conjuntos geométricos. Esta representación realza las caracter´sticas visuales de bajo nivel más complejas, como son las esquinas y terminaciones, y suprime otras caracter´sticas relativamente más sencillas, como los bordes. Con esta mejorada representación de imágenes, el rendimiento de nuestro modelo de atención mejora en las dos bases de datos.

En la Parte 2 de la tesis, investigamos el problema del análisis estético visual. Mientras la mayor´a de investigación se ha llevado a cabo creando descriptores estéticos de forma manual, ha sido poca la atención dedicada a la colección, anotación y distribución de datos de ground-truth. Debido a que la estética de imágenes es algo complejo y subjetivo, las bases de datos existentes, que proveen unas pocas imágenes y anotaciones, tienen importantes limitaciones. Para tratar estas limitaciones, hemos presentado una base de datos a gran escala para llevar a cabo actividades de análisis estético visual, que llamamos AVA. AVA contiene más de 250,000 imágenes, junto con una rica variedad de anotaciones. Hemos investigado cómo la riqueza de los datos en AVA puede ser usada para abordar el dif´cil problema de entender y evaluar la estética visual, en el contexto de diversos problemas relevantes para el análisis

estético. Hemos demostrado que aprovechando los datos en AVA, y usando caracter´sticas genéricas de bajo nivel, como SIFT e histogramas de color, podemos superar el estado-del-arte en tareas de predicción de la calidad estética.

Finalmente, consideramos la hipótesis de que la información visual de bajo nivel en nuestro modelo de atención puede también ser usada para predecir la estética visual. Para ello, capturamos las caracter´sticas locales de la imagen como contraste, agrupaciones y aislamiento de caracter´sticas, que se suponen relacionadas con reglas universales de la estética. Usamos las respuestas del centre-surround que forman la base de nuestro modelo de atención, para crear un vector de caracter´sticas que describe la estética. También introducimos un nuevo espacio de color, para representaciones de grano fino. Para terminar, demostramos que las caracter´sticas resultantes alcanzan la precisión del estado-del-arte en el problema de clasificación de la calidad estética.

Una contribución prometedora de esta tesis es demostrar que diversas experiencias de la visión - percepción de color a bajo nivel, atención visual, y estimación de la estética visual - pueden ser satisfactoriamente modeladas usando un marco de trabajo unificado. Esto sugiere una arquitectura similar en el área V1 del cerebro para la percepción del color y la atención, y añade evidencias a la hipótesis que la apreciación estética está influenciada, en parte, por información de bajo nivel.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The viewing of a visual scene may elicit a variety of reactions in a human observer. One region of the scene may attract focused attention while large regions are completely ignored. The scene may elicit pleasant emotions or feelings of revulsion. It may make a lasting and memorable impression on the observer or may never again be recalled. It seems reasonable to hypothesize that some of these reactions, for example the attention we give to a visual stimulus and our ability to recall having seen that stimulus, may share similar or even common perceptual mechanisms.

The mechanisms that give rise to these reactions and impressions in human observers are so multitudinous and interconnected that discovering and deciphering them may seem an impossible task. And yet, researchers in fields as wide-ranging as psychology, machine learning, art history, neuroscience and computer vision have been, independently and in collaboration, expanding our knowledge about the reasons why we attend, ignore, enjoy or dislike some and not other visual stimuli.

These reasons are related to factors which may vary greatly across individuals, such as emotional states and educational history. For example, when observing artwork, those with a formal education on the subject have a very different pattern of visual attention than do those without formal training [71,92, 124]. Due to their inherently subjective and variable nature, it is difficult to study visual experience by analysing such factors. However, visual experience is also a product of mechanisms which *vary much less across individuals* and are more easily understood. Such mechanisms are involved in the perception of relatively objective characteristics of the elements of a scene such as spatial frequency, orientation and color.

Numerous brain regions participate in perceiving the subjective and objective visual characteristics that ultimately lead to an experience such as attention, aesthetic appreciation or image memorization. These brain areas process information from a variety of sources. Visual attention for instance engages the visual cortex, which processes visual information [53], the inferior temporal cortex, which accesses memory [30], in addition to many other areas. Aesthetic appreciation is a function of, among other factors, perception of form and content in the visual cortex, and emotional responses, processed in areas such as the anterior medial temporal lobe and the orbito-frontal cortex [19].

However, while the information sources involved are diverse, data captured by the retinae must necessarily play a critical role in each type of visual experience. In the human visual system, this data is transmitted to higher cortical areas almost entirely via the primary visual cortex or area V1. As such, the retina, intermediary areas, and eventually the primary visual cortex, form a common visual front-end [115] for various visual experiences. This common visual front-end reminds one of the previously mentioned hypothesis of shared perceptual mechanisms. An obvious question then arises: are different visual experiences determined, to some significant and measurable degree, by common perceptual mechanisms found in the visual front-end?

1.1 Motivation

The existence of common mechanisms in the visual front-end that directly affect different visual experiences is an intriguing hypothesis as it would allow the experiences to be (partially) explained in a unified framework. In spite of this, to my knowledge there have been no works that explicitly test this hypothesis by using a generic computational model of low-level vision to predict visual phenomena and quantitatively evaluating its performance.

This dissertation presents the attempt to do just that, by adapting a state-of-the-art computational model of low-level color perception [93, 94] and applying its modified version to different visual tasks. This color perception model follows the standard architecture of the visual front-end and is thus a good candidate for testing this hypothesis. Two different aspects of how an observer experiences a natural image are investigated in this dissertation:

- where we look, that is, where attention is guided. In particular, we develop a bottom-up visual attention model which predicts the eye-fixations of observers who were given a free-viewing task.
- what we like, that is, whether or not the image is aesthetically pleasing. Here, we develop a model of aesthetics which we then use to predict human annotations.

These two experiences are the subjects of increasing research efforts in computer vision. The ability to predict visual attention has wide applications, from object recognition to marketing. Aesthetic quality prediction is becoming increasingly important for organizing and navigating the ever-expanding volume of visual content available online and elsewhere.

Different dimensions of visual experience, including color perception, visual attention, and visual aesthetics appreciation, are widely understood as having two types of interacting mechanisms: those that are "top-down", and those that are "bottom-up". So-called top-down components are thought to be cognitive processes that may be knowledge, memory, or task-guided. These correspond to the individualistic or subjective components of visual experience mentioned previously. Bottom-up components correspond to the more objective visual percepts described earlier. Such components involve low-level visual mechanisms and features, and are driven by data received through the retinae.

Here, the term "low-level" is used in the sense explained by Sukuzi *et al.* [113]: low-level mechanisms refer to mechanisms used in the early stages of visual processing while low-level features are those image features thought to be processed at these stages. Bottom-up or low-level vision processes are found in the visual front-end and, as mentioned previously, are more extensively studied and understood than the more elusive top-down mechanisms. For this reason a bottom-up perspective is adopted in this work, as it is here that the links between color perception, visual attention and visual aesthetics may be more obvious and/or more easily studied.

1.2 Contributions

The major contribution of this dissertation is to show that several visual experiences - low-level color perception, visual saliency and visual aesthetics estimation - may be successfully modeled using a unified framework. This unified framework is based on a model of color perception which has been shown to successfully reproduce several visual illusions related to color and brightness induction phenomena.

The first step was to fit the parameters of the color perception model. These parameters are fit using data obtained from psychophysical experiments related to brightness and color induction [88].

Slight adaptations to this model are then made and the resulting saliency model is used to predict eye-fixations of observers viewing images of natural scenes [88]. Although the visual stimuli used to fit the model parameters are quite different to those typical of natural scenes, the adapted model, which has been termed SIM (Saliency by Induction Mechanisms), outperforms state-of-the-art saliency models at predicting eye-fixations. Moreover, the psychophysically-tuned parameters are shown to be

optimal for both eye-fixation prediction and color perception modeling. This indeed suggests a similar architecture in area V1 for both color perception and saliency. In addition, because the model inherits a principled selection of parameters and an innate spatial pooling mechanism from the color perception model on which it is based, it addresses key criticisms of and unresolved issues with biologically-inspired saliency estimation models. The main criticisms are that (i) such models are difficult to tune owing to their myriad parameters; and (ii) such models do not have a principled manner of pooling information gleaned across different spatial scales.

SIM was highly responsive to edges as well as more complex features created by superpositions of edges, such as corners and junctions. However, complex features have been shown to be preferentially fixated upon in comparison to simpler features. Therefore, an image representation for which the response amplitudes of complex features are enhanced relative to simpler features such as edges was desirable. To this end an image decomposition termed the grouplet transform, which was originally used for image de-noising, was incorporated into the proposed saliency model. This image representation essentially extends the extent of the region over which spatial competition occurs for each local feature response. This new representation had the desired effect of enhancing complex features [89].

After developing the SIM model, the subject of image aesthetics was studied in a computational framework. Computational modeling of image aesthetics is a nascent research field and not as well studied as visual attention. Most research efforts to date have focused on designing features that correlate with techniques used by professional photographers for capturing high-quality photographs. Because such models are overwhelmingly trained in a supervised learning framework, rich and diverse training images and annotations are critical to the success of such models, moreover because aesthetics itself is a multi-faceted concept without a single interpretation. However, as this is a new area of research, there is a dearth of robust and diverse datasets for training, evaluation and analysis of computational models of aesthetics. To address this issue the next contribution was made: the assembly and in-depth analysis of a large-scale database for image aesthetics analysis, which has been named AVA [86, 87]. AVA contains over 200,000 images, with hundreds of score annotations each. These score annotations form score distributions over a rating scale, allowing one to gain an idea of the degree of consensus among users. In addition, the images have many associated textual comments given by annotators, providing detailed feedback on an image's aesthetic characteristics and attributes.

In [85–87], it was demonstrated, through several applications, how the large scale and diverse annotations of AVA can be leveraged to improve performance on existing preference tasks and inspire new ones. In particular, models were trained to perform binary classification into "high-quality" and "low-quality" aesthetic categories, to perform aesthetic score prediction, and to perform image ranking. It was shown that the large scale of training data in AVA enabled significant improvement in model training. It was also shown that by judiciously selecting training images from among those in AVA, one can preserve model performance even when fewer training images are used.

At this stage, armed with a suitable dataset and baseline methods, we returned to the central theme of the : the plausibility of using a common low-level vision model to predict different complex visual experiences. We again made slight adaptations to the color perception model and were able to extract image features which can predict aesthetics labels given to images by human annotators. The extracted features perform at a state-of-the-art level when compared with features extracted using procedures that have been hand-crafted especially for aesthetics and also when compared with sophisticated generic low-level visual features. We believe that this is because low-level visual features in our saliency model capture local image characteristics such as feature contrast, grouping and isolation, characteristics thought to be related to universal aesthetic laws.

Thus, our saliency model and aesthetics features, both of which have been directly derived from a model of low-level color perception, achieve state-of-the-art performance on related predictive tasks. Their success adds evidence to the hypothesis that color perception, bottom-up visual attention and visual aesthetics appreciation are driven in significant part by cell responses from a common neural substrate in the early human visual system.

1.3 Organization

The is organized into three parts.

The topic of part 1 of the is visual saliency. A brief introduction to the field is given in chapter 2, situating it first within the wider scope of visual attention, and then paying particular attention to bottom-up visual attention or saliency. Seminal works in computational models are described and the common components and limitations among the approaches are described in detail, as the components are also shared with our proposed low-level models. A basic understanding of the architecture and known properties of the human visual system is assumed. In chapter 3 we validate our hypothesis on the relationship between low-level color perception and visual saliency. We describe in detail the implementation of our saliency models and describe experimental results which demonstrate its state-of-the-art performance at predicting eye-fixations on two datasets. After demonstrating the effectiveness of our basic saliency model we introduce, in chapter 4, the improved image representation, based on geometrical grouplets. We describe how the image representation is constructed using a modified Haar wavelet transform, and we show through quantitative evaulations that, with this improved image representation, the performance of our saliency model in predicting eye-fixations increases for both datasets.

In part 2 of the , we investigate the problem of image aesthetic analysis. In chapter 5 we describe the state of the field, focusing on computational methods for learning models of image aesthetics. We discuss current state-of-the-art aesthetics features and the popular paradigms for learning aesthetic models. We describe our database for image aesthetics analysis in chapter 6. We explain the provenance of the data and we discuss the context in which the aesthetics and other annotations were made. We also compare AVA to other existing image aesthetics databases. In chapter 7 we investigate how the wealth of data in AVA can be used to tackle the challenge of understanding and assessing visual aesthetics by looking into several problems relevant for aesthetic analysis, including binary classification into “high-quality” and “low-quality” categories, aesthetic score prediction, and image ranking.

Finally, in part 3 we investigate the hypothesis that low-level visual features in our saliency model are informative about the aesthetic characteristics of images. In chapter 8 we explain our aesthetic feature extraction process and our novel color space representation. We also provide extensive quantitative evaluation of the proposed features. Conclusions and future directions of research in the work presented in this are described in chapter 9.

Part I

Visual Saliency

Chapter 2

A Brief Review of Visual Saliency Modeling

Although many factors may determine what image features are selected or discarded by our attentional processes, it has been useful to separate these into two categories of processes: top-down and bottom-up [116]. Top-down processes are dependent on the organism's internal state and are often task-driven, so that the areas of a scene to which attention is given varies as a function of the motivation for viewing the scene. Therefore, if the organism is searching for a specific object, its attention will be guided to different scene elements than would be the case were it simply navigating its environment. Bottom-up processes on the other hand, comprise unconscious and instantaneous processes, usually thought to be driven by data captured by the retinae and relayed through the lateral geniculate nuclei to the early stages of the human visual system. Bottom-up visual attention, termed saliency, may be thought of as visual attention in the absence of conflicting top-down cues.

In his seminal work on cognitive psychology, Neisser championed the now widely-accepted view of visual perception as resulting from an interplay between bottom-up and top-down factors [90]. Computational models of visual attention based on this view have proliferated in fields of vision-related research, including cognitive psychology, computational neuroscience and biological and computer vision. The majority of these works are of mostly theoretical interest and have only been tested on synthetic visual stimuli. Such works are out of the scope of this discussion.

Models for predicting visual attention towards a natural scene typically make these predictions in the form of a topographical map of the scene. This map charts the degree to which each location in the scene is likely to attract visual attention. Such maps are termed visual attention maps if they are computed in part or in whole by using top-down mechanisms. They are termed saliency maps when only bottom-up mechanisms are used in their computation [12, 15, 34, 39, 53, 60, 63, 135]. As this dissertation takes a bottom-up perspective, we center our discussion on visual saliency modeling.

2.1 Visual Saliency Modeling

A good working definition of saliency is that given by Koch & Ullman [65]:

> "Saliency at a given location is determined primarily by how different this location is from its surround in color, orientation, motion, depth, etc."

The "feature-integration theory of attention" of Treisman & Gelade [117] advocated what has become the dominant paradigm for modeling saliency. This theory holds that low-level features (or dimensions in their terminology) such as color, orientation and motion, are processed in parallel by the

visual system before being integrated into a "master map" using some attentional mechanism. Koch & Ullman [65] proposed an attentional framework in which these features are encoded in separate topographical, cortical maps which preserve their spatial relationships. These maps would exist at different spatial frequencies, reflecting the evidence for multiple spatial frequency channels [13, 123], as well as at different feature values. This means that to represent color for example, red features would be encoded in a separate map to blue features. In the proposed framework, the information encoded in the different elementary feature maps are combined into what Koch & Ullman coined the "saliency map", a topographical map of the conspicuity at each location of the visual scene. This saliency map was hypothesized to be located in the early visual system, perhaps in the lateral geniculate nucleus or the primary visual cortex (indeed, more recent work by Li Zhaoping suggests that the outputs of area V1 constitute a saliency map [73], in that V1 cells fire more rapidly when their receptive fields contain salient features to which they are tuned). The saliency map locations with the highest elevations would be the located to which visual attention was guided.

The first implementation of a saliency model in the conceptual framework proposed by Koch & Ullman is that of Niebur & Koch [91]. In this model, maps of different features, for multiple spatial frequencies, were generated using Gaussian pyramids. Center-surround operations were performed on these channels in order to mimic the receptive field properties of cortical cells. Specifically, the value of a pixel in a given location of a feature map was treated as the response of the center of a receptive field, while the pixel value in the corresponding location of the feature map at a lower spatial frequency was treated as the surround. By comparing the center and surround values, by for example subtracting them, local feature contrast, or conspicuity, was estimated for that feature value at that spatial frequency. The contrast information across different features and spatial frequencies was pooled additively, using identical weights.

The model of Itti *et al.* [54] follows in the vein of that of Nieber & Koch and has become one of the most influential models in computer vision. It uses a neural network to output a saliency map after training the network with center-surround excitation responses of feature maps obtained after a single layer of linear filters are applied to the input image. Each feature map contains information from one of three cues: orientation, color or scale. This model has been deployed in many practical applications including video summarization, image compression, and designing advertising materials.

Saliency Map Evaluation

A natural approach to evaluating a saliency map is to compare its predictions of salient image locations to the behavior of human observers when viewing the image. However, a non-trivial question arises: what are the behavioral correlates of bottom-up visual attention to which saliency maps may be compared?

One such correlate is reaction time (RT) when performing visual search tasks. In such tasks, observers are instructed to locate a target feature among several distractor features. The RT of the observer is the time taken to locate the target. This is typically measured as the time interval between the beginning of the search (when the experimenter indicates to the observer to begin and simultaneously displays the search array for example) and the response of the observer (for example by pressing a button on a keyboard or game pad). The assumption here is that more salient targets will have a short reaction time as compared to less salient targets. When used as a correlate of attention, RT has conventionally been measured in visual search experiments involving synthetic visual stimuli arranged into what is termed a search array. In these arrays, the target is designed to be salient and "pop-out" at the observer from among the distractors (see Figure 2.1 for an representative example). Unfortunately, RT is a poor saliency correlate when visual search involves familiar targets and natural scenes. This is because many top-down processes such as memory and prior experience may be engaged [20]. For example, if tasked with locating a deer in an image of a landscape, the observer is more likely to attend first to ground regions rather than sky regions, due to prior knowledge that a deer is unlikely to found in the sky. A

further draw-back of RT is that it includes both the time taken to locate the target and the time taken for response execution (i.e. pressing the button).

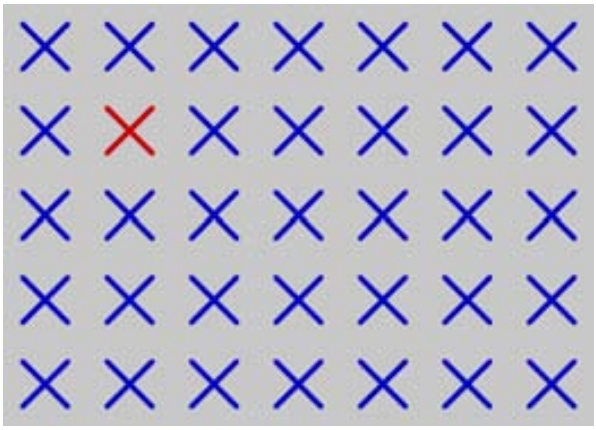
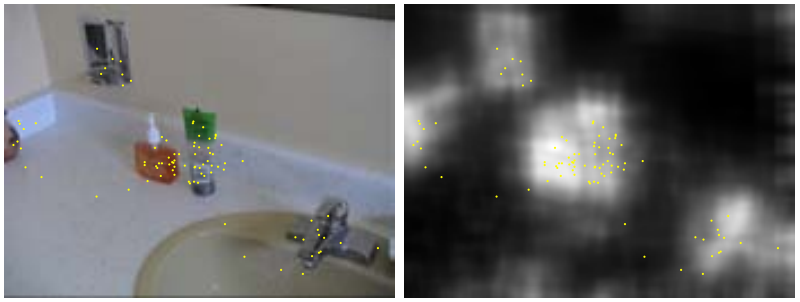


Figure 2.1: A typical search array for investigating color saliency. The target red cross should be more salient than the distractor blue crosses.

Another, more widely-used, behavioral correlate is eye fixation. Saccadic eye movements, perhaps one of the most defining characteristics of the human visual system, allow us to rapidly sample images by changing the point of fixation. Eye-fixations are guided by both top-down and bottom-up visual processes, and there is a decided lack of consensus about the quantitative proportions in which these two processing modalities contribute. However, various studies [37, 52, 96] have shown that there is some contribution, and that this contribution is stronger in the absence of task-driven cues. As such, when eye-fixations are to be used as correlates of saliency, observers are typically instructed to study images, but are not given a specific task. An eye fixation may refer to the movement of the eye to re-orient the fovea, but here we view an eye-fixation as the point between two saccades, in which the eye is relatively motionless [64]. The most widely-used methods for recording eye-fixation coordinates and duration is eye-tracking technology (an accessible guide to which may be found in [32]).

An example of an image, associated eye-fixations and an estimated saliency map is shown in Figure 2.2. Now that such pairs of predictions and behavioral correlates can be made, how are they compared? For fixations, several popular procedures exist. In one, the saliency map values at fixation locations may be used to form a probability distribution which is then compared to the probability distribution of saliency map values sampled randomly from the same or a different saliency map using the Kullback-Leibler distance. The saliency map may also be used to classify image locations into fixated and non-fixated categories, after which the area under the ROC curve is computed. In another procedure, the fixations are used to create a saliency map, using for example a kernel density estimator and the correlation between that map and the model predicted one is computed. Further details on these and several other evaluation procedures are discussed in detail in [8].

Computing saliency maps is still an open problem whose interest is growing in computer vision [8, 9]. Many models are inspired in major part by the computational framework of Niebur & Koch [91] (and eventually Itti *et al.* [54]), and contain common stages as a result. In this dissertation, we will explore saliency map estimation using these common stages, which form what we term the general biologically-inspired bottom-up framework. We describe this framework in the next section.



(a) Input Image (b) Estimated Saliency Map

Figure 2.2: An example of the saliency map for an image (yellow dots indicate eye-fixations). In the saliency map, greater lightness indicates higher saliency.

2.2 General biologically-inspired bottom-up framework

The general biologically-inspired bottom-up framework mimics the standard architecture of cortical area V1. In V1, mutually suppressive interactions between cortical cells, competing for representation in later stages of the visual pathway, begin in earnest [53, 118]. As a result of these suppressive interactions, the stimulus regions in the receptive fields of the cells with the least suppression, or the most facilitation, are eventually fixated upon [53]. The locations of such features of a visual scene correspond to the peak locations in the saliency map of that scene.

The first stage in this common framework involves representing the image in an opponent-color space. Next, a scale-space decomposition of the input image is performed using a set of linear filters. This is followed by a center-surround operation over the decomposition, after which spatial pooling is performed to build the final saliency map. Each of these stages is described next.

2.2.1 Color-space representation

Inspired by color-opponent cells in the lateral geniculate nucleus and cortical area V1, many saliency models choose to represent images in a color-opponent space. This space has three components: red-green or $O1$, yellow-blue or $O2$ and intensity or $O3$. Many manners of computing these three components have been suggested [29, 50, 79, 79]. Among the simplest is the following:

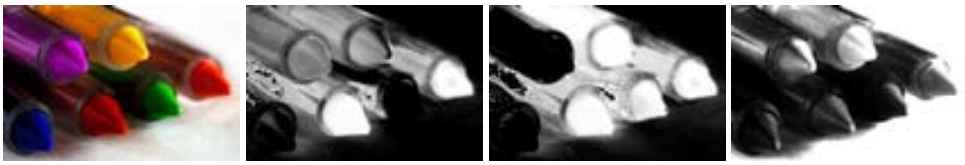
$$O1 = \frac{R-G}{R+G+B}, O2 = \frac{R+G-2B}{R+G+B}, \text{and } O3 = R+G+B \tag{2.1}$$

, where R, G, and B are the familiar red, green and blue color components. The chromatic channels $O1$ and $O2$ have both been normalized by the intensity channel $O3$.

Lab space [50] is a popular color-opponent space for saliency modeling, as it was designed to be more perceptually uniform than existing color spaces. Here, perceptual uniformity signifies that the distance between two colors represented in Lab space should be fairly proportional to the perceived difference between the two colors.

2.2.2 Multi-resolution decomposition

After feature channels containing color, orientation or intensity information are obtained, a multi-resolution decomposition is performed on each channel in order to extract edge information at different



(a) image (b) red-green channel ($O1$) (c) yellow-blue channel ($O2$) (d) intensity channel ($O3$)

Figure 2.3: A simple color-opponent space image representation.

spatial frequencies. There are several popular techniques for doing so, each of which uses a cascade of linear filters. Such filters are Laplacian of Gaussians (LoG) filters, the related Difference of Gaussians (DoG) filters and Gabor-like wavelet basis functions. Filters of these types have become canonical in vision literature for modeling the receptive fields of simple cells in area V1. The response to such filters are consequently used to model the response of such cells to visual stimuli within their receptive fields. The cascade of filters results in multiple image "subbands" which enhance structural information such as edges, ridges and blobs, features popular in works aligned with feature-integration theory.

Laplacian and Difference of Gaussians

A 2-D Laplacian of Gaussian operation over an image gives an isotropic measure of the 2nd-order spatial derivative of that image. It is often approximated using the difference of two isotropic Gaussians, as in the work of David Lowe on keypoint detection [75]. To create a multi-resolution image decomposition using DoG filters, a spatial pyramid of blurred images is first created using a cascade of two-dimensional Gaussian filters. The 2-D Gaussian filter is often decomposed into two 1-D filters, using the separability property of Gaussians, in order to increase computational efficiency in the convolution step. A 1-D Gaussian $G(x\ \sigma)$ may be defined as

$$G(x\ \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-x^2}{2\sigma^2}} \tag{2.2}$$

The image $I(x\ y)$ is succesively blurred by a Gaussian function such that the content of each blurred image, $B(x\ y\ \sigma) = G(x\ y\ \sigma) * I(x\ y)$, differs in scale by a factor $k = 2^{(1/S)}$, where S is the number of scales in each octave. For an initial blurring σ_0, when $k = 2\sigma_o$, the blurred image is down-sampled. Because the image has been passed through a low-pass filter (the Gaussian filter) before down-sampling (resampling at at half the original rate), the resulting decimation ensures no aliasing, and no introduction of new, false structures in the down-sampled image. The decimation increases the efficiency of the algorithm, as the number of elements in the image signal decreases by a factor of two when traversing the cascade of filters.

The scale space can therefore be defined as follows:

$$\sigma(o\ s) = \sigma_o 2^{(o+s/S)} \qquad o = 0\ \cdots\ O-1 \qquad s = 0\ \cdots\ S-1 \tag{2.3}$$

where o is the octave index, s is the scale index and O is the number of octaves created (1 + number of decimations). Because a cascade of Gaussians is being used, each sucessive blurred image $B(x\ y\ \sigma_{s+1})$ is created by convolving the previous blurred image, $B(x\ y\ \sigma_s)$ with a Gaussian

$$G(x\ \sigma_s\ \sigma_{s+1}) = \frac{1}{\sqrt{2\pi(\sigma_{s+1}^2 - \sigma_s^2)}} e^{\frac{-x}{2(\sigma_{s+1}^2 - \sigma_s^2)}} \tag{2.4}$$

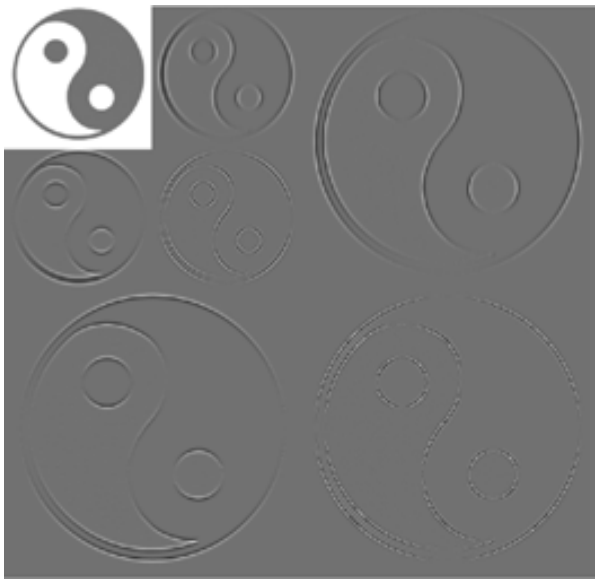


Figure 2.4: Decomposition of image into horizontal, vertical and diagonal wavelet planes for two spatial scales. Light and dark areas of the wavelet planes have high absolute responses to the wavelet kernel.

taking into account the fact that:

$$G(x\ \sigma_{s+1}) * (G(x\ \sigma_s) * I(x\ y)) = G(x\ \sigma_{s+1} + \sigma_s) * I(x\ y) \tag{2.5}$$

so that:

$$\begin{aligned} B(x\ y\ \sigma_{s+1}) &= G(x\ \sigma_{s+1}) * (G(x\ \sigma_s) * I(x\ y)) \\ &= \frac{1}{\overline{2\pi}\sigma_{s+1}} e^{\frac{-x^2}{2\sigma_{s+1}^2}} * I(x\ y) \end{aligned}$$

To create the DoG pyramid, each sucessive blurred image $B(x\ y\ \sigma_{s+1})$ is subtracted from the previous blurred image, $B(x\ y\ \sigma_s)$. Therefore:

$$D(x\ y\ \sigma_s) = B(x\ y\ \sigma_{s+1}) - B(x\ y\ \sigma_s) \tag{2.6}$$

As such, there are one less DoG images than blurred images.

Discrete wavelet transform

When a discrete wavelet transform (DWT) is applied to an image, it is decomposed into a series of new image subbands, termed wavelet planes, with respect to spatial scale s and orientation o (vertical, horizontal and diagonal) [2]. The wavelet planes, w_s^h, w_s^v and w_s^d, contain the response of the image intensities at that orientation to the wavelet kernel corresponding to the scale, s. Figure 2.4 illustrates one such multi-resolution wavelet decomposition. One can see that the variations of the image in different orientations and scale are captured in different wavelet planes. Image decompositions based on wavelet decompositions with Gabor-like basis functions are often used in biologicallly-inspired models of low-level vision as they are well-suited to representing parvo-cellular spatial frequency channels and cortical orientation-selective receptive fields in the HVS [72].

2.2.3 Center-surround response

The center-surround response at a location is at the heart of saliency modeling and is a measure of the degree to which the features at the location are conspicuous or distinctive with respect to those in its surrounding environment. This surround can be along a spatial frequency dimension or the 2-D space defined over an x-y plane. Itti *et al.* [55] proposed to model the center-surround response at a location and spatial scale as the difference between values at that location and the corresponding location at the next finest spatial scale. As such, the surround in this case is at a different spatial frequency. Approaches which measure local center-surround responses within an x-y plane tend to define and compare a local central region and a surrounding region. The central region is typically defined to be circular with a concentric surround annular ring, as illustrated in Figure 2.5. In this case, the center-surround response are calculated by comparing the values lying within the center region to the values lying with the surround region. This comparison may be performed by a divisive normalization of the mean of the center values by the surround values, or by measuring statistical differences between the values in the central region and the surround region.

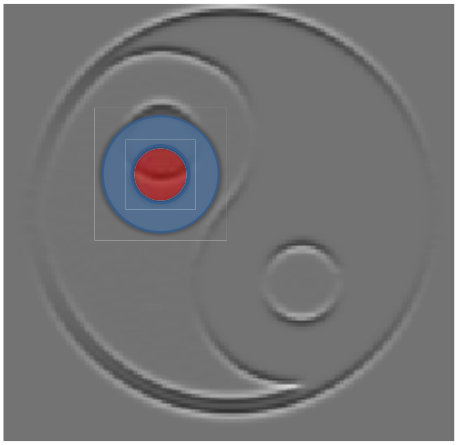


Figure 2.5: Center and surround spatial regions in a wavelet plane, defined by a circle (in red) and a concentric annular ring (in blue) respectively.

2.2.4 Spatial pooling

Once center-surround responses are obtained for each x-y location in each image subband, they must be pooled in order to form a single saliency map of the input image. The pooling is performed typically by linear (weighted or unweighted) summation, or by summation after exponentiation. For image decomposition which involved successive decimations of the image signal, the subbands are interpolated where necessary.

2.3 Saliency estimation in the recent literature

There is a wide spectrum of approaches for modeling visual attention [8] in static scenes, from data-driven methods to biologically-inspired ones. When modeling top-down factors, the difficulties of understanding internal states are usually dealt with by machine learning techniques trained on general

prior knowledge. Bottom-up factors may be incorporated into saliency models by using machine learning techniques or by deriving inspiration from models of low-level vision mechanisms in the human visual system (HVS). As our work deals with saliency modeling, we focus our review on saliency estimation paradigms, for static scenes, that are related to bottom-up factors. An extensive review of saliency estimation and salient object detection may be found in [8].

A typical data-driven method is that of Keinzle *et al.* [63], who sampled small image patches at eye-fixation locations and learned which of these patches classify fixation locations well, by learning patch weights with a support vector machine (SVM). The result method has few free parameters, in contrast with most biologically-inspired models. Their resulting system maximally exitatory stimuli had a center-surround structure, in agreement with several other works [49]. The model of Judd *et al.* [60] combined the information contained in different saliency methods to produce a single saliency map, by using an SVM. High-level information, such as the presence of people and cars in images, were also incorporated in the form of binary maps with non-zero values in detection bounding boxes. Feature vectors for training were constructed by sampling each saliency map at fixation locations and concatenating the values at these locations. The common thread in these works is the use of eye-fixation data for training the models, and formulating saliency estimation as a classification problem. Therefore background or non-fixated regions were also sampled in order to provide negative training examples for the SVM. In all, about 24,000 training samples were used in Kienzle *et al.*and 18,060 samples were used in Judd *et al.*.

The more bio-inspired models of saliency are often based on spatial contrast or information-theoretic formulations. Gao *et al.* [39] considered the saliency of a local region to be quantified by the discriminatory power of a set of features describing that region to distinguish the region from its surrounding context. Bruce & Tsotsos [12] approached local saliency as the self-information of local patches with respect to its surrounding patches, where the surround could be considered a localized surround region or the remainder of the entire image. In [12], an ICA basis set of filters was learned from RGB patches extracted from images and used to represent the local patches. As was also found by Hou & Zhang [49] in a similar approach, the basis set consisted mainly of oriented Gabor-like patches with opponent color properties. Zhang *et al.* [135] also proposed a method which uses self-information, but in this case a spatial pyramid was used to produce local features and a database of natural images, rather than a local neighborhood of pixels or a single image, provided contextual statistics. In addition, Zhang *et al.* extracted features from a spatial pyramid of each of the three opponent color channels. Seo & Milanfar [106] used kernel regression-based self-resemblance to compute saliency, and considered a region to be salient when its curvature was different to that of its surround. Perhaps the most similar model to ours is that of Le Meur *et al.* [83]. This model is based on the early HVS, and models phenomena such as selective contrast sensitivity and visual masking.

2.4 Open questions in the general bottom-up framework

The above-mentioned biologically-inspired methods all follow the general biologically-inspired bottom-up framework to a high degree and have been quite successful models of attention. However, several questions at the core of this framework remain unresolved:

- Which are the optimal feature maps for estimating saliency and how should they be generated? It is unclear whether the filter profiles, color spaces, orientations and other parameters currently used to create feature maps are optimal [63].
- How can the saliency information contained in these feature maps, which have been extracted from multiple scales, orientations, etc., be holistically combined? Current methods either perform linear un-weighted [53] or weighted [136] summations over the maps. Linear weighting is ad-hoc and weights learned with machine-learning introduce additional parameters to the model which must be tuned.

- How can parameters related to model components such as the center-surround mechanisms and non-linear normalizations be fitted in a principled manner? [100].

In chapters 3 and 4, we address the above questions by adapting a low-level model of color perception for the problem of saliency estimation.

Chapter 3

Saliency Estimation Using a Low-Level Color Perception Model

In this chapter, we propose a computational model of saliency that follows the typical three-step architecture described in section 2.2, while trying to address its limitations through a combination of simple, neurally-plausible mechanisms that remove nearly all arbitrary variables. Our proposal in this paper generalizes a particular low-level model developed to predict color appearance [94] and has three main levels:

In the rst stage, the visual stimuli are processed in a manner consistent with what is known about the early human visual pathway (color-opponent and luminance channels, followed by a multi-scale decomposition). The bank of filters used (Gabor-like wavelets) and the range of spatial scales (in octaves) are biologically justified [6, 122, 131] and commonly used in low-level vision modelling.

The second stage of our model consists of a simulation of the inhibition mechanisms present in cells of the visual cortex, which effectively normalize their response to stimulus contrast. The sizes of the central and normalizing surround windows were learned by training a Gaussian Mixture Model (GMM) on eye-fixation data.

The third stage of our model integrates information at multiple scales by performing an inverse wavelet transform directly on weights computed from the non-linearization of the cortical outputs. This non-linear integration is done through a weighting function similar to that proposed by Otazu *et al.* [94] and named *Extended Contrast Sensitivity Function* ($ECSF$), but optimized to fit psychophysical color matching data at different spatial scales.

Our fitted $ECSF$ is at the core of our proposal and represents its most novel component. It had been previously adjusted by fitting the same low-level model to predict matching of color inductive patterns by human observers. The fact that this function can also model saliency provides support for the hypothesis of a unique underlying low-level mechanism for different visual tasks. This mechanism can be modelled either to predict color appearance (by applying the inverse wavelet transform onto the decomposed coefficients modulated by the $ECSF$ weights) or visual salience (by applying the transform to the weights themselves instead). In addition, we introduce a novel approach to selecting the size of the normalization window, which reduces the number of parameters that must be set in an ad-hoc manner.

Our two main contributions can be summarized as follows:

1. We adapt a low-level color induction model in order to predict saliency. The resultant saliency model inherits an extended Contrast Sensitivity Function (termed the $ECSF$), which provides a biologically-plausible manner of integrating scale, orientation and color.

2. A reduction of ad-hoc parameters by including an $ECSF$ which has been fitted to psychophysical data and has no free parameters.

The proposed model exceeds the performance of state-of-the-art saliency estimation methods in predicting eye-fixations for two datasets and using two metrics. Its success in predicting eye-fixations suggests a similar architecture for both the low-level visual saliency machinery and the colour perception machinery in humans.

The rest of this chapter is organized as follows. In section 3.1 we present the low-level color vision model and our fitted ECSF. In section 3.2, we use the resulting weights of the model to compute saliency while in section 3.2.1 we evaluate the model's performance. Section 3.2.2 summarizes the results and section 3.3 discusses further work.

3.1 A low level vision model

Two decades ago, a modular paradigm arose in biological vision, similar to that described in section 2.1 for saliency, stating that color perception occurs in the visual system in a specific cortical area, V4 [133]. This modular paradigm has been challenged in recent years by research supporting the view of a more interlinked processing of color and form in the human visual cortex [107]. Accordingly, both the spatial layout and spectral reflectances of surfaces are processed simultaneously by the same neurons in V1 and other areas.

The saliency estimation method we propose in this work is an extension of a computational model of color perception developed by Otazu *et al.* [94]. The model is based on a non-modular approach to combining color, scale and orientation and has been designed to predict well-known color perception phenomena. Color perception is the result of several adaptation mechanisms which cause the same patch to be perceived differently depending on its surround. Areas A and B of both images in Figure 3.1 are perceived as having different brightness (in panel a) and/or different color (in panel c) respectively, although in both cases they are physically identical (intensity and RGB color channel profiles are plotted as solid lines in the corresponding panels (b) and (d)). These illusions [1] are predicted by the color model of Otazu *et al.* [94], shown in dashed lines in Figure 3.1 (panels (b) and (d)). For example, area A is darker in graphic (b) and area B is more orange-ish in graphic (d).

The model of [94] captures the effect of three key properties on the perceived color of stimuli. In the following paragraphs we describe these effects and how they have been incorporated into our saliency model.

First, the perceived color of a stimulus is influenced by the *surround spatial frequency*. Fig. 3.2(a) shows how surround spatial frequency affects the perceived colors of 4 identical stimuli. In a high-frequency background the color of the stimulus approaches that of the surround (top left stimulus becomes more greenish while the bottom left becomes yellowish). In a low-frequency background the stimulus's perceived color moves away from the surround color (top right stimulus becomes more yellowish when surrounded by green; bottom right more greenish when surrounded by yellow). These induction effects are termed assimilation and contrast respectively.

Second, *orientation* also influences color appearance. In Fig. 3.2(b) we can observe that the relative orientation between the stimulus and the surround provokes a perceptual change. While the top left and right stimuli clearly undergo assimilation (a greenish perception when surrounded by pink, and a bluish perception when surrounded by blue), the stimuli at bottom appear closer to their true cyan color. This is because assimilation is greatest when the stimulus and background have the same orientation.

These two effects are incorporated by representing images using a wavelet decomposition, which jointly encodes the spatial frequency and orientation of image stimuli. In the first stage of Otazu *et al.*'s model, an image is convolved with a bank of filters using a multi-resolution wavelet transform.

[1] the Checkershadow and Beau-lotto illusions were created by E.H. Adelson and Beau Lotto respectively.

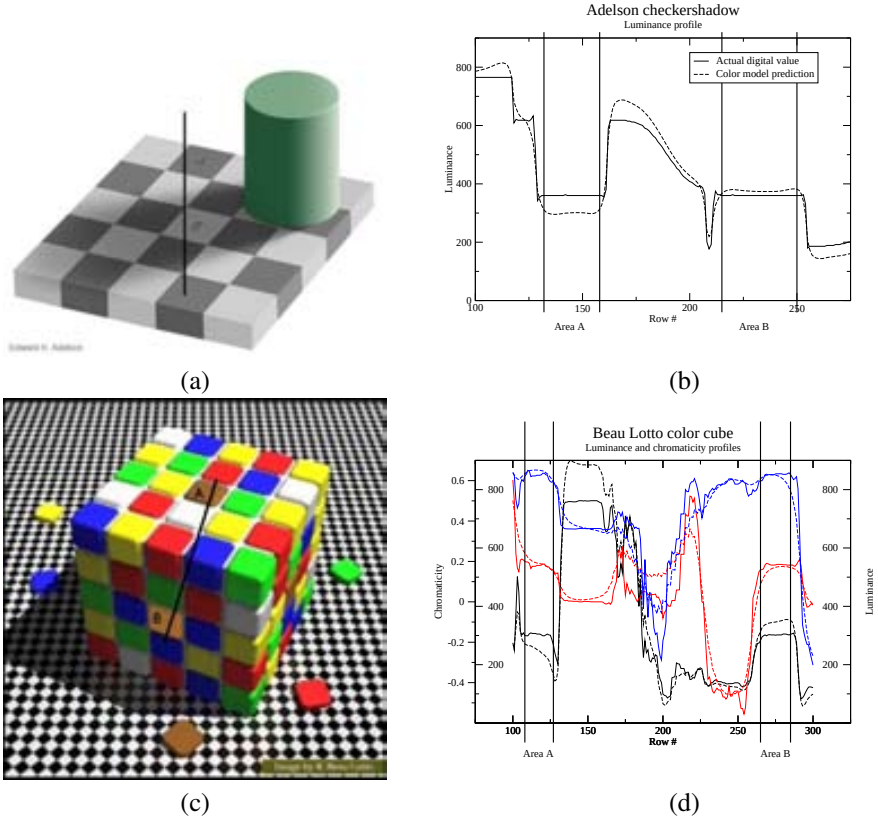


Figure 3.1: Brightness and color visual illusions with their corresponding image profiles (continuous lines, panels b and d) and model predictions profiles (broken lines, in panels b and d).

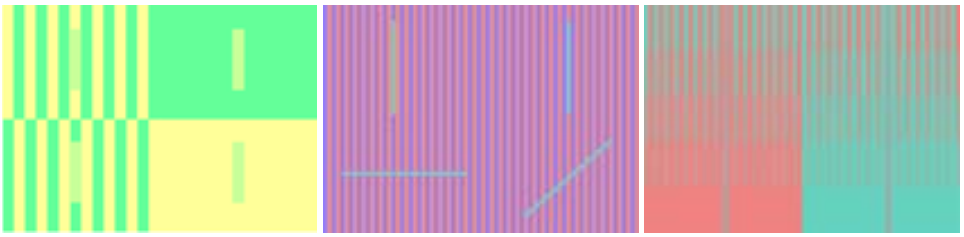
The resulting spatial pyramid contains wavelet planes oriented either horizontally (h), vertically (v) or diagonally (d). The coefficients of the spatial pyramid obtained using the wavelet transform can be considered an estimation of the local oriented contrast. For a given image I, the wavelet transform is denoted as

$$WT(I_c) = \quad w_{s,o} \quad {}_{s=1,2,\ldots,n\,;\,o=h,v,d} \tag{3.1}$$

where $w_{s,o}$ is the wavelet plane at spatial scale s and orientation o and I_c represents one of the opponent channels $O1$, $O2$ and $O3$ of image I, computed as:

$$O1 = \frac{R-G}{R+G+B}, O2 = \frac{R+G-2B}{R+G+B}, \text{and } O3 = R+G+B \tag{3.2}$$

Each opponent channel is decomposed into a spatial pyramid using the wavelet transform, WT. This transform contains Gabor-like basis functions, as Gabor functions resemble the receptive fields of neurons in the cortex. The number of scales used in the decomposition is given by $n = log_2 D$ for an image whose largest dimension is size D.



(a) (b) (c)

Figure 3.2: Perceived color of the stimulus depends on the (a) color and frequency of the surround; (b) relative orientation of the stimuli to the surround; (c) self-contrast of the surround.

Third, *surround contrast* also plays a crucial role in how color is perceived. As shown in Fig. 3.2(c), chromatic assimilation is reduced and chromatic contrast is increased when the surround contrast decreases. Therefore the amount of induction at an image location is modulated by the surround contrast at that location.

Surround contrast is computed in the second stage of the induction model. The surround contrast of a stimulus at position x, y can be modeled as a divisive normalization, which we term the normalized center contrast, $z_{x\ y}$, around a wavelet coefficient $w_{x\ y}$. It is estimated as a normalization of the variance of the coefficients of the central region $a^{cen}_{x\ y}$ normalized by the variance of the coefficients of the surround region $a^{sur}_{x\ y}$:

$$z_{x\ y} = \frac{(a^{cen}_{x\ y})^2}{(a^{cen}_{x\ y})^2 + (a^{sur}_{x\ y})^2}. \tag{3.3}$$

so that $z_{x\ y} \in [0, 1]$. When $z_{x\ y} \to 0$, central activity $a^{cen}_{x\ y}$ is much lower than surround activity $a^{sur}_{x\ y}$. Similarly, when $z_{x\ y} \to 1$, central activity is much higher than surround activity. Therefore, $r_{x\ y}$ may be interpreted as a saturated approximation to the relative central activity $a^{cen}_{x\ y}$. The size of central and surround regions are used to define the size of the corresponding h_j filters.

Divisive normalization has been shown by Simoncelli and Schwartz [110] to remove statistical dependencies present in wavelet decompositions of natural scenes and, in this instance, may be viewed as a center-surround contrast mechanism.

The variance of the coefficients of the central region $a^{cen}_{x\ y}$ is estimated by convolving the local region with a binary filter h. The shape of the filter varies with the orientation of the wavelet plane on which it operates, as shown in Figure 3.5. For example, for a horizontal wavelet plane, $a_{x\ y}$ is computed by

$$a_{x\ y} = \sum_j \omega_{x-j\ y} 2h_j \ << FIX >> \tag{3.4}$$

where h_j is the j-th coefficient of the one-dimensional filter h. The filter h_j defines a region around the central wavelet coefficient $\omega_{x\ y}$ where the activity $a_{x\ y}$ is calculated.

The energy of the surrounding regions, $a^{sur}_{x\ y}$, is computed in an analogous manner to $a^{cen}_{x\ y}$, with the only difference being the definition of the filter h, also shown in Figure 3.5.

The three effects mentioned above, spatial frequency, relative orientation, and surround contrast, are integrated using an extended Contrast Sensitvity Function ($ECSF$). The $ECSF$ determines the type of induction depending on the orientation at a specific spatial frequency, and the amount of induction depending on the surround contrast. This function is inspired by the well-known CSF that was measured in [84] for luminance and colour contrast. Otazu *et al.* defined an $ECSF$ which is parametrized by spatial scale s and center-surround contrast energy. Spatial scale is inversely proportional to spatial frequency ν such that $s = log_2(1/\nu) = log_2(T)$, where T is the period and thus denotes one frequency

cycle measured in pixels. The function $ECSF$ is defined as

$$ECSF(z\ s) = z \cdot g(s) + k(s) \tag{3.5}$$

where the function $g(s)$ is defined as

$$g(s) = \begin{array}{ll} \beta e^{-\frac{s^2}{2\sigma_1^2}} & s \leq s_0^g \\ \beta e^{-\frac{s^2}{2\sigma_2^2}} & otherwise \end{array} \tag{3.6}$$

Here s represents the spatial scale of the wavelet plane being processed, β is a scaling constant, and σ_1 and σ_2 define the spread of the spatial sensitivity of $g(s)$. The s_0^g parameter defines the peak spatial scale sensitivity of $g(s)$. In Equation 3.5, the center-surround activity z of wavelet coefficients are modulated by $g(s)$. An additional function, $k(s)$, was introduced to ensure a non-zero lower bound on $ECSF(z\ s)$:

$$k(s) = \begin{array}{ll} e^{-\frac{s^2}{2\sigma_3^2}} & s \leq s_0^k \\ 1 & otherwise \end{array} \tag{3.7}$$

Here, σ_3 defines the spread of the spatial sensitivity of $k(s)$ and s_0^k defines the peak spatial scale sensitivity of $k(s)$.

The function $ECSF$ is used to weight the center-surround contrast energy $z_{x,y}$ at a location, producing the final response $\alpha_{x,y}$:

$$\alpha_{x,y} = ECSF(z_{x,y}\ \ s_{x,y}) \tag{3.8}$$

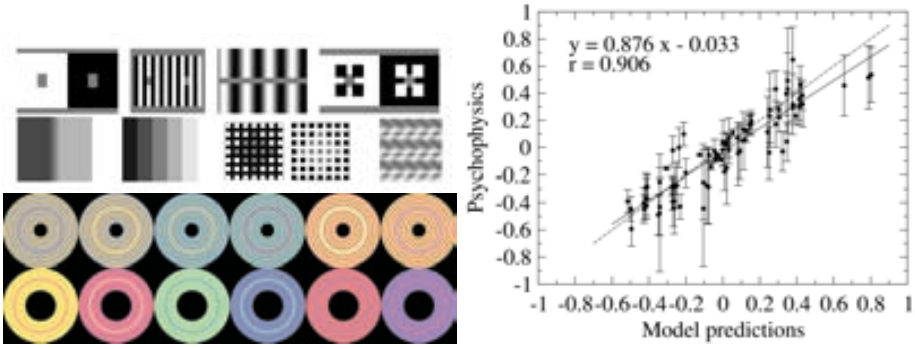
$\alpha_{x,y}$ is the weight that modulates the wavelet coefficient $_{x,y}$. The perceived image channel $I_c^{perceived}$ that contains the color appearance illusions are obtained by performing an inverse wavelet transform on the wavelet coefficients $_{x,y}$ at each location, scale and orientation, after the coefficients have been weighted by the $\alpha_{x,y}$ response at that location:

$$I_c^{perceived}(x\ y) = \sum_{s} \sum_{o} \alpha_{x,y,s,o} \cdot {}_{x,y,s,o} + C_r \tag{3.9}$$

Here o represents the orientation of the wavelet plane of $_{x,y,s,o}$ and C_r represents the residual image plane obtained from WT.

The model of Otazu *et al.* was capable of replicating the psychophysical data obtained from two separate experiments. In the first experiment, by Blakeslee *et al.* [7], observers performed asymmetric brightness matching tasks in order to match the illusions present in regions of the stimuli. Some example brightness stimuli are shown in Figure 3.3(a). The second experiment was performed by Otazu *et al.* [94] in an analogous fashion, but with observers performing asymmetric color matching tasks rather than tasks involving brightness. Some example color stimuli used in these experiments are shown in Figure 3.3(a).

Our saliency estimation model is based on the induction model we have just described. However, to obtain parameters for the intensity and color $ECSF(z\ s)$ functions, we used the psychophysical data from two experiments, one involving color and the other brightness. In the first experiment, by Blakeslee *et al.* [7], observers performed asymmetric brightness matching tasks in order to match the illusions present in regions of the stimuli. The second experiment was conducted by Otazu *et al.* [94] in an analogous fashion, but with observers performing asymmetric color matching tasks rather than tasks involving brightness. The data, provided to us by the authors of [7] and [94], were used to perform a least squares regression in order to select the parameters of the functions. Two different $ECSF$ functions were fitted, one for the achromatic channel and another for the two chromatic channels. Our fitted parameters are given in table 3.1. Both fitted $ECSF(z\ s)$ functions maintain a high correlation rate ($r = 0\ 9$) with the color and lightness psychophysical data, as shown in Figure 3.3(b). Note



(a) (b)

Figure 3.3: (a) Examples of images used in psychophysical experiments. (b) Correlation between model prediction and psychophysical data. The solid line represents the model linear regression fit and the dashed line is the ideal fit. Since measurements involve dimensionless measures and physical units, they were arbitrarily normalized to show the correlation.

Parameter	σ_1	σ_2	σ_3	β	s_0^g	s_0^k
Intensity	1.021	1.048	0.212	4.982	4.000	4.531
Color	1.361	0.796	0.349	3.612	4.724	5.059

Table 3.1: Parameters for $ECSF(z\ s)$ obtained using least square regression.

that both chromaticity channels share the same $ECSF(z\ s)$ function. The profiles of the resulting optimized $ECSF(x\ s)$ functions for brightness and chromaticity channels are shown in Figure 3.4. These $ECSF$s have peak spatial scales in the wavelet decomposition that correspond to peak spatial frequencies between 2-5 cpd, which agree with previous psychophysical estimations [84].

In the induction model of [94], the output of the $ECSF$ was used to weight wavelet coefficients, after which an inverse wavelet transform was performed, producing a new "perceived" image. This reconstructed image replicates color induction phenoma perceived by human observers. For our saliency model, we use these *induction weights* output by the $ECSF$ as a measure of the *saliency* of a feature given its orientation, spatial frequency and center-surround contrast properties.

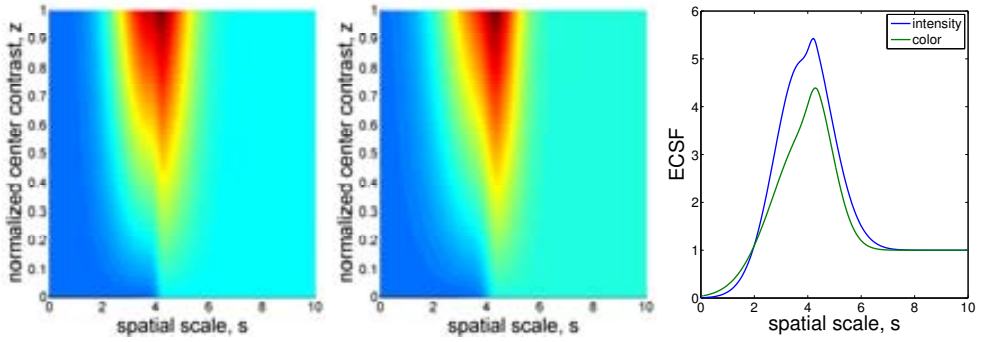
3.2 Building saliency maps

In the previous section we described a low-level color perception model that predicts color appearance phenomena. This model concluded with equation 3.9 which can be re-formulated as

$$I_c^{perceived}(x\ y) = WT^{-1}\ \alpha_{x,y,s,o} \cdot \ _{x,y,s,o} \tag{3.10}$$

where $I_c^{perceived}$ is a new version of the original channel in which image locations may have been modified by the α weight, either by a blurring or an enhancing effect. The colors of modified locations have either been assimilated (averaged) to be more similar to the surrounding color or contrasted (sharpened) to be less similar to the surround.

To obtain predictions of saliency using this color representation, we hypothesize that image locations that undergo enhancement are salient, while locations that undergo blurring are non-salient. In



(a) (b) (c)

Figure 3.4: Weighting functions for (a) intensity and (b) chromaticity channels: Bluer colors represent lower $ECSF$ values while redder colors indicate higher $ECSF$ values. (c) shows slices of both $ECSF(z\ \ s)$ functions for z = 0.9. For a wavelet coefficient corresponding to a scale between approximately 3 and 6, z is boosted. Coefficients outside this passband are either suppressed (for low spatial scales) or remain unchanged (for high spatial scales).

this sense we can define the saliency map of an specific image channel by the inverse wavelet transform of the α weight. Thus the saliency map, S_c, of the image channel I_c at the location $x\ \ y$ can be easily estimated as

$$S_c(x\ \ y) = WT^{-1}\ \ \alpha_{x,y,s,o} \tag{3.11}$$

By removing the wavelet coefficients $\ _{x,y,s,o}$ and performing the inverse transform solely on the weights computed at each image location we provide an elegant and direct method for estimating image saliency from a generalized low level visual representation.

To combine the maps for each channel into the final saliency map, S, we compute the Euclidean norm $S = \sqrt{S_{O1}^2 + S_{O2}^2 + S_{O3}^2}$. The steps of the saliency model are illustrated in Figure 3.5.

Designing the center and surround regions

In stage III of the method, normalized center contrast is measured. The number of pixels spanning the center region and the extended region, comprising both the center and surround regions, are critical parameters. They were chosen so as to resemble the receptive and extra-receptive fields of V1 cortical cells respectively, in a similar fashion to Gao *et al.* [38]. Various studies [14, 112] estimate the central region of the receptive field in V1 cells to correspond on average to a visual angle, β, of approximately 1°. The size of a feature, l, that subtends this visual angle when shown on a screen is computed as $l = d \cdot \tan\beta$, where d is the distance from the observer to the screen. Therefore, the number of pixels P_c that correspond to feature l is $P_c = (d \cdot \tan\beta)\ \ (\frac{mon}{res})$, where mon is the size of the monitor and res is the average of the horizontal and vertical resolution of the displayed image. We used this P_c value as the diameter of the central region.

The diameter of the extra-receptive field has been estimated to be at least 2 to 5 times that of the receptive field [18, 120]. We experimented with diameters in this range and found a size of 5.5 times that of the central region to perform well.

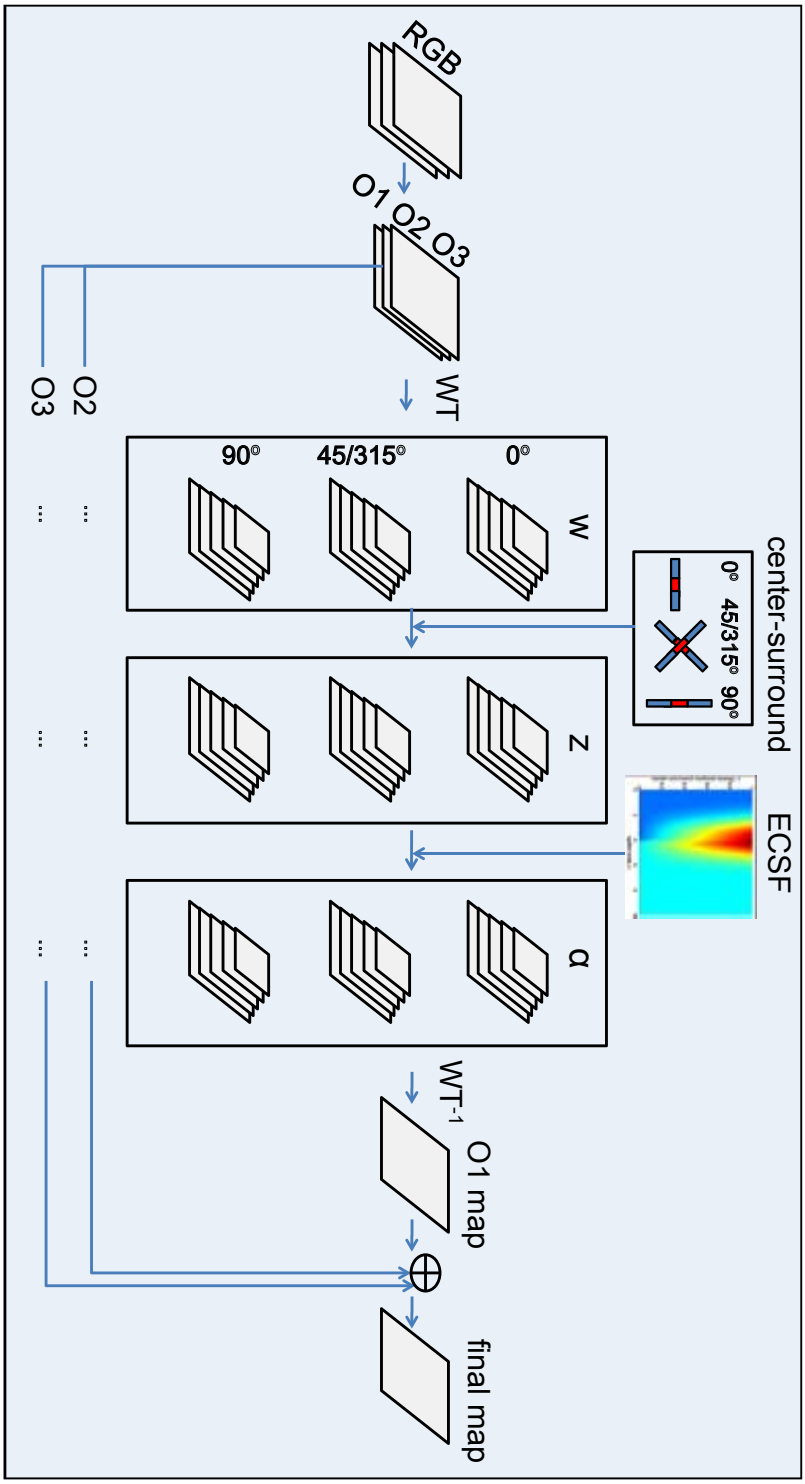


Figure 3.5: Schematic of our saliency approach. Red sections of the center-surround filters correspond to the central filters while blue sections correspond to the surround filters.

3.2.1 Experimental results

We evaluated our model's performance with respect to predicting human eye fixation data from two image datasets. To assess the accuracy of our model we used both the well-known receiver operating characteristic (ROC) and Kullback-Leibler (KL) divergence as quantitative metrics. The ROC curve indicates how well the saliency map discriminates between fixated and non-fixated locations for different binary saliency thresholds while the KL divergence indicates how well the method distinguishes between the histograms of saliency values at fixated and non-fixated locations in the image. For both of these metrics, a higher value indicates better performance.

Zhang *et al.* noted that several saliency methods have image border effects which artificially improve the ROC results [135]. To avoid this issue and ensure a fair comparison of saliency methods we adopt the evaluation framework described by Zhang *et al.* [135], which involves modified metrics for both the area under the ROC curve (AROC) and KL divergence. For each image in the dataset, true positive fixations are fixations for that image, while false positive fixations are fixations for a *different* image from the dataset, chosen randomly. This avoids the true positive fixations having a center bias with respect to the false positive fixations. Because the false fixations for an image are randomly chosen, a new calculation of the metrics is likely to produce a different value. Therefore we computed the metrics 100 times in order to compute the standard error. The saliency maps are shuffled 100 times. On each occasion, the KL-divergence is computed between the histograms of saliency values at unshuffled fixation points and shuffled fixation points. When calculating the area under the ROC curve, we also used 100 random permutations of the fixation points.

The first dataset we use was provided by Bruce & Tsotsos in [12]. This popular dataset is commonly used as the benchmark dataset for comparing visual saliency predictions between methods. The dataset contains 120 color images of indoor and outdoor scenes, along with eye-fixation data for 20 different subjects. The mean and the standard error of each metric are reported in Table 3.2. We performed this evaluation on five state-of-the-art methods as well as our proposed method and as Table 4.1 shows, our method exceeds the state-of-the-art performance as measured by both metrics.

Model	KL (SE)	AROC (SE)
Itti [54]	0.1913 (0.0019)	0.6214 (0.0007)
AIM [12]	0.3228 (0.0023)	0.6711 (0.0006)
SUN [135]	0.2118 (0.0019)	0.6377 (0.0007)
GBVS [46]	0.1909 (0.0015)	0.6324 (0.0006)
Seo [106]	0.3558 (0.0027)	0.6783 (0.0007)
DVA [49]	0.3227 (0.0024)	0.6795 (0.0007)
SIGS [48]	0.3679 (0.0025)	0.6868 (0.0007)
SIM	0.4456 (0.0031)	0.7077 (0.0007)

Table 3.2: Performance in predicting human eye fixations from the Bruce & Tsotsos dataset.

The second dataset we used was introduced by Judd *et al.* in [60]. This dataset contains 1,003 images of varying dimensions, along with eye fixation data for 15 subjects. In order to be able to compare fixations across images, only those images whose dimensions were 768x1024 pixels were used, reducing the number of images examined to 463. This dataset is more challenging than the first as its images contain more semantic objects which are not modeled by bottom-up saliency, such as people, faces and text. Therefore, as would be expected, the AROC and KL divergence metrics are lower for all bottom-up visual attention models. The results, obtained using the same evaluation method described previously, are shown in Table 3.3 and indicate that once again our method exceeds state-of-the-art performance.

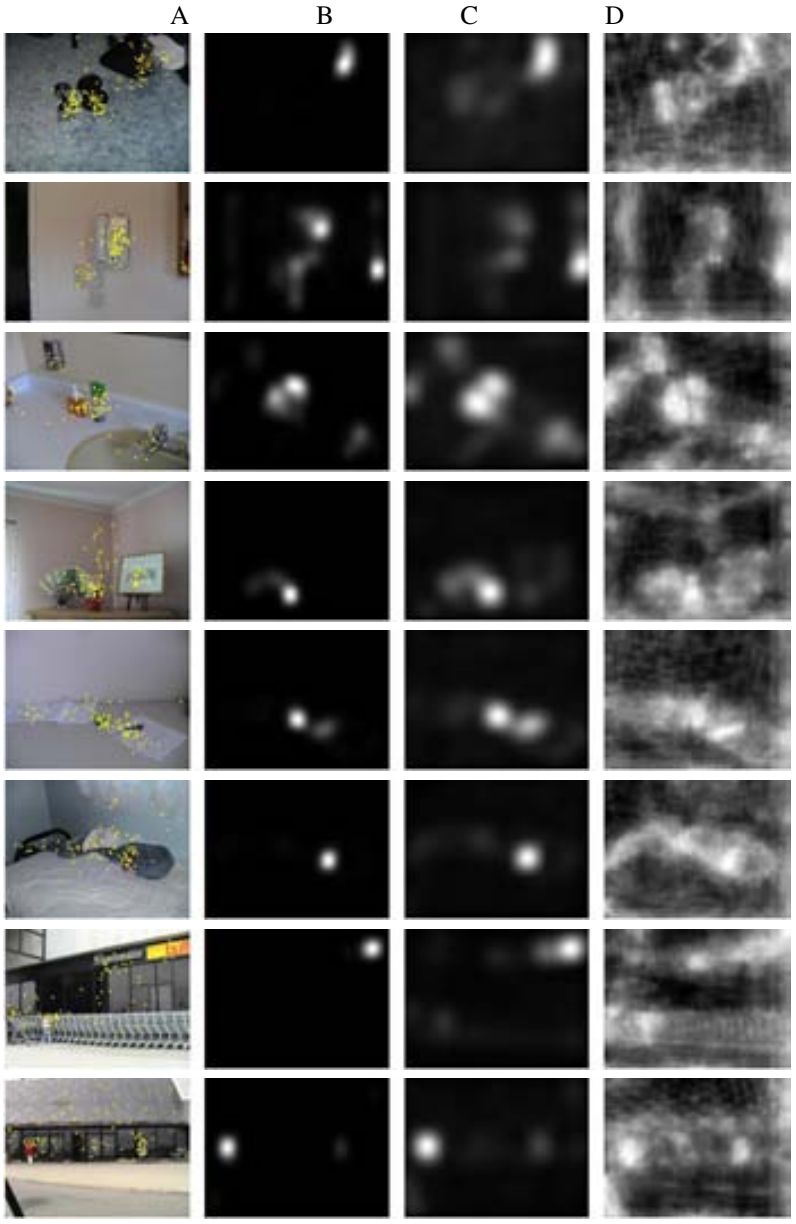


Figure 3.6: Qualitative analysis of results for Bruce & Tsotsos dataset: Column A contains original image. Columns B, C, and D contain saliency maps obtained from Bruce & Tsotsos, Seo & Milanfar and our method, respectively. Yellow markers indicate eye fixations. Our method is seen to be less sensitive to low-frequency edges such as street curbs and skylights, which is in line with human eye fixations.

Model	KL (SE)	AROC (SE)
Itti [54]	0.2073 (0.0014)	0.6285 (0.0005)
AIM [12]	0.2647 (0.0016)	0.6506 (0.0004)
SUN [135]	0.1832 (0.0012)	0.6244 (0.0004)
GBVS [46]	0.1207 (0.0008)	0.5880 (0.0003)
Seo [106]	0.2749 (0.0015)	0.6479 (0.0004)
DVA [49]	0.2924 (0.0016)	0.6565 (0.0005)
SIGS [48]	0.2953 (0.0014)	0.6555 (0.0004)
SIM	0.3021 (0.0017)	0.6695 (0.0005)

Table 3.3: Performance in predicting human eye fixations from the Judd *et al.* dataset.

3.2.2 Discussion

Figure 3.6 illustrates the benefit of our method when compared to Bruce & Tsotsos [12] and Seo & Milanfar [106]. The saliency maps have each been thresholded to their top 10% most salient locations and show that the most salient regions of our saliency map better correspond to the fixations of human observers. In addition, the ROC curves for the three methods in Figure 3.8 show that our method has fewer false positives at higher thresholds, indicating that the proposed method is better able to detect the most salient regions of the image.

Figure 3.7 shows qualitative results for the second dataset, provided by Judd *et al.* [60]. Here there is also a higher correlation between the most salient regions of our saliency map, and human eye fixations, when compared with Bruce & Tsotsos and Seo & Milanfar.

We attribute our model's success to the fact that it is less sensitive to low-frequency edges in the images, such as skylines and road curbs. In addition, we avoid excessive sensitivity to textured regions by suppressing high-frequency information using the weighting functions $ECSF(z\ s)$. As Figure 3.4 shows, the weighting function is more sensitive to mid-range frequencies. The previous methods included in Table 4.1 either select information at one scale or combine scale information from subband pyramids by an unweighted linear combination while in our method, $ECSF(z\ s)$ acts as a bandpass filter in the image's spatial frequency domain, and provides a biologically plausible mechanism for combining spatial information.

Integrating scale information is of particular importance as salient features in a scene may occupy different spatial frequencies, as shown in Figure 3.9. Therefore a mechanism to locate salient features at different levels of the spatial pyramid and combine these features into a final map is critical.

3.3 Conclusions and further work

The proposed saliency model can be summarized by the following pipeline:

$$I_c \underline{\ }^{WT} \quad _{s,o} \quad \underline{\ }^{CS} \quad z_{s,o} \quad \underline{\ }^{ECSF} \quad \alpha_{s,o} \quad \underline{\ }^{WT^{-1}} \quad S_c$$

where CS represents the center-surround mechanism and $ECSF$ is the extended contrast sensitivity function. The main advantage of our formulation is the use of a scale-weighting function that is less sensitive to non-salient edges and provides a biologically plausible mechanism for integrating scale information contained in the spatial pyramid. In the following chapter, we will describe how the introduction of an image representation based on geometric grouplets improves the performance of SIM.

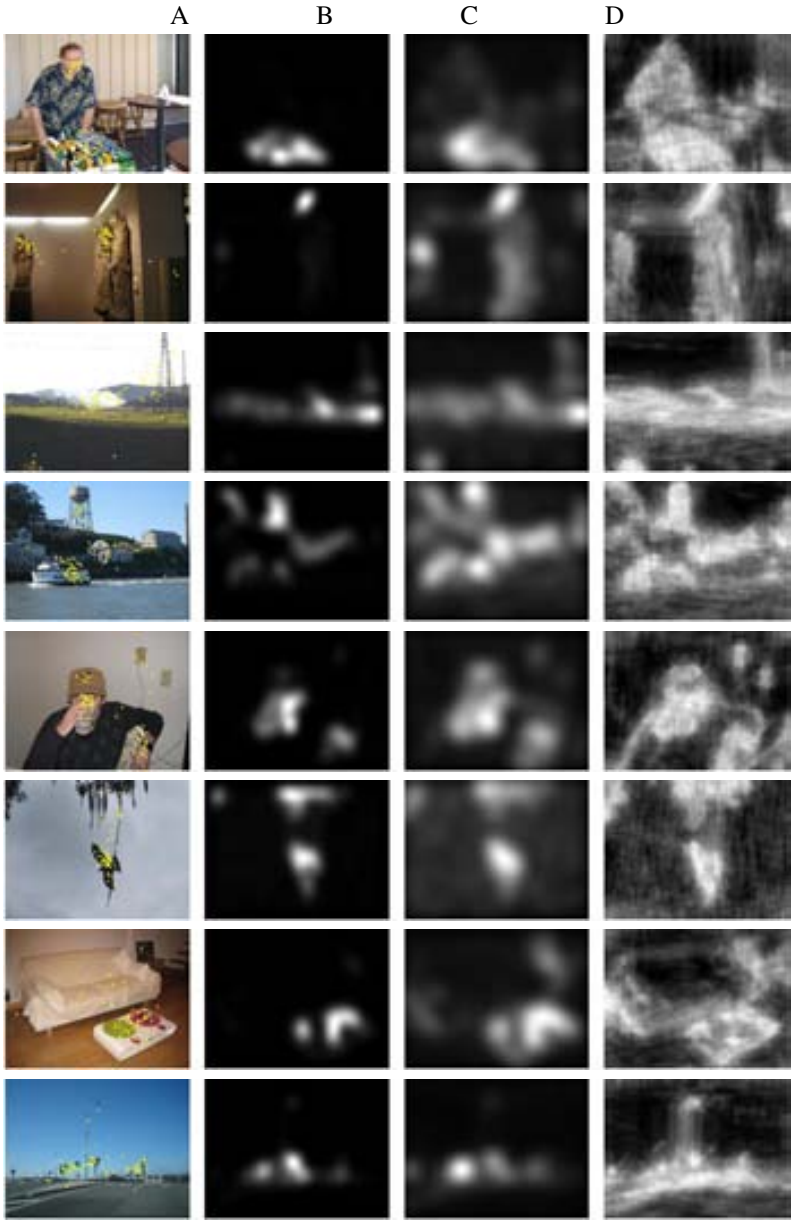


Figure 3.7: Qualitative analysis of results for Judd *et al.* dataset: Column A contains original image. Columns B, C, and D contain saliency maps obtained from Bruce & Tsotsos, Seo & Milanfar and our method, respectively.

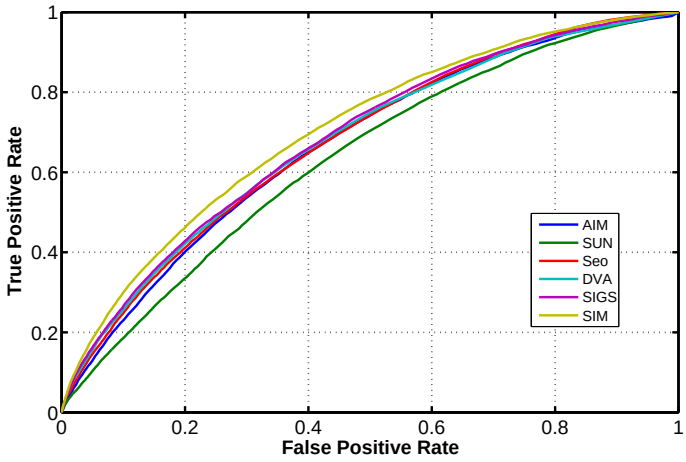
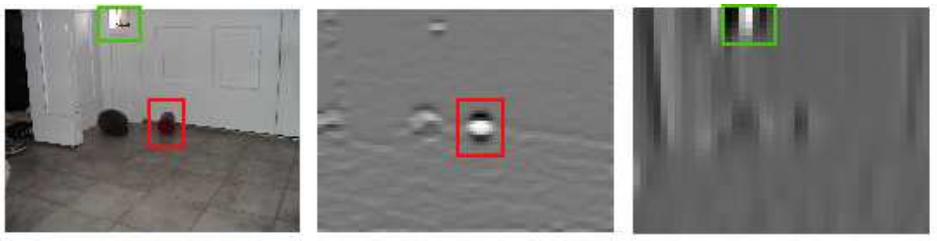


Figure 3.8: ROC curves for state-of-the-art methods and SIM, for the Bruce & Tsotsos dataset.



(a) (b) (c)

Figure 3.9: (a) Two salient features of a scene outlined in green and red. In (b) and (c) we show the spatial scale and orientations at which each object is most prominent. Because these scales and orientation are different for the two features, integrating information contained in the spatial pyramid is critical.

Chapter 4

Grouplets: A Sparse Image Representation for Saliency Estimation

4.1 Introduction

As described in section 3.2, we use a wavelet transform as an image representation. This representation agrees with a long-standing view of the early human sensory system as an efficient information processing system [3, 4, 53]. In this view, one of the objectives of early sensory coding is to transform the visual signal into a sparse, statistically independent representation such that *redundancy has been removed*.

Wavelet decompositions are highly sensitive to edges, in addition to more complex features resulting from super-imposed orientations, such as corners and terminations. However, in comparision with edges, complex features are preferentially fixated on when humans free-view natural images, [5, 99, 134]. Therefore, to estimate saliency, an image representation with higher responses for complex features, relative to the responses for simple features, is desirable.

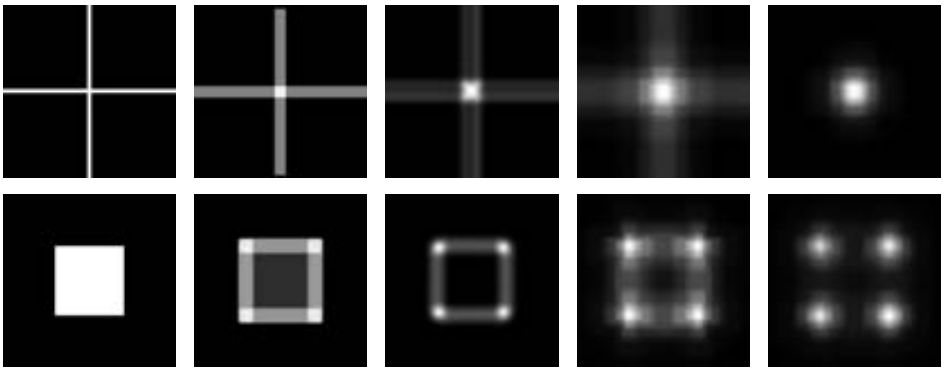
In this chapter, we propose to enhance SIM by introducing an additional stage of the image representation that renders it more responsive to complex features. To generate such a representation we apply a Grouplet Transform (GT) [80] to each wavelet plane $w_{s,o}$. The GT produces a sparse and efficiently-computed image representation that selects for features known to guide visual attention and suppresses non-salient features, as illustrated in Figure 4.1.

The proposed model exceeds the performance of state-of-the-art saliency estimation methods in predicting eye-fixations for two datasets and using two metrics. Its success in predicting eye-fixations suggests a similar architecture for both the low-level visual saliency machinery and the colour perception machinery in humans.

The remainder of this chapter is organized as follows: in section 4.2 we describe our sparse image representation based on geometrical grouplets. Our modified saliency estimation framework is detailed in section 4.3. In section 4.4 we discuss quantitive and qualitative experimental results and we draw several conclusions in section 4.5.

4.2 The grouplet transform for image representation

The GT is constructed a modified Haar transform, computed using a lifting scheme. The Haar transform (HT) decomposes a signal into a residual (lower-frequency) component and a detail (higher-frequency) component. When the signal is a wavelet plane $w_{s,o}$, its residual data $r_{s,j,o}$ is initialized to $w_{s,o}$. The



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

Figure 4.1: The proposed method selects for visually salient features such as junctions and corners. Column (a) contains the original image. Columns (b), (c), (d), and (e) contain saliency maps obtained from Bruce & Tsotsos, Seo & Milanfar, SIM without the GT and SIM with the GT, respectively.

grouplet scale j increases from 1 to J, where J is the number of scales. For a horizontal wavelet support, the HT groups consecutive residual coefficients $r_{s,j,o}(2x-1\ y)$ and $r_{s,j,o}(2x\ y)$ at scale j to compute the residual at the subsequent scale $j+1$:

$$r_{s,j+1,o}(x\ y) = \frac{r_{s,j,o}(2x-1\ y) + r_{s,j,o}(2x\ y)}{2} \tag{4.1}$$

The detail data is computed as a normalized difference of the consecutive residual coefficients:

$$d_{s,j+1,o}(x\ y) = \frac{r_{s,j,o}(2x\ y) - r_{s,j,o}(2x-1\ y)}{2^j} \tag{4.2}$$

A GT is a Haar transform in which the residual and detail coefficients are computed between pairs of elements which are not necessarily consecutive, but are *paired along the contour to which they both belong*. To ascertain the contour along which coefficients should be paired, an "association field" is defined using a block matching algorithm. In this field, associations occur between points and their neighbors in the direction of maximum regularity. In this way, the association field encodes the anistropic regularities present in the image. The regularities in $r_{s,j,o}$ are suppressed in $d_{s,j+1,o}$ by equation 4.2. Therefore, the GT is in essence a differencing operator applied to neighboring wavelet responses along a contour. Neighbors with similar values produce low responses in $d_{s,j+1,o}$ while those with differing values or singularities produce high responses, as illustrated in Fig. 4.2. By computing $d_{s,j,o}\ \ j=1 \qquad J$, points are grouped across increasingly long distances. Each resultant grouplet plane is a sparser representation that contains comparatively higher coefficients for complex geometrical features, whilst simple features are suppressed.

In our saliency model, we apply the GT to wavelet coefficients in order to obtain this improved representation in which salient features are more prominent. It has been suggested that the hierarchical application of the GT to wavelet coefficients may mimic long-range horizontal connections between simple cells in area V1 [80].

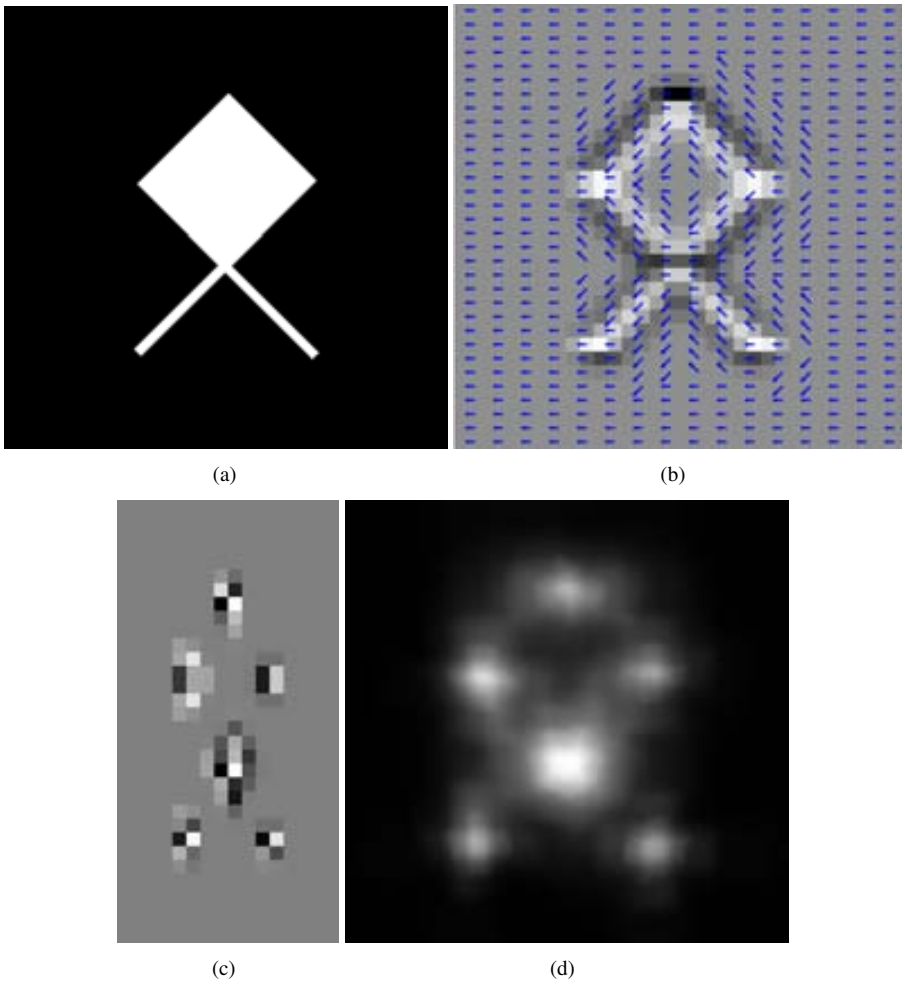


Figure 4.2: Grouping associated wavelet coefficients: (a) shows the input image; (b) shows the association field at j = 1 over a vertically orientated wavelet plane (dark coefficients in the wavelet plane are negative, bright coefficients are positive and gray coefficients are close to zero). The association field (arrows) groups coefficients. The resultant grouplet detail plane in (c) is more sparse than the wavelet plane, preserving only the variations occurring at the corners and terminations; (d) shows the final saliency map (see section 4.3).

4.3 Saliency estimation

We claimed that complex image features such as corners, terminations or crossings emerging from contours are salient. We proposed that a grouplet transform be used to enhance these complex features in the image representation. The grouplet transform furthers distill the information present in the wavelet decomposition of an image.

Considering this hypothesis, here we propose a 6-stage model that estimates saliency by enhancing image locations with certain local spatio-chromatic properties and/or contour singularities. Our model contains the main stages of a color induction model [94], which uses a wavelet decomposition and a function that modulates wavelet coefficients according to their local properties. We introduce a grouplet transform that enables the grouping of simple features whilst maintaining singularities. Below, we describe the stages of our saliency model.

Stage (I): Color representation Three opponent color channels are obtained from image I by converting each (RGB) value, after γ correction, to the opponent space so that:

$$O1 = \frac{R-G}{R+G+B}, O2 = \frac{R+G-2B}{R+G+B}, \text{and } O3 = R+G+B \tag{4.3}$$

Stage (II): Spatial decomposition Each channel is decomposed in two successive steps. The first one uses the wavelet transform in equation 3.1, obtaining $w_{s,o}$. Subsequently, on each wavelet plane the grouplet transform in equation 4.2 is applied:

$$I_c \stackrel{WT}{-} \quad _{s,o} \quad \stackrel{GT}{-} \quad d_{s,j,o} \tag{4.4}$$

where $d_{s,j,o}$ denotes the detail plane at scale j. For a wavelet plane whose largest dimension is size D, $J = log_2 D$. To group features, the association field for a wavelet plane is initialized perpendicularly to its orientation o. Thus for a horizontal wavelet plane, the Haar differencing in equation 4.2 is conducted column-wise and vice versa.

Stage (III): Normalized Center Contrast (NCC) We compute the NCC, $z_{s,j,o}(x\ y)$, for every grouplet coefficient $d_{s,j,o}(x\ y)$ using equation 3.3. The number of pixels spanning the center region and the extended region was set as described in section 3.2.

Stage (IV): Induction weights ($ECSF$) The $ECSF$ function is used to compute induction weights $\alpha_{s,j,o}(x\ y)$ for every grouplet coefficient $d_{s,j,o}(x\ y)$:

$$\alpha_{s,j,o}(x\ y) = ECSF(z_{s,j,o}(x\ y)\ s) \tag{4.5}$$

The $\alpha_{s,j,o}(x\ y)$ weight gives a measure of saliency for location $(x\ y)$ in $d_{s,j,o}$. The $ECSF$ acts so that $z_{s,j,o}$ values with scales s in the passband of the $ECSF$ are enhanced, while those with scales outside of this passband are suppressed.

Each $\alpha_{s,j,o}$ plane is resized to the size of its corresponding wavelet plane $w_{s,o}$ using bicubic interpolation, and then summed to produce $\alpha_{s,o}$ for that wavelet plane:

$$\alpha_{s,o}(x\ y) = \quad _{j} \quad (\alpha_{s,j,o}(x\ y)) \tag{4.6}$$

where $(\cdot)$ denotes bicubic interpolation.

Stages (V)-(VI): Saliency Map Recovery Finally, an inverse wavelet transform is performed on the spatial pyramid of $\alpha_{s,o}$ planes to produce the final saliency map S_c for an image channel. At this point the pipeline of the model may be summarized as

$$I_c \overset{WT}{-} \quad _{s,o} \overset{GT}{-} d_{s,j,o} \overset{NCC}{-} z_{s,j,o} \overset{ECSF}{-} \alpha_{s,j,o} \overset{\varphi}{-} \alpha_{s,o} \overset{WT^{-1}}{-} S_c \tag{4.7}$$

The saliency maps for all three image channels are combined to form the final saliency map S using the Euclidean norm $S = \sqrt{S_{O1}^2 + S_{O2}^2 + S_{O3}^2}$. The method is summarized schematically in Fig. 4.3.

4.4 Experiments

To evaluate our model, we applied it to the problem of predicting eye-fixations in the two image datasets described in section 3.2.1: that of Bruce & Tsotsos [12] and that of Judd *et al.* [60]. We also follow the same experimental procedure detailed in that section. That is, the accuracy of the predictions were quantitatively assessed using both the Kullback-Leibler (KL) divergence and the receiver operating characteristic (ROC) metrics. The KL divergence measures how well the method distinguishes between the histograms of saliency values at fixated and non-fixated locations in the image. The ROC curve measures how well the saliency map discriminates between fixated and non-fixated locations for different binary saliency thresholds. For both metrics, a higher value indicates better performance.

Results for the Bruce & Tsotsos dataset are reported in Table 4.1. We see that, with or without the GT, SIM exceeds the state-of-the-art performance as measured by both metrics. Further, the addition of the GT improves upon SIM's performance.

Model	KL (SE)	AROC (SE)
Itti [54]	0.1913 (0.0019)	0.6214 (0.0007)
AIM [12]	0.3228 (0.0023)	0.6711 (0.0006)
SUN [135]	0.2118 (0.0019)	0.6377 (0.0007)
GBVS [46]	0.1909 (0.0015)	0.6324 (0.0006)
Seo [106]	0.3558 (0.0027)	0.6783 (0.0007)
DVA [49]	0.3227 (0.0024)	0.6795 (0.0007)
SIGS [48]	0.3679 (0.0025)	0.6868 (0.0007)
SIM w/o GT	0.4456 (0.0031)	0.7077 (0.0007)
SIM with GT	0.4925 (0.0034)	0.7136 (0.0007)

Table 4.1: Performance in predicting human eye fixations from the Bruce & Tsotsos dataset.

Results for the Judd *et al.* dataset, shown in Table 4.2 indicate that once again the addition of the GT improves upon SIM's state-of-the-art performance.

Implementation Details

The Bruce & Tsotsos dataset was collected on a 21 inch monitor with $d = 29\ 5$ inches. For images with 511x681 resolution, the diameter of the central region, P_c, $= 18$ pixels. The Judd *et al.* dataset was collected on a 19 inch monitor with $d = 24$ inches. For images with 768x1024 resolution, $P_c = 24$ pixels. For a MATLAB implementation running on an Intel Core 2 Duo CPU at 3.00 GHz with 2GB RAM, typical run times for color images of sizes 128x128, 256x256 and 512x512 pixels are 0.6, 1.2 and 3.2 seconds respectively.

4.4.1 Discussion

Qualitative comparisons between two state-of-the-art methods [12, 106] and SIM are displayed in Figs. 4.4 and 4.5. One can see that for the proposed method (column (d)), the most salient regions

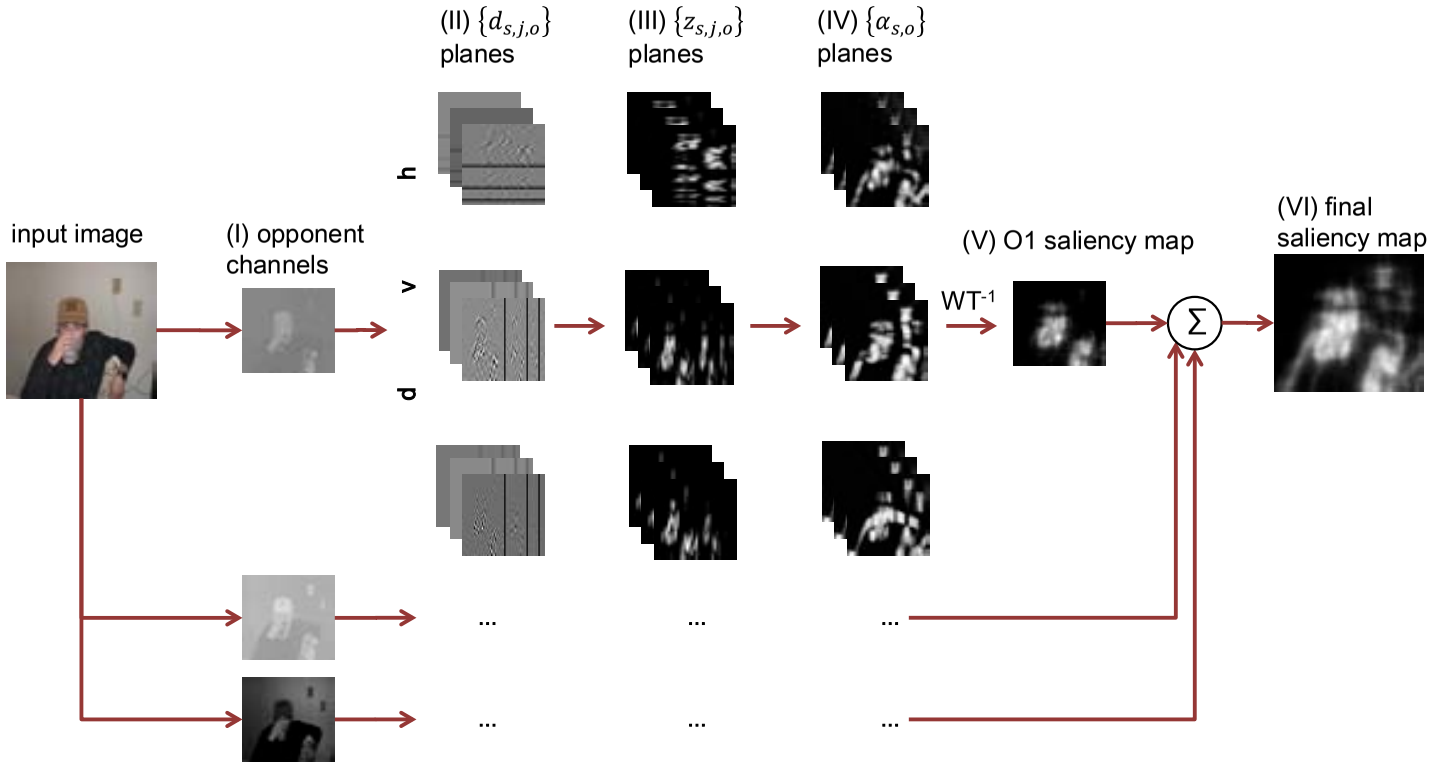


Figure 4.3: Schematic of our saliency method: (I) The image is converted to the opponent space. (II) Each opponent color channel is decomposed using a wavelet transform, after which each wavelet plane is decomposed into grouplet planes. (III) Contrast responses from grouplet planes are calculated and combined to produce the contrast response plane. (IV) The $ECSF$ is used to produce the plane of induction weights $\alpha_{s,o}$. (V) The $\alpha_{s,o}$ planes are combined by an inverse wavelet transform to produce the final saliency map for the channel. (VI) The 3 channels maps are combined using the Euclidean norm.

Model	KL (SE)	AROC (SE)
Itti [54]	0.2073 (0.0014)	0.6285 (0.0005)
AIM [12]	0.2647 (0.0016)	0.6506 (0.0004)
SUN [135]	0.1832 (0.0012)	0.6244 (0.0004)
GBVS [46]	0.1207 (0.0008)	0.5880 (0.0003)
Seo [106]	0.2749 (0.0015)	0.6479 (0.0004)
DVA [49]	0.2924 (0.0016)	0.6565 (0.0005)
SIGS [48]	0.2953 (0.0014)	0.6555 (0.0004)
SIM w/o GT	0.3021 (0.0017)	0.6695 (0.0005)
SIM with GT	0.3678 (0.0020)	0.6788 (0.0005)

Table 4.2: Performance in predicting human eye fixations from the Judd *et al.* dataset.

correspond better to eye-fixations and highly salient features are located at a variety of spatial frequencies.

One can also see in the figures that regions of high saliency are more clearly distinguished from background regions. This is reflected in the large improvements in KL divergence achieved for both datasets. The increased discriminative power is due to the fact that the background features present in the wavelet planes are attenuated by the grouplet transform, as illustrated in Fig. 4.6. These background features tend to be small, isolated features which, while present in wavelet planes, do not persist beyond the first few grouplet planes.

The grouplet transform itself may be considered a center-surround mechanism, as it measures the difference in amplitude between a coefficient and its neighbor. Consequently, regions of the wavelet plane with similar amplitudes, and therefore low contrast, are attenuated in their grouplet planes, while regions of the wavelet plane with large differentials between their amplitudes are enhanced. Therefore the grouplet transform acts to further distill the information present in the wavelet transform, preserving only features which are spatially extensive and strongly contrasting with their surroundings.

Our model required parameters to be set for the $ECSF$ and the center-surround regions. The $ECSF$ parameters were set using psychophysical data and are dataset-independent. Therefore our only free parameters are the center-surround region sizes. As mentioned in section 3.2, the center regions's size was set to correspond to 1° of visual angle, and the surround size was set to be 5.5 times the size of the center region. We found results to be very stable for surround-to-center region ratios from 3-6 and for center sizes of $1^\circ \pm 0$ 2. As such, our model is robust to uncertainty in the choice of free parameters.

We also investigated the effect of changing s_0, the spatial scale for which the $ECSF(z\ s)$ gives the highest response. We varied s_0 for the $ECSF$ of the intensity channel, the channel containing the majority of the saliency information. Fig. 4.7 shows that the model performs best when mid-range frequencies are enhanced and low or high frequencies are inhibited. Furthermore, the best scale range for these metrics, between 4 and 6, is consistent with the value determined using psychophysical data, $s_0 = 4$ 2 (see Fig. 3.4(a)).

4.5 Conclusions

In this work we propose a saliency model based on a biologically-plausible low-level spatio-chromatic representation. Our model measures saliency using the result of the perceptual integration of color, orientation, local spatial frequency and surround contrast. The parameters of our integration mechanisms have been fitted to psychophysical data. In addition, we have shown that prediction of saliency

is improved if we insert a further grouping stage that suppresses simple edges, thereby avoiding strong saliency responses for such features. We demonstrate that the model exceeds state-of-the-art performance in predicting eye-fixations using two metrics and when evaluated with two datasets.

As saliency models cannot hope to replicate visual attention, which is highly susceptible to semantic cues such as faces and text, we would like to expand the model to include such cues. Lastly, we would like to explore the application of grouplet-based representations to other computer vision problems, such as feature detection, which typically involve scale-space decompositions.

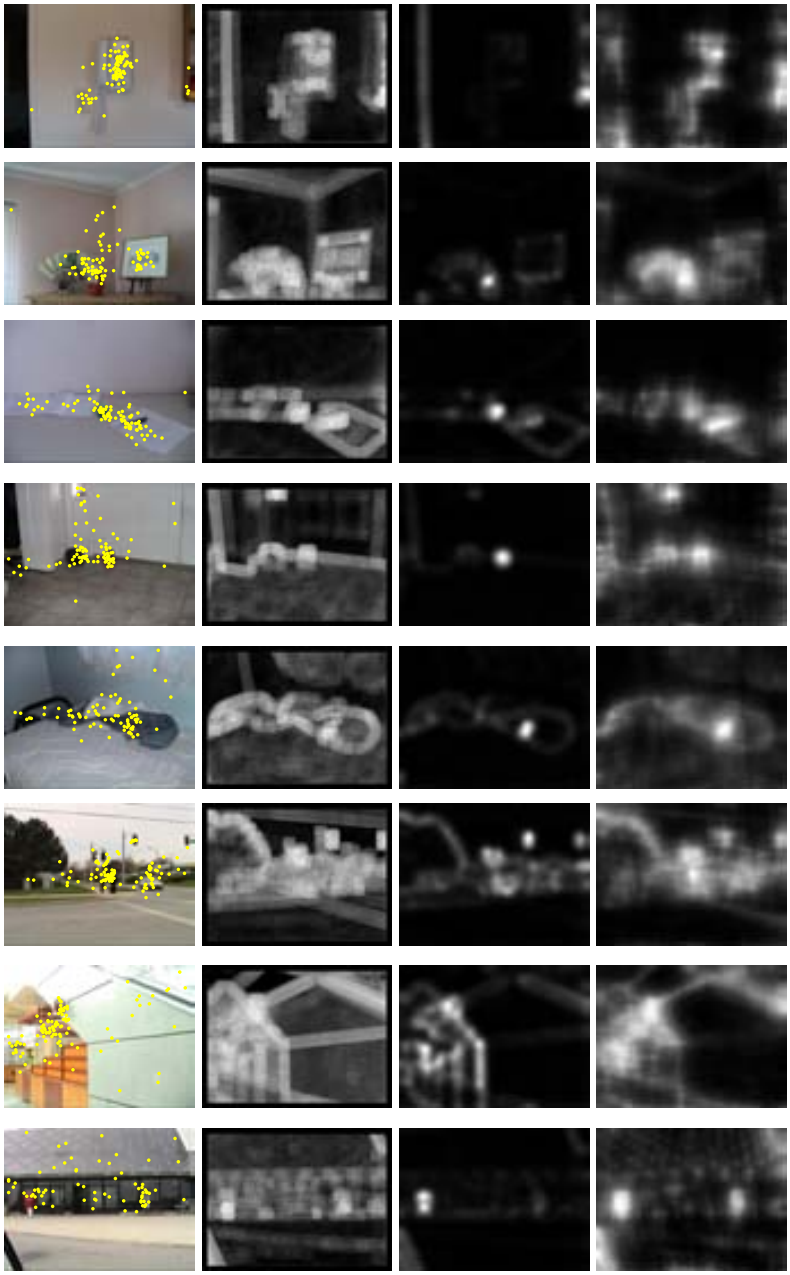


Figure 4.4: Qualitative results for Bruce & Tsotsos dataset: Column (a) contains the original image. Columns (b), (c), and (d) contain saliency maps obtained from [12], [106] and SIM respectively. Yellow markers indicate eye fixations. Our method is seen to more clearly distinguish salient regions from background regions and to better estimate the extent of salient regions.

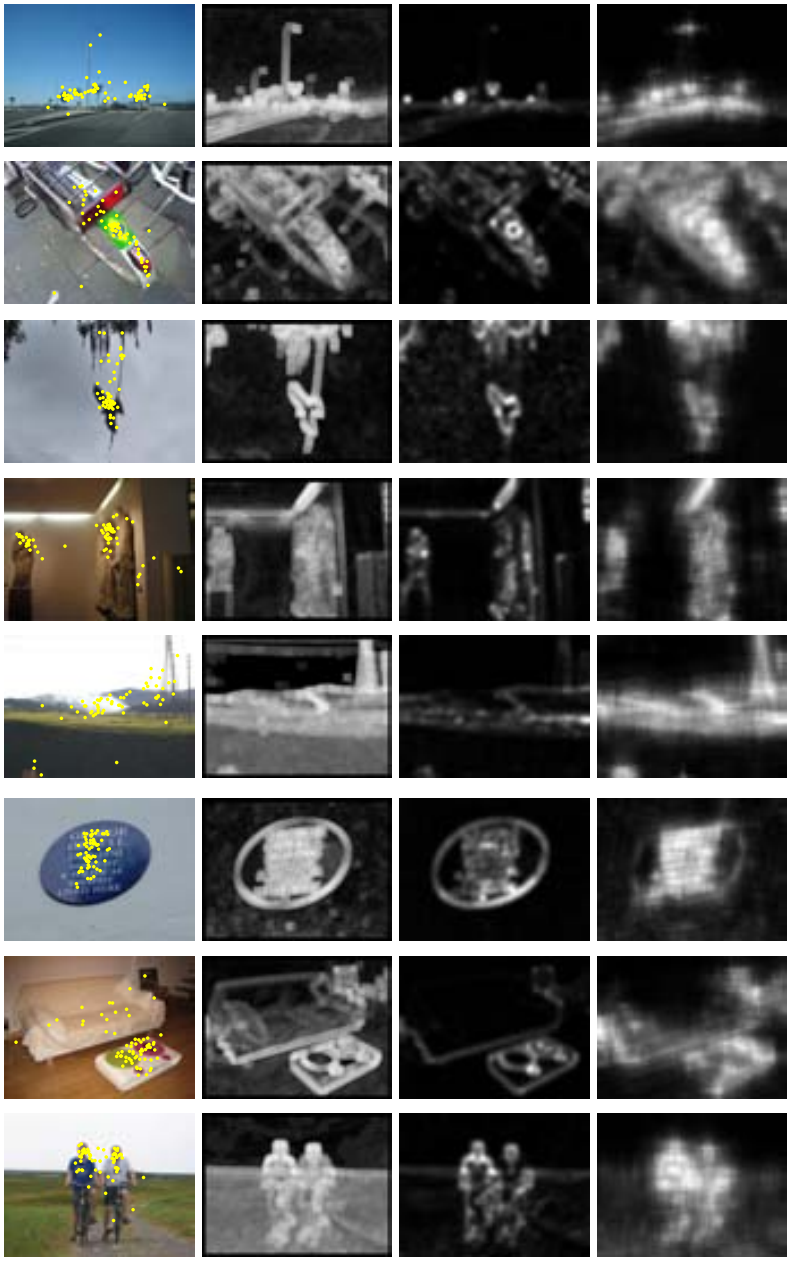


Figure 4.5: Qualitative results for Judd *et al.* dataset: Column (a) contains the original image. Columns (b), (c), and (d) contain saliency maps obtained from [12], [106] and SIM respectively. Yellow markers indicate eye fixations.

(a) Input image (b) Result without GT (c) Result with GT

Figure 4.6: The GT attenuates spatially isolated features.

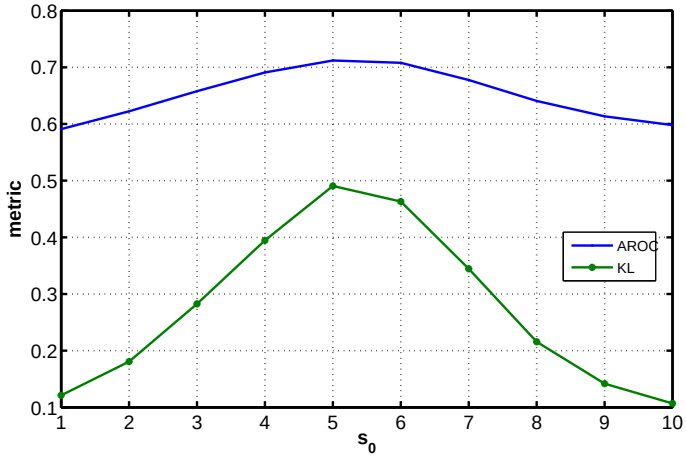


Figure 4.7: Change in AROC and KL metrics with change in s_0 for intensity $ECSF(z\ s)$, for the Bruce & Tsotsos dataset: The best s_0 for both these metrics are in line with the value determined using psychophysical experiments.

Part II

Aesthetic Visual Analysis

Chapter 5

A Brief Review of Image Aesthetics Analysis

With the ever-expanding volume of visual content available, the ability to organize and navigate such content by aesthetic preference is becoming increasingly important. In the case of semantic retrieval, for instance using multi-media search engines, semantic relevance is currently perceived by users as a commoditized feature. This was confirmed by a recent user evaluation [40] performed to determine the key differentiating factors of an image search engine. The top five factors were reported to be: "High-quality" (13%), "Colorful" (10%), "Semantic Relevance" (8%), "Topically clear" (7%) and "Appealing" (5%). Semantic relevance is only ranked as the third factor, whereas features related to the quality and aesthetics rank first and second.

The concept of a "high-quality" or "colorful" can be readily defined. There has been a great deal of research in the vision community into inferring and even improving the quality of an image, where quality in this sense refers to factors such as image resolution, presence or absence of compression artifacts. However, how does one infer whether or not an image is "appealing"? In other words, how does one infer the aesthetics of an image?

Aesthetics has been studied since antiquity by philosophers such as Plato and continues to be the subject of vibrant scholarly exchange today. These exchanges occur in a diverse array of fields, including philosophy, psychology, and more recently, neuroscience [19, 71, 109]. Studies into aesthetics raise such questions as "What are the principles driving aesthetic appreciation?", "Are there universal aesthetic laws?", and "What are the contributions of sensory input, prior knowledge and other factors to aesthetic experiences?".

The philosopher Alexander Gottlieb Baumgarten appropriated the term aesthetics, which had always been connotative of sensations and perception, to give it the meaning in which it is used today, as referring to the sense or perception of beauty [45]. It is defined in the The American Heritage® Dictionary of the English Language [1] as:

> "the study of the mind and emotions in relation to the sense of beauty."

Baumgarten advocated the study of aesthetics as a "science of sensual cognition" [45], and that aesthetic appreciation was the result of objective reasoning. This view was in direct opposition to those of David Hume and Edmund Burke [43, 108], who believed that aesthetic appreciation was a result of induced feelings. Immanuel Kant, however, believed that aesthetic appreciation of an object was a result of the interplay between the perception of its empirical features and the imagination [41]. These differing views are echoed in modern times by the contemporary debate between "internalists", who view aesthetic experience as owing to subjective factors, and "externalists", who typically describe

aesthetic experience as due to objective features of the stimulus under consideration [109].

In the particular case of pictorial art, such as paintings or photographs, there are visual characteristics related to accepted aesthetic principles that transcend subjective factors. For example, certain combinations of colors form what are called “color harmonies” and are held to be more appealing than others as a rule [57]. As another example, the “rule-of-thirds” is a compositional principle that is thought to guide attention [67].

These types of visual characteristics and aesthetic principles are more evident and accessible than the cultural and other subjective influences that govern aesthetic experiences. As a result, they have recently been brought to bear in image aesthetics research conducted in the computer vision community. In the past few years, this community has demonstrated a growing interest in the data-driven analysis of pictorial art, especially photographs and paintings. A representative analysis paradigm is exemplified in Figure 5.1 where, given a set of images, the goal is to classify images into ”good” and ”bad” aesthetic classes.

Figure 5.1: Representative computational framework for image aesthetics analsyis: Binary classification of landscape images into “high-quality” and “low-quality” classes.

In computer vision, most of the research on image aesthetics analysis has focused on feature design. Typically, image features are proposed that aim to represent the visual characteristics related to specific aesthetic principles. For example, features have been designed to detect photographic rules and practices such as the golden ratio, the rule of thirds and color harmonies [25, 31, 59, 62, 76, 77, 105]. Such features are extracted from images and used to train statistical models to discriminate between ”high quality” and ”low quality” images [26, 31, 62, 76, 77, 81], to predict the aesthetic score of an image [25, 125], or to rank images by their aesthetic quality [105]. We describe these two elements of aesthetic prediction - feature representations and discriminative model learning - in the following sections.

5.1 Feature representations

5.1.1 Aesthetics-specific visual features

In the short time span between Datta *et al.*’s seminal work on the topic [25] in 2006, a plethora of aesthetic features have been proposed [25, 31, 62, 76, 77, 105]. Datta *et al.*proposed 56 visual features which could be extracted from an image. “Colorfulness” features were extracted by comparing the distribution of its colors to a reference distribution. Average pixel intensity was used to represent light exposure. Average pixel saturation and hue were also used as features. These averages were computed for pixels lying within the inner rectangle of an image segmented according to the rule-of-thirds. Several other features related to familiarity, texture, size and aspect ratio, region composition, low depth-of-field, and shape convexity were designed.

More recent works have been largely derivative. The features proposed by Ke *et al.* [62] were designed to describe the spatial distribution of edges, color distribution, simplicity, blur, contrast and brightness. Luo & Tang [77] first segmented the subject region of an image before extracting from it high-level semantic features related to composition, lighting, focus controlling and color. Dhar *et al.* [31] aimed to predict aesthetics using attributes describable by humans. These attributes were compositional, content-related and related to illumination.

5.1.2 Generic visual features

Many such techniques are quite high-level and difficult to model. Consequently a gap between the representational power of hand-crafted aesthetic features and the aesthetic quality of an image has emerged. In a recent work, it was shown that generic image descriptors, *i.e.*descriptors which were not specifically designed for aesthetic image analysis, could yield state-of-the-art results [81]. One such descriptor is the Bag-Of-Visual-words descriptor [22, 111], which is quite possibly the most widely used image descriptor for semantic tasks. Another successful generic descriptor was, the Fisher Vector (FV, [97, 98]), a recent extension of the BoV feature vector. Fisher vectors have been shown to yield state-of-the-art results for tasks such as image retrieval and image classification. Both the BoV and FV descriptors were calcuated using SIFT [75] features and color histogram features, before being applied to aesthetic quality prediction.

These generic descriptors implicitly encode the aesthetic characteristics of an image by describing the distribution of intensity and color gradients in local image patches. BoV is a representation of the discrete distribution of patches of various gradient profiles, while the FV represents a continuous distribution of the same patches. Each (color gradient or SIFT) patch contains a great deal of information about the local properties of an image such as the degree of color saturation in the local region or the degree of blur. By summarizing this patch-level information into a single image signature (BoV or FV) signature, one can have a global idea of the proportion of blur and the distribution of color in an image, in addition to the relation between these and other aesthetic characteristics. In addition, although BoV-based signatures by definition discard spatial layout information, this type of information can still be included in a limited way by using the spatial pyramid framework [70]. This type of strategy may enable such descriptors to capture composition information, such as the presence of absence of a rule-of-thirds layout.

5.1.3 Textual features

Textual data associated with an image often contains a great deal of information about the content and aesthetics of the image. In fact, the information contained in text on webpages has been used by many search engines to return images relevant to a given query. For images on social networks such as Flickr, the comments made by users often express their impressions on the aesthetic and artistic qualities of the images. The use of textual data for image aesthetic analysis is a new approach but one which has already given promising results [40, 105]. Standard textual feature vectors, such as word frequency or TF-IDF vectors, may be created from the data associated with images. Such feature vectors are analogous to, and in fact were used to inspire BoV descriptors. In this case, textual words are bagged, rather than visual words.

5.2 Learning discriminative models of visual aesthetics

The features described above, or combinations thereof, have been used to train, via supervised learning, discriminative models for various aesthetics-related image annotation problems.

5.2.1 Binary classification

One such problem is labeling images as belonging to one of two classes: "good/high aesthetic quality" or "bad/low aesthetic quality". This problem is relevant for applications such as automated photo-book construction or culling sets of duplicated images in photo-shoots. Collecting annotations to use for training models is a non-trivial task. Humans tend to disagree when annotating images by their semantic content. Annotator disagreement is significantly greater for the problem of labeling an image as aesthetically pleasing or not. In the case of semantic annotations, multiple annotations per image are collected in order to gain an idea of the general consensus with respect to an image. In the case of aesthetic annotations, a larger number of annotations may be required per image, which would be expensive and laborious to collect. In addition, the annotation task may be difficult or ambiguous when images are neither particular appealing or unappealing.

To deal with this issue, some researchers have simplified the problem by only considering images whose annotations have a high degree of consensus, so that fewer annotations per image are required [62,76]. Others have collected crowd-sourced annotations from social networks for photography enthusiasts, where hundreds of users rate each image [25, 105]. Once collected, ground-truth annotations are used to train discriminative models such as SVMs or decision trees [25,31,81].

5.2.2 Aesthetic score prediction

Another important annotation task is predicting the aesthetic score of an image on some numerical scale. Score predictions can be useful for example, when incorporated into consumer cameras to provide online feedback. For a human annotator, this task is easier for than binary annotations, as the annotations are more granular. However, annotator consensus is still an issue in this case. With these annotations, support vector regression models may be learned. The distribution of scores given to images has also been used to train a structured prediction model to predict such distributions for unseen images [125].

5.2.3 Aesthetics-aware image retrieval

As mentioned previously, aesthetic quality is increasingly important for applications such as content-based image search. When searching for images containing specific contents, users desire that semantically-relevant images that are also *aesthetically pleasing* are returned at the top of the search results. Few works in the literature have tackled this problem [40, 105]. In these works, standard ranking SVMs are trained using annotations obtained by thresholding the aesthetic scores of images into 3 or 4 relevance levels.

5.3 Online feedback systems

The encouraging results obtained by several aesthetics models have lead to the development of a few prototypes for assessing and improving image aesthetics [59]. One such system, ACQUINE [27], has been deployed via a web interface. On the website, an images or the url to an image may be uploaded and a score from 1 to 100 is returned. To date, more than 300,000 images have been uploaded to ACQUINE. Another system, OSCAR [130], may be deployed to a mobile device such as a smart-phone and offers on-line feedback to help the user improve the composition or colorfulness of an image.

5.4 Objectives

As discussed above, rich and representative annotations are essential for successfully training supervised models of image aesthetics but are non-trivial to collect. However, while significant effort has

been dedicated to designing image descriptors for aesthetics, little attention so far has been dedicated to the collection, annotation and distribution of ground-truth data. We believe that *novel datasets shared by the community will greatly advance the research around this problem*. This has been the case for semantic categorization, where successful datasets such as Caltech 101 [69] and 256 [44], PASCAL VOC [36] and Imagenet [28] have contributed significantly to the advancement of research. Such databases are typically composed of images obtained by web-crawling and annotated by crowd-sourcing. In the specific case of aesthetic analysis, having rich and large-scale annotations is a key factor.

However, a major complication of aesthetic analysis in comparison to semantic categorization is the highly subjective nature of aesthetics. To our knowledge, all the image datasets used for aesthetic analysis were obtained from on-line communities of photography amateurs such as `www.dpchallenge.com` or `www.photo.net`. These datasets contain images as well as aesthetic judgments they received from members of the community. Collecting ground truth data in this manner is advantageous primarily because it is an inexpensive and expedient way to obtain aesthetic judgments from multiple individuals who are generally "prosumers" of data: they produce images and they also score them on dedicated social networks.

The interpretation of these aesthetic judgments, expressed under the form of numeric scores, has always been taken for granted. Yet a deeper analysis of the context in which these judgments are given is essential. The result of this lack of context is that it is difficult to understand what the aesthetic classifiers *really* model when trained with such datasets.

While still in its nascent stage, research into computational models of aesthetic preference already shows great potential. However, to advance research, realistic, diverse and challenging databases are needed. To this end, we introduced a new large-scale database for conducting Aesthetic Visual Analysis: AVA. It contains over 250,000 images along with a rich variety of meta-data including a large number of aesthetic scores for each image, semantic labels for over 60 categories as well as labels related to photographic style. In chapter 6, we show the advantages of AVA with respect to existing databases in terms of scale, diversity, and heterogeneity of annotations. We also describe several key insights into aesthetic preference afforded by AVA. In chapter 7 we investigate how this wealth of data can be used to tackle the problem of understanding and assessing visual aesthetics by looking into several problems relevant for aesthetic analysis, in particular image classification, image aesthetic score prediction and image ranking. We demonstrate how the large scale of AVA can be leveraged to improve performance on these tasks.

Chapter 6

AVA: A Large-Scale Database for Aesthetic Visual Analysis

For the problem of semantic categorization, datasets such as Caltech 101 [69] and 256 [44], PASCAL VOC [36] and Imagenet [28] have contributed significantly to the advancement of research. Such databases are typically composed of images obtained by web-crawling and annotated by crowdsourcing.

In the specific case of visual aesthetic analysis, having rich and large-scale annotations is a key factor. However, little attention so far has been dedicated to the collection, annotation and distribution of ground truth data for studying visual aesthetics.

A major complication of aesthetic analysis in comparison to semantic categorization is the highly subjective nature of aesthetics. To our knowledge, all the image datasets used for aesthetic analysis were obtained from on-line communities of photography enthusiasts such as *photo.net*[1], *DPChallenge*[2], *Flickr*[3] or *Terra Galleria*[4]. In these communities, a large number of professional and amateur photographers share, view and judge photos. These photographers also agree on the most appropriate annotation policy to score the images. Such policies can include textual labels ("like it", "don't like it") or a scale of numerical values (ratings). From these annotations, images can be labeled as being visually appealing or not. These datasets contain images as well as aesthetic judgments they received from members of the community.

Collecting ground truth data in this manner is advantageous primarily because it is an inexpensive and expedient way to obtain aesthetic judgments from multiple individuals who are generally "prosumers" of data: they produce images and they also score them on dedicated social networks. The interpretation of these aesthetic judgments, expressed under the form of numeric scores, has usually been taken for granted. The few analyses performed on such datasets have been preliminary and on a small scale [59]. Yet a deeper analysis of the context in which these judgments are given is essential. The result of this lack of context is that it is difficult to understand what aesthetic classifiers *really* model when trained with such datasets.

Additional limitations and biases of current datasets may be mitigated by performing analysis on a much larger scale than is presently done. To date, at most 20,000 images have been used to train aesthetic models used for classification and regression. In chapter 6, we describe AVA (Aesthetic Visual Analysis), a database we assembled which contains more than 250,000 images, along

[1] http://www.photo.net
[2] http://www.dpchallenge.com
[3] http://www.flickr.com
[4] http://www.terragalleria.com

with a rich variety of annotations. We investigate how this wealth of data can be used to tackle the problem of understanding and assessing visual aesthetics. The database is publicly available at `www.lucamarchesotti.com/ava`.

6.0.1 AVA and Related Databases

In addition to AVA, there exist several public image databases in current use which contain aesthetic annotations. In this section, we compare the properties of these databases to those of AVA and discuss the features that differentiate AVA from such databases. A summary of this comparison is shown in in Table 6.1.

	AVA	PN	CUHK	CUHKPQ	CLEF
Large scale	Y	N	N	N	N
Score distr.	Y	Y	N	N	N
Rich annotations	Y	N	Y	Y	Y
Semantic labels	Y	N	N	Y	Y
Style labels	Y	N	N	N	Y

Table 6.1: Comparison of the properties of current databases containing aesthetic annotations. AVA is large-scale and contains score distributions, rich annotations, and semantic and style labels.

Photo.net (PN) [25]: PN contains 3,581 images gathered from the social network *Photo.net*. In this online community, members are instructed to give two scores from 1 to 7 for an image. One score corresponds to the image's aesthetics and the other to the image's originality. The dataset includes the mean aesthetic score and the mean originality score for each image. As described in [25], the aesthetic and originality scores are highly correlated, with little disparity between these two scores for a given image. This is probably due to the difficulty of separating these two characteristics of an image. As the two scores are therefore virtually interchangeable, works using PN have restricted their analysis to the aesthetic scores. The users are provided by the site administrators with the following guidelines for judging images: *"Reasons for a rating closer to 7: a)it looks good, b)it attracts/holds attention, c)it has an interesting composition, d)it has great use of color, e)(if photojournalism) contains drama, humor, impact, f)(if sports) peak moment, struggle of athlete"*. Figure 6.1 shows sample photos of high quality with their scores and number of votes. At visual inspection of PN, we have noticed a correlation between images receiving a high grade and the presence of frames manually created by the owners to enhance the visual appearance (see examples in Figure 6.2). In particular, we manually detected that more than 30% of the images are framed. In addition to this bias, many images in PN have been scored by very few users. In fact, the images were included on the condition that they had received scores from at least two users. In contrast, each image included in AVA has at least 78 votes. In addition, AVA contains approximately 70$\times$ the number of images.

CUHK [62]: CUHK contains 12,000 images, half of which are considered high quality and the rest labeled as low quality. [62] observed the same bias for images with border as we did for PN, so they removed all the frames from the images they released. The images were obtained by retaining the top and bottom 10% (in terms of mean scores) of 60,000 images randomly crawled from `www.dpchallenge.com`. Our dataset differs from CUHK in several ways. While AVA includes more ambiguous images, CUHK only contains images with a very clear consensus on their score. As a consequence, the images in CUHK are not representative of the range of images, in terms of aesthetic quality, that one would find in a real-world application such as re-ranking images returned by a search

Figure 6.1: Photos highly rated by peer voting in an on-line photo sharing community (*photo.net*).

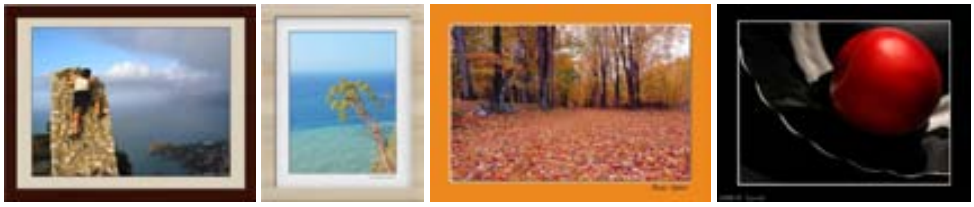


Figure 6.2: Sample images from PN with borders manually created by photographers to enhance the photo visual appearance.

on the web. In addition, CUHK is no longer a challenging dataset for classification; recent methods achieved accuracies superior to 90% on this dataset [81]. Finally, CUHK provides only binary labels (1=high quality images, 0=low quality images) whereas AVA provides an entire distribution of scores for each image.

CUHKPQ [76]: CUHKPQ consists of 17,690 images obtained from a variety of on-line communities and divided into 7 semantic categories. Each image was labeled as either high or low quality by at least 8 out of 10 independent viewers. Therefore this dataset consists of very high consensus images and their binary labels. Like CUHK, it is not a challenging dataset for the problem of binary classification: the method of [76] obtained AROC values between 0.89 and 0.95 for all semantic categories. Also like CUHK, the images in the dataset do not span the full range of images, in terms of aesthetic quality, that one is likely to find in a real-world aesthetic prediction application. In addition, despite the fact that AVA shares similar semantic annotations, it differs in terms of scale and also in terms of consistency. In fact, CUHKPQ was created by mixing high quality images derived from photographic communities and low quality images provided by university students.

MIRFLICKR/Image CLEF: Visual Concept Detection and Annotation Task 2011 [47]: MIRFLICKR is a large dataset introduced in the community of multimedia retrieval. It contains 1 million images crawled by Flickr, along with textual tags, aesthetic annotations (Flickr's interestingness flag) and EXIF meta-data. A sub-part of MIRFLICKR was used by CLEF (the Cross-Language Evaluation Forum) to organize two challenges on "Visual Concept Detection". For these challenges, the basic annotations were enriched with emotional annotations and with some tags related to photographic style.

It is probably the dataset closest to AVA but it lacks rich aesthetic preference annotations. In fact, only the "interestingness" flag is available to describe aesthetic preference. Some of the 44 visual concepts available might be related to AVA photographic styles but they focus on two very specific aspects: exposure and blur. Only the following categories are available: neutral illumination, over-exposed, under-exposed, motion blur, no blur, out of focus, partially blurred. In addition, the number of images with such style annotations is limited.

6.1 Creating AVA

AVA is a collection of images and meta-data derived from `www.dpchallenge.com`. To our knowledge, it represents the first attempt to create a large database containing a unique combination of heterogeneous annotations. The peculiarity of this database is that it is derived from a community where images are uploaded and scored in response to photographic challenges. Each challenge is defined by a title and a short description (see Fig. 6.3 for a sample challenge). Using this interesting characteristic,

Figure 6.3: A sample challenge entitled "Skyscape" from the social network `www.dpchallenge.com`. Users submit images that should conform to the challenge description and be of high aesthetic quality. The submitted images are voted on by members of the social network during a finite voting period. After this period, the images are ranked by their average scores and the top three images are awarded ribbons.

we associated each image in AVA with the information of its corresponding challenge. This information can be exploited in combination with aesthetic scores or semantic tags to gain an understanding of the context in which such annotations were provided. We created AVA by collecting approximately 255,000 images covering a wide variety of subjects on 1,447 challenges. We combined the challenges with identical titles and descriptions and we reduced them to 963. Each image is associated with a single challenge.

In AVA we provide three types of annotations:

Aesthetic annotations: Each image is associated with a distribution of scores which correspond to

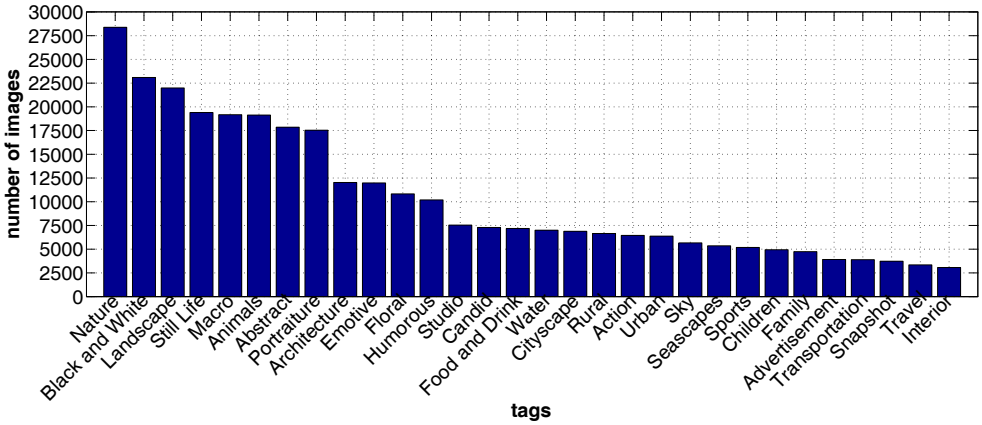


Figure 6.4: Frequency of the 30 most common semantic tags in AVA.

individual votes. The number of votes per image ranges from 78 to 549, with an average of 210 votes. Such score distributions represent a gold mine of aesthetic judgments generated by hundreds of amateur and professional photographers with a practiced eye. We believe that such annotations have a high intrinsic value because they capture the way hobbyists and professionals understand visual aesthetics.

Semantic annotations: We provide 66 textual tags describing the semantics of the images. Approximately 200,000 images contain at least one tag, and 150,000 images contain 2 tags. The frequency of the most common tags in the database can be observed in Fig. 6.4.

Photographic style annotations: Despite the lack of a formal definition, we understand photographic style as a consistent manner of shooting photographs achieved by manipulating camera configurations (such as shutter speed, exposure, or ISO level). We manually selected 72 Challenges corresponding to photographic styles and we identified three broad categories according to a popular photography manual [66]: *Light*, *Colour*, *Composition*. We then merged similar challenges (e.g. "Duotones" and "Black & White") and we associated each style with one category. The 14 resulting photographic styles along with the number of associated images are: Complementary Colors (949), Duotones (1,301), High Dynamic Range (396), Image Grain (840), Light on White (1,199), Long Exposure (845), Macro (1,698), Motion Blur (609), Negative Image (959), Rule of Thirds (1,031), Shallow DOF (710), Silhouettes (1,389), Soft Focus (1,479), Vanishing Point (674).

6.1.1 Aesthetic preference in AVA

Aesthetic preference can be described either as a single (real or binary) score or as a distribution of scores. In the first case, the single value is obtained by averaging all the available scores and by eventually binarizing the average with an appropriate threshold value. The main limitation of this representation is that it does not provide an indication of the degree of consensus or diversity of opinion among annotators. The recent work of [125] proposed a solution to this drawback by learning a model capable of predicting score distributions through structured-SVMs. However, they use a dataset composed of 1,224 images annotated with a limited amount of votes (on average 28 votes per image). We believe that such methods can greatly benefit from AVA where much richer scores distributions (consisting on average of approximately 200 votes) are available. AVA also enables us to have a deeper understanding of such distributions and of what kind of information can be deduced from them.

Score distributions are largely Gaussian. Table 6.2 shows a comparison of Goodness-of-Fit (GoF), as measured by RMSE, between top performing distributions we used to model the score distributions of AVA. One sees that Gaussian functions perform adequately for images with mean scores between 2 and 8, which constitute 99.77% of all the images in the dataset. In fact, the RMSEs for Gaussian models are rarely higher than 0.06. This is illustrated in Fig. 6.5. Each plot shows 8 density functions obtained by clustering the score distributions of images whose mean score lies within a specified range. Clustering was performed using k-means. The clusters of score distributions are usually well approximated by Gaussian functions (see Figures 6.5(b) and 6.5(c)). We also fitted Gaussian Mixture Models with three Gaussians to the distributions but we only found minor improvement with respect to one Gaussian. Beta, Weibull and Generalized Extreme Value distributions were also fitted to the score distributions, but gave poor RMSE results.

Non-Gaussian distributions tend to be highly-skewed. This skew can be attributed to a floor and ceiling effect [21], occurring at the low and high extremes of the rating scale. This can be observed in Figures 6.5(a) and 6.5(d). Images with positively-skewed distributions are better modeled by a Gamma distribution $\Gamma(s)$, which may also model negatively-skewed distributions using the transformation $\Gamma\ (s) = \Gamma((s_{min} + s_{max}) - s)$, where s_{min} and s_{max} are the minimum and maximum scores of the rating scale.

Mean score	**Average RMSE**		
	Gaussian	Γ	Γ
1-2	0.1138	**0.0717**	0.1249
2-3	0.0579	**0.0460**	0.0633
3-4	**0.0279**	0.0444	0.0325
4-5	**0.0291**	0.0412	0.0389
5-6	**0.0288**	0.0321	0.0445
6-7	0.0260	**0.0250**	0.0455
7-8	**0.0268**	0.0273	0.0424
8-9	0.0532	0.0591	**0.0403**
Average RMSE	**0.0284**	0.0335	0.0429

Table 6.2: Goodness-of-Fit per distribution with respect to mean score: The last row shows the average RMSE for all images in the dataset. The Gaussian distribution was the best-performing model for 62% of images in AVA.

Standard Deviation is a function of mean score. Box-plots of the variance of scores for images with mean scores within a specified range are shown in Fig. 6.6. It can be seen that images with "average" scores (scores around 4, 5 and 6) tend to have a lower variance than images with scores greater than 6.6 or less than 4.5. Indeed, the closer the mean score gets to the extreme scores of 1 or 10, the higher the probability of a greater variance in the scores. This is likely due to the non-Gaussian nature of score distributions at the extremes of the rating scale.

Images with high variance are often non-conven -tional. To gain an understanding of the additional information a distribution of scores may provide, we performed a qualitative evaluation of images with low and high variance. Table 6.3 displays our findings. The quality of execution of the styles and techniques used for an image seem to correlate with the mean score it receives. For a given mean value however, images with a high variance seem more likely to be edgy or subject to interpretation, while images with a low variance tend to use conventional styles or depict conventional subject matter. This is

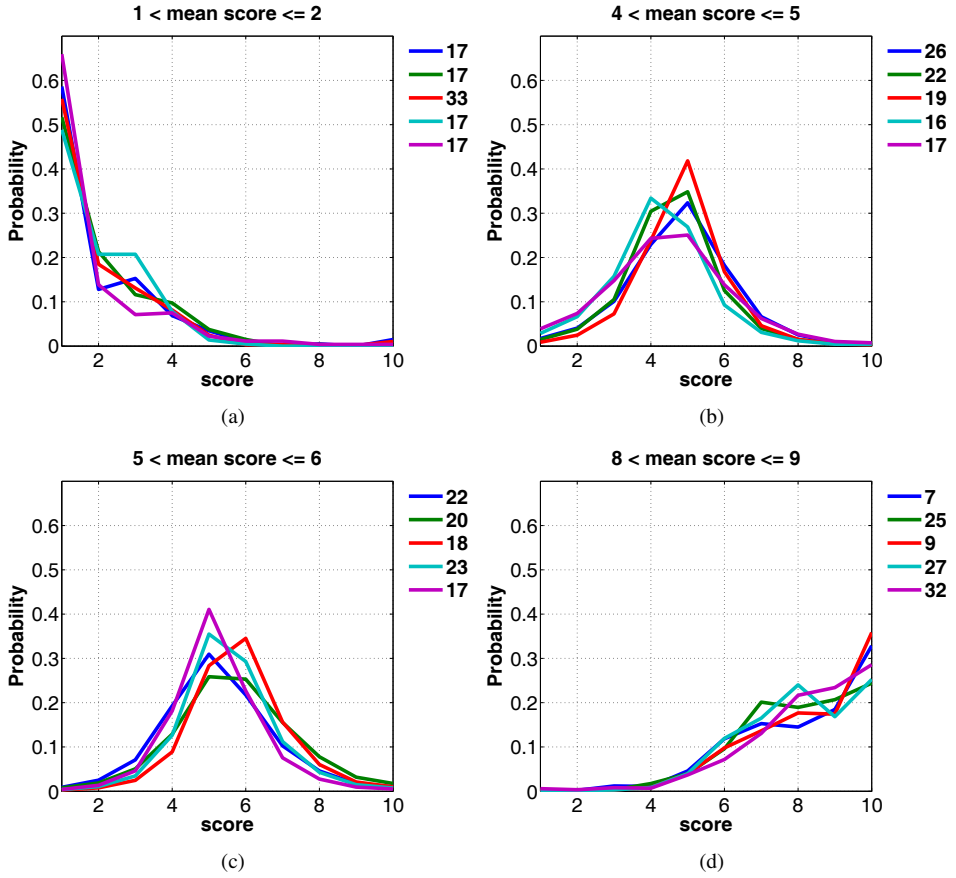


Figure 6.5: Clusters of distributions for images with different mean scores. The legend of each plot shows the percentage of these images associated with each cluster. Distributions with mean scores close to the mid-point of the rating scale tend to be Gaussian, with highly-skewed distributions appearing at the end-points of the scale.

consistent with our intuition that an innovative application of photographic techniques and/or a creative interpretation of a challenge description is more likely to result in a divergence of opinion among voters. Examples of images with low and high score variances are shown in Fig. 6.7. The bottom-left photo in particular, submitted to the challenge "Faceless", had an average score of 5.46 but a very high variance of 5.27. The comments it received indicate that while many voters found the photo humorous, others may have found it rude.

6.1.2 Semantic content and aesthetic preference

We evaluated aggregated statistics for each challenge using the score distributions of the images that were submitted. Fig. 6.8 shows a histogram of the mean score of all challenges. As expected, the mean scores are approximately normally distributed around the mid-point of the rating scale. We inspected the titles and associated descriptions of the challenges at the two extremes of this distribution. We did

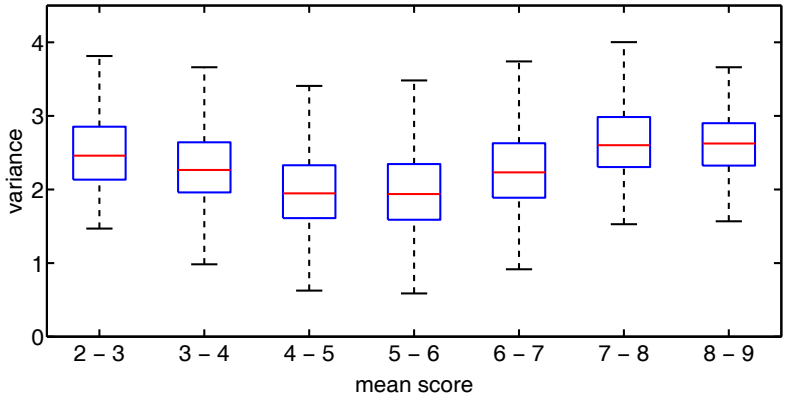


Figure 6.6: Distributions of variances of score distributions, for images with different mean scores. The variance tends to increase with the distance between the mean score and the mid-point of the rating scale.

		variance	
		low	high
mean	low	poor, conventional technique and/or subject matter	poor, non-conventional technique and/or subject matter
	high	good, conventional technique and/or subject matter	good, non-conventional technique and/or subject matter

Table 6.3: Mean-variance matrix. Images can be roughly divided into 4 quadrants according to conventionality and quality.

not observe any semantic coherence between the challenges in the right-most part of the distribution. However, it is worth noticing that two "masters' studies" (where only members who have won awards in previous challenges are allowed to participate) were among the top 5 scoring challenges. We use the arousal-valence emotional plane [104] to plot the challenges on the left of the distribution (the low-scoring tail). The dimension of valence ranges from highly positive to highly negative, whereas the dimension of arousal ranges from passive to active. In particular, among the lowest-scoring challenges we identified: #1 "At Rest" (av. vote= 4.747), #2 "Despair" (av. vote=4.786), #3 "Fear" (av.vote=4.801), #4 "Bored" (av. vote=4.8060), # 6 "Pain" (av. vote=4.818), #23 "Conflict" (av. vote= 4.934), #25 "Silence" (av. vote= 4.948), #30 "Shadows" (av. vote= 4.953), #32 "Waiting" (av. vote.=4.953), #39 "Obsolete" (av.vote= 4.9740). In each case, the photographers were instructed to depict or interpret the emotion or concept of the challenge's title. This suggests that themes in the left quadrants of the arousal-valence plane (see Fig. 6.8) bias the aesthetic judgments towards smaller scores.

We investigated the relationship between the title and description of a challenge and the mean of the variance of the score distributions of images submitted to that challenge. We found that the majority of free study challenges were among the bottom 100 challenges by variance, with 11 free studies

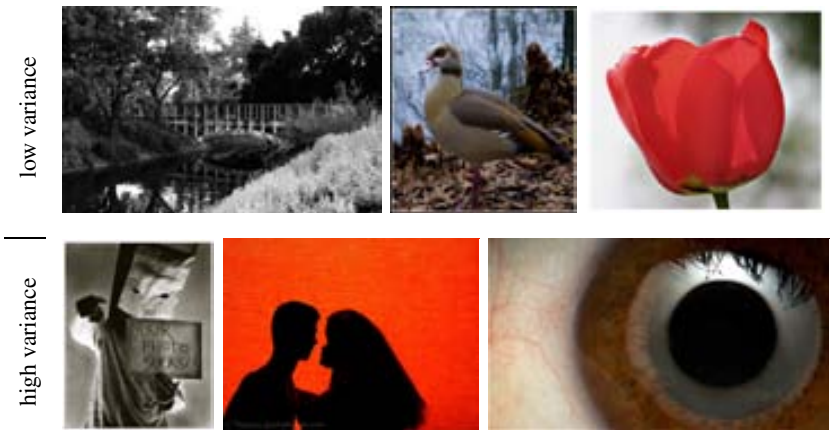


Figure 6.7: Examples of images with mean scores around 5 but with different score variances. High-variance images have non-conventional styles or subjects.

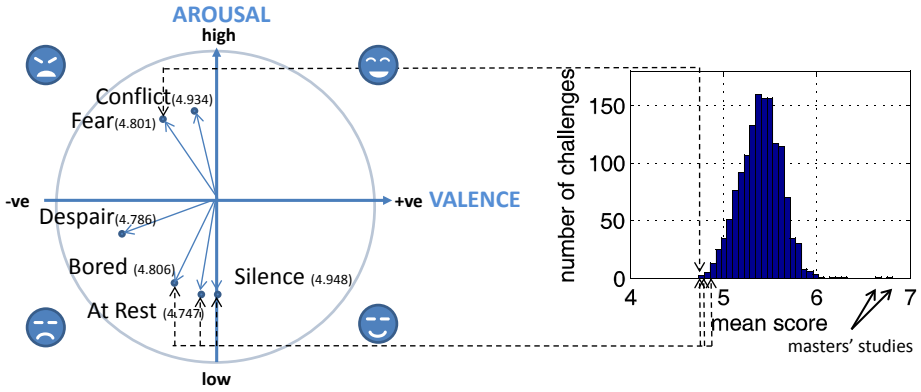


Figure 6.8: Challenges with a lower-than-normal average vote are often in the left quadrants of the arousal-valence plane. The two outliers on the right are masters' studies challenges.

among the bottom 20 challenges. Free study challenges have no restrictions or requirements as to the subject matter of the submitted photographs. The low variance of these types of challenges suggests that challenges with specific requirements tend to lead to a greater variance of opinion, probably with respect to how well entries adhere to these requirements.

6.1.3 Textual comments in AVA

Of the 255,530 images in AVA, most of them (253,903) received at least one comment from a member of the social network. There are two phases in which comments may be given. In the first phase, the challenge is ongoing and the comments and votes given to images are not yet visible to the community. In this phase, a user is allowed to give a comment to an image after giving that image a score. Comments given in this phase should therefore be unbiased with respect to the opinions of other members. In the second phase, the challenge has been completed and the results are public. Comments given in this

phase are therefore likely to be biased in at least two ways. First, images which performed well during the challenge are likely to have a greater number of comments as they are more visible, being high in the rankings for that challenge. Second, the comments given to an image in this period may be influenced by the results of the challenge and the comments it has already received.

The guidelines for commenting [126] encourage the users to leave comments when voting and, as the site focuses on improving skills, asks users to include advice for improving the work. As such, comments typically express the member's opinion on the quality of the photograph, their justifications for giving a certain score, as well as critiques of the strengths and weaknesses of the photograph. For example, the top right image in Fig. 6.7 received the following comment:

> "Like the shot. One thing I think it could be helped by is a bit more contrast, make the colors more rich and stand out that much more. I like the [square] crop...good choice."

These comments are a rich source of information about the *reasons* for which an individual may assign a particular aesthetic score to an image.

We investigated several properties of the comments given to images in AVA:

- the number of available comments;
- the commentators' activity; and
- the quality of available comments.

Number of comments: Statistics on the number and length of comments given to images are shown in Table 6.4. On average, an image tends to have about 11 comments, with a comment having about 18 words on average. However, the mean number of comments given during a challenge is greater than the mean number of comments given after. Interestingly, the length of comments given during a challenge is on average much shorter than those given after the challenge. Our observations lead us to believe that this is due to a "critique club" effect. The critique club comprises volunteer members who give a detailed critique of images which they have been assigned to review. The website states that [127]:

> "...the Critique Club critiques should be significantly longer than your average challenge comment and they should contain details about why the viewer feels a certain way about a photograph."

For an image to be critiqued, its author must request a critique when submitting the image. These critiques are then posted to the image's page *after* voting has finished. As such comments are detailed and long, they likely increase the average length of comments given after challenge completion.

As shown in Fig. 6.5, the number of comments made about an image varies significantly with respect to the mean score given to that image. Unsurprisingly, high-scoring images have a large number of comments with respect to other images. This bias is more pronounced when comparing the number of comments given during voting to the number of comments given after. Images with mean scores close to the midpoint of the rating scale tend to have very few comments, perhaps because it is difficult to form an opinion about an image that is neither clearly bad nor clearly good. However, the mean length of the comments given to such images is much higher than the global average. This may be because critique club comments are often one of the few comments given to such images, and bias the mean length towards a higher number.

Statistic	During challenge	After challenge	Overall
Mean number of comments per image	9.99	1.49	11.49
Std. dev. of number of comments per image	8.41	4.77	11.12
Mean comment length (in number of words)	16.10	43.51	18.12
Std. dev. of comment length	8.24	61.74	11.55

Table 6.4: Statistics on comments in AVA.

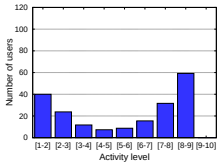
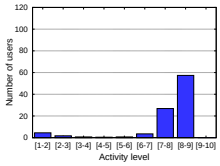
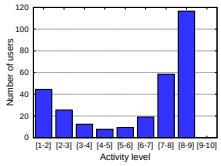
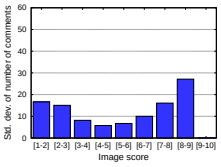
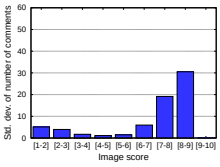
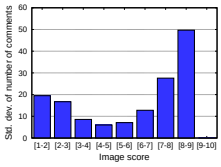
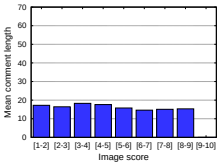
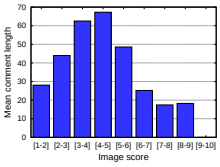
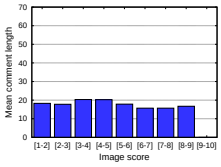
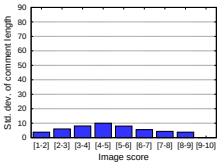
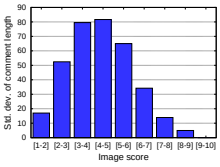
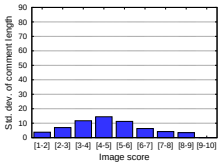
Statistic	During challenge	After challenge	Overall
Mean number of comments per image			
Std. dev. of number of comments per image			
Mean comment length			
Std. dev. of comment length			

Table 6.5: Number of comments in the AVA database and their length (in number of words) for images within the given score range.

Commentators activity: For the images in AVA, 27,557 unique members made 2,934,728 comments. Fig. 6.9 shows the commenting activity of these commentators. We found that approximately 86% of users write comments only occasionally, while the remaining 3,983 users are regular commentators who have authored at least 100 comments.

Technical content in comments: We investigated the words present in comments to determine how many comments contained technical content related to photographic techniques and aesthetic quality. We manually selected the technical words found among the 1,000 most frequently used words in the set of comments. We found 149 such words, examples of which are “exposure”, “lighting”, “vivid” and “texture”. We note that this was a non-exhaustive list of the technical terms included in the corpus of comments. Even so, we found that 77% of commments include at least one of these technical words,

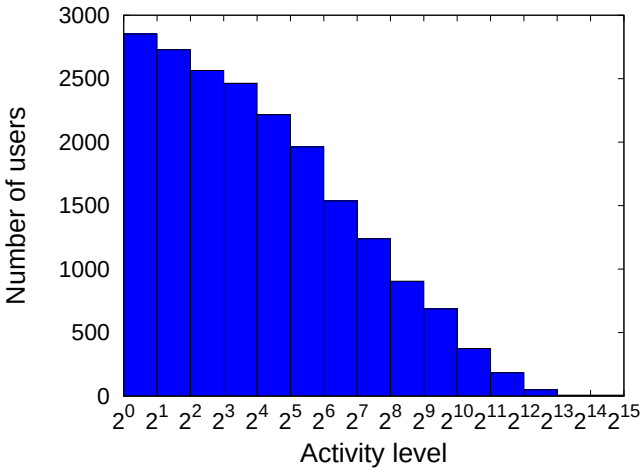


Figure 6.9: Histogram of number of users for different activity levels, where activity level is denoted by number of comments made. The activity level ranges from 1 and 24,232 comments.

and among these comments, 2.8 words were used on average.

Chapter 7

Addressing Problems in Aesthetics Prediction using the AVA Dataset

In this chapter we investigate how the wealth of data contained in AVA can be used to tackle the problem of understanding and assessing visual aesthetics by looking into several applications relevant for aesthetic analysis. These applications illustrate the advantages of the AVA dataset not only for classic problems such as aesthetic categorization, but also for gaining a deeper understanding of what makes an image appealing, *e.g.*what are the respective roles of the semantic content and the photographic technique. The applications also demonstrate how the large scale of AVA can be leveraged to improve performance on these tasks.

In section 7.1, we show the classification performance gains we achieve using a large amount of training data and a judicious selection of training data. In section 7.2 we present a scenario where AVA can be used to classify the photographic style of an image. Finally, in section 7.3 we explore in depth aesthetics-aware content-based image retrieval.

7.1 Binary aesthetic categorization

Most approaches to the problem of aesthetic categorization involve fully-supervised learning. Typically, a classification model is trained to assign "high quality" or "low quality" labels to images [31,59,62,76, 77,81]. This framework is particularly interesting because preference information is currently collected at a web-scale through binary ratings (such as Facebook's "Like" button or Google's "+1" button). However, recent works [125] have interpreted this problem as a regression problem, which is possible only if appropriate annotations are available. To investigate the performance gains afforded by the large scale of AVA, we performed categorization experiments using SIFT and Color-based Fisher Vectors (FV) [56,97]. These features were shown in [81] to give state-of-the-art performance in this task.

The FV $\mathcal{G}_\lambda^X$ characterizes a sample $X = \quad x_t \;\; t = 1 \quad \; T \;\;$ by its deviation from a distribution u_λ (with parameters λ):

$$\mathcal{G}_\lambda^X = L_\lambda G_\lambda^X \tag{7.1}$$

G_λ^X is the gradient of the log-likelihood with respect to λ:

$$G_\lambda^X = \frac{1}{T} \quad _\lambda \log u_\lambda(X) \tag{7.2}$$

L_λ is the Cholesky decomposition of the inverse of the Fisher information matrix F_λ of u_λ, *i.e.*$F_\lambda^{-1} =$

$L_\lambda L_\lambda$ where by definition:

$$F_\lambda = E_{x\sim u_\lambda}\left[\quad _\lambda \log u_\lambda(x) \quad _\lambda \log u_\lambda(x) \right] \tag{7.3}$$

Here, X is the set of T local descriptors extracted from an image and $u_\lambda = \quad _{i=1}^{N} w_i u_i$ is a GMM which models the generative process of local descriptors.

To construct the FV for an image, local patches of size 32×32 are extracted regularly on grids every 4 pixels at 5 scales. For SIFT, the local patch is divided into a 4x4 grid, and a histogram of oriented gradients in each bin of the grid is computed. Similarly, the color descriptor divides the patch into a 4x4 grid and computes simple statistics per color channel for each bin of the grid. This produces 128-dimensional SIFT descriptors [75] and 96-dimensional color descriptors [98]. Both are reduced with PCA to 64 dimensions. The probabilistic visual vocabulary, *i.e.*the GMM, is learned using a standard EM algorithm. For the FV we use a GMM with 256 Gaussians. For the spatial pyramid, we follow the splitting strategy adopted by the winning system of PASCAL VOC 2008 [35]. We extract 8 vectors per image: one for the whole image, three for the top, middle and bottom regions and four for each of the four quadrants. The pyramid was introduced by [81] with the aim of encoding information about the image composition.

Figures 7.1(a) and 7.1(b) show the learning curves with color and SIFT features respectively for a variable number of training samples and for more or less complex models. The model complexity is set by the number of Gaussians, $ngauss$, used to compute the FV as the FV dimensionality is directly proportional to $ngauss$. All the models in this chapter were learned using stochastic gradient descent (SGD) [10]. We chose to use SGD because of its scalability. As expected, for both types of features, we consistently increase the performance with more training images but with diminishing returns. Also, more Gaussians lead to better results although the difference between $ngauss = 64$ and 512 remains limited (on the order of 1%).

Reducing scale of training data by careful selection of training images: We introduce a parameter δ to discard ambiguous images from the training set. More precisely, we discard from the training set all those images with an average score between $5 - \delta$ and $5 + \delta$. As δ increases, we are left with increasingly unambiguous images. On the other hand, when $\delta = 0$, we use the full training set. This is somewhat similar to the protocol of [26, 81]. However, there is a major difference: in those works, δ was used to discard ambiguous images from the training *and the test set*, thus making the problem easier with larger values of δ. In our case, the test set is left unchanged, *i.e.*it includes both ambiguous and unambiguous images. Figures 7.1(c) and 7.1(d) show the classification results for color and SIFT descriptors respectively, as δ increases. There are two points to note. First, for the same number of training images, the accuracy increases with δ.

Second, the same level of accuracy that is achieved by increasing the number of training samples can also be achieved by increasing δ. In this way, accuracy is preserved and computational cost is reduced by selecting the "right" training images.

Generalization to other datasets: Different image datasets and photography social networks may contain images with different characteristics, due to the specific criteria for selecting images in the case of curated datasets, and because of the different community guidelines and cultures in the case of social networks. For this reason, it is reasonable to wonder how well models of aesthetic quality trained using data from one corpus generalize when applied to a different image corpus. We investigate this issue by conducting several cross-database experiments.

For these experiments, 198,000 images from AVA were selected for training aesthetic models using SIFT-based features and color-based features. These models were applied to several test sets. The first, called "Free Study", contains 22,000 images from "Free Study" challenges, which are challenges that do not have specific instructions for photographic content. The models were also tested on the CUHK

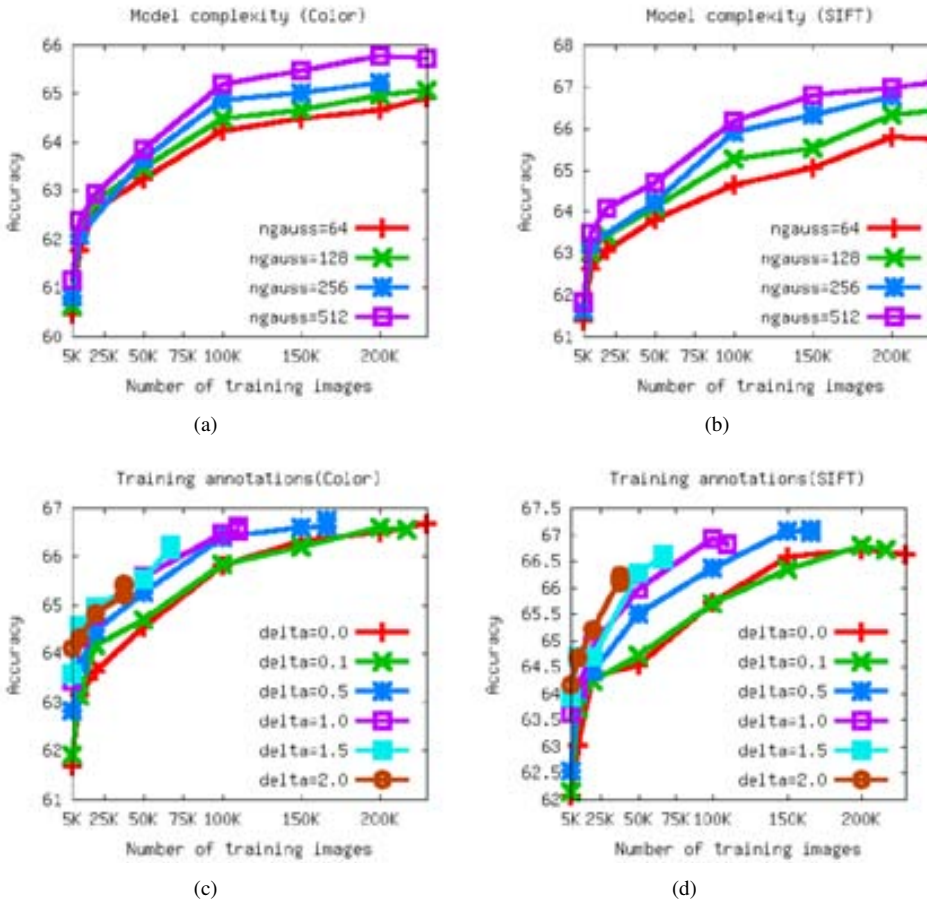


Figure 7.1: Results for large-scale aesthetic quality categorization for increasing model complexity ((a) and (b)) and increasing values of δ ((c) and (d)).

and CUHK-PQ datasets. In addition, we created a small-scale dataset of 22,000 images from photo.net, called PNSS, in order to test cross-social-network generalization performance.

We also created another training dataset from photo.net, called PNLS, containing 198,000 images, in order to investigate how models trained using images from this social network perform when applied to images from dpchallenge.com.

We performed both classification and regression experiments, using the features and optimization procedure described previously. In the case of regression, we optimized ridge regression parameters using the same SGD framework.

For classification, ground truth labels for images in AVA and PNLS were obtained by thresholding their mean scores by the global mean score across all the images in their respective training corpora. The ground truth labels for the Free Study and PNSS test databases were also obtained using the global mean score of the AVA and PNLS test databases respectively. The labels for CUHK and CUHK-PQ are provided by their distributors. Because these two databases do not include aesthetic scores, they were not included in the regression experiments. The aesthetic scores used as annotations in the regression

experiments were normalized to lie in the range -1 to 1.

Results for classification experiments are shown in Table 7.1. We note two main findings. First, models trained on both the AVA and PNLS training sets generalize well to test images from different social networks. Second, AVA-trained models generalize better to the PNSS dataset (in the sense that their performances are closer to those of PNLS-trained models) than PNLS-trained models generalize to the three dpchallenge.com-derived test sets. These findings also hold for the regression results, shown in Table 7.2.

Feature	Train Set	Test Set			
		Free Study	CUHK	CUHK-PQ	PNSS
ORH	AVA	68.8185	71.5833	75.6652	66.0455
	PNLS	66.0815	72.5417	75.2038	65.3046
COL	AVA	68.3074	72.1417	76.4870	64.8959
	PNLS	65.1593	71.6917	74.7861	64.6617
ORH+COL	AVA	70.7296	73.6833	78.0229	67.1461
	PNLS	67.7370	74.8333	77.3324	66.3888

Table 7.1: Cross-dataset classification experiments using different features: accuracy (in %).

Feature	Train Set	Test Set	
		Free Study	PNSS
ORH	AVA	0.0154	0.0182
	PNLS	0.0310	0.0143
COL	AVA	0.0154	0.0182
	PNLS	0.0309	0.0145
ORH+COL	AVA	0.0142	0.0175
	PNLS	0.0294	0.0138

Table 7.2: Cross-dataset regression experiments using different features: Mean Squared Error (MSE).

7.2 Style Categorization

When asked for a critique, experienced photographers not only say *how* much they like an image. In general, they also explain *why* they like or dislike it. This is the behavior that we observed in social networks such as `www.dpchallenge.com`. Ideally, we would like to replicate this qualitative assessment of the aesthetic properties of the image. This represents a novel goal that can be tackled using the style annotations of AVA.

To verify this possibility, we trained 14 classification models using the 14 photographic style annotations of AVA and their associated images (totaling 14,079). We trained 14 one-versus-all linear SVMs

using SGD. We computed separate FV signatures using SIFT, color histogram and LBP (Local Binary Patterns) features and combined them by late fusion.

Results are summarized in Figure 7.2. Not surprisingly, the color histogram feature is the best performer for the "duotones", "complementary colors", "light on white" and "negative image" challenges. SIFT and LBP perform better for the "shallow depth of field" and "vanishing point" challenges. Late fusion significantly increases the mean average precision (mAP) of the classification model, leading to a mAP of 53.85%. The qualitative results shown in Figure 7.3 illustrate that top-scored images are quite consistent with their respective styles, even while their semantic content differed.

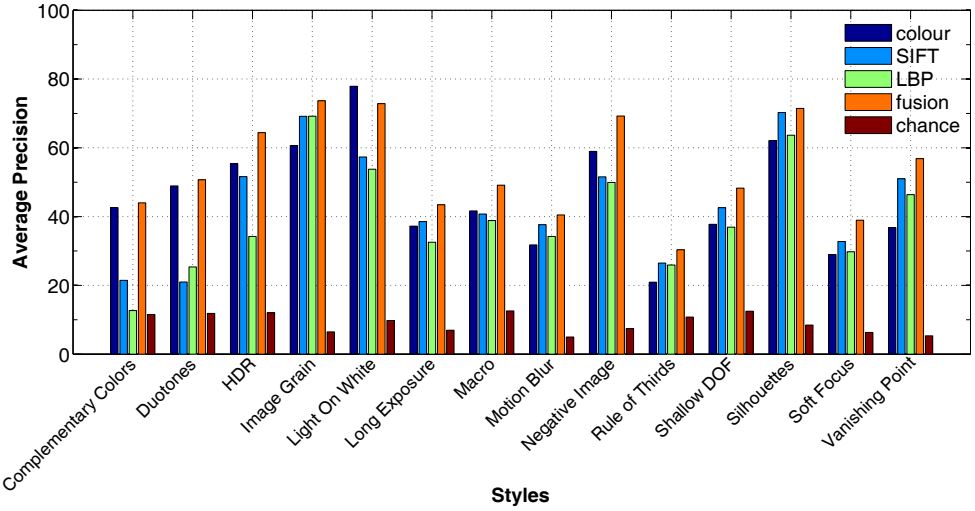


Figure 7.2: Mean average precision (mAP) for challenges. Late fusion results in a mAP of 53.85%.

7.3 Combined Semantic and Aesthetic Retrieval

Semantic retrieval is currently perceived by users as a commoditized feature of multimedia search engines. This is confirmed by a recent user evaluation [40] performed to determine the key differentiating factors of an image search engine. The top five factors were reported to be: "High-quality" (13%), "Colorful" (10%), "Semantic Relevance" (8%), "Topically clear" (7%) and "Appealing" (5%). Semantic relevance is only ranked as the third factor, whereas features related to the quality and aesthetics rank first and second. For this reason, the ability to assess the aesthetic quality of an image is an increasingly important differentiating factor for search engines. This has lead to recent interest in methods for retrieving images which are *both relevant and aesthetically pleasing* in response to a semantic (textual query). In [105], textual and visual features are used to predict the aesthetic scores of images retrieved using textual queries. The retrieved images are then re-ranked by the sum of their aesthetic score and their query relevance score. Geng *et. al* [40] propose to train a ranking-SVM using visual, textual and contextual features. Like [105], textual features are used for determining semantic relevance. For a given query, [40] enforces relevant high-quality images to rank higher than relevant low-quality images which should themselves rank higher than irrelevant images (whatever their quality). See their section 7.2 for more details. We believe that a significant limitation of this approach is that the model mixes both sources of variability (semantic and aesthetic), thus making the job of the ranker significantly more

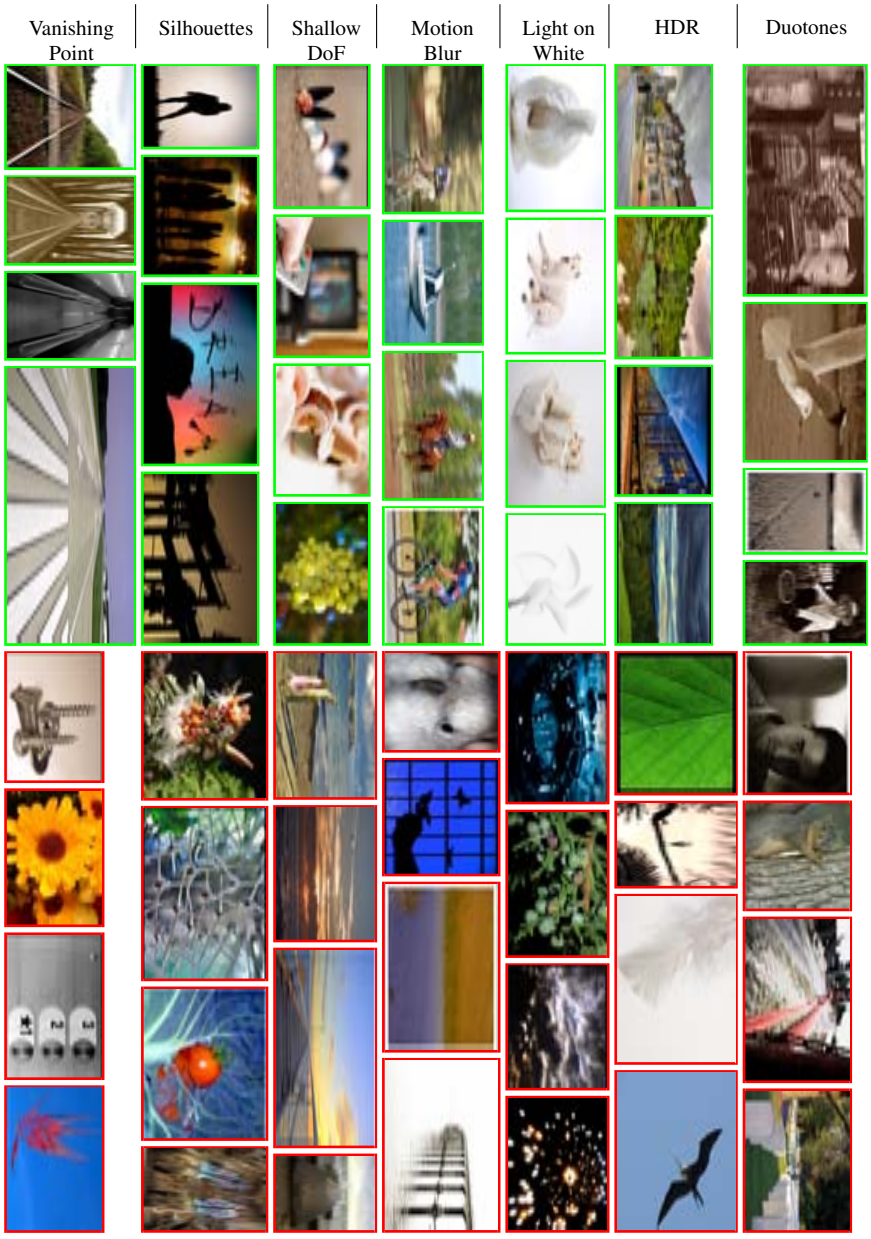


Figure 7.3: Qualitative results for style categorization. Each row shows the top 4 (green) and bottom 4 (red) ranked images for a category. Images with very different semantic content are correctly labeled.

difficult.

In this chapter, we demonstrate that the heterogeneous annotations in AVA can be used, in conjunction with low-level visual features, to learn models for ranking images by both aesthetic quality and semantic relevance. We advocate models which treat these two sources of variability separately. In addition, we do not assume the availability of textual features to score the semantic relevance of a new image.

We make three main contributions:

- Through a statistical analysis, we show that aesthetic rankings cannot be directly inferred from crowd-sourced aesthetic scores and we provide a strategy to derive meaningful relevance levels from these scores.
- We show that the ranking approach of [40] can be significantly improved by an appropriate re-weighting of the training samples inspired by the re-weighting of positive and negative examples when learning binary classifiers.
- We propose two simple models which, as opposed to [40], separate the semantic and aesthetic components. In the case of the first model, the aesthetic part is independent of the semantic part while in the second case, the aesthetic part depends on the semantic part.

Our experimental results demonstrate that it is preferable to train separate components for semantics and aesthetics rather than include them into a single model.

This chapter is organized as follows: in section 7.3.1 we describe the data we use for learning and evaluation. In sections 7.3.2 and 7.3.3 we describe and evaluate the three approaches for learning to rank images using aesthetic and semantic labels. Lastly, we provide a qualitative analysis of the results in section 7.3.4

7.3.1 Extracting heterogeneous annotations from AVA

To perform supervised learning of a model of both semantics and aesthetics, training images require annotations for both these types of labels. AVA contains such images for a large number of images.

Semantic labels. Semantic information is available in the form of textual tags (at most 2 per image) and from the textual description of each challenge. Tags are assigned by photographers while challenges are created by the website moderators. To have an idea of the kind of semantic information that can be deduced from AVA, we manually inspected the textual description and title of each challenge. We discovered that most of the challenges are dedicated to themes (e.g. vintage, spooky, Halloween), concepts (e.g. poverty, trance), or photographic techniques (e.g. rule of thirds, macro, high dynamic range). Semantic categories are present in a smaller amount. In addition, the variety of semantic subjects is limited, as well as the number of images per challenge. Because of these limitations, we used the semantic information present in the form of the 33 textual tags listed in the horizontal axis of Fig. 7.4. On average, 8,000 images are available for each tag.

Aesthetic labels. Each image in AVA is associated with a distribution of scores in a pre-defined range (1=lowest score, 10=highest score) that we normalized between -1 and 1. We averaged the distributions of scores per semantic tag and obtained the box-plots in Fig. 7.4. As can be seen, such averaged distributions are rather stable across the various semantic tags. However, we are confronted with a fundamental problem: how to represent the aesthetic information compactly and efficiently. The objective is to find a representation suitable for learning different types of statistical models (such as discriminative classifiers or rankers).

A reasonable representation would be to derive binary labels ($High-quality$ and $Low-quality$) from the mean scores of images. However, deciding on a threshold for binarization is non-trivial. Following a common approach in computer vision we could interpret classification as a retrieval problem. This decision would ultimately lead to the definition of image ranks as ground truth. Since we have scores distributions associated with each image, a natural approach to derive such ranks would

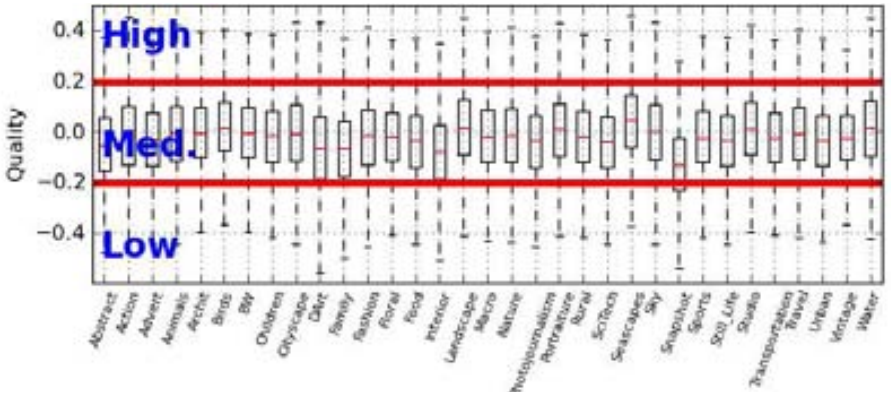


Figure 7.4: Mean distributions of scores for AVA images labeled with the 33 textual tags. Two thresholds define the aesthetic labels used to train the aesthetic models.

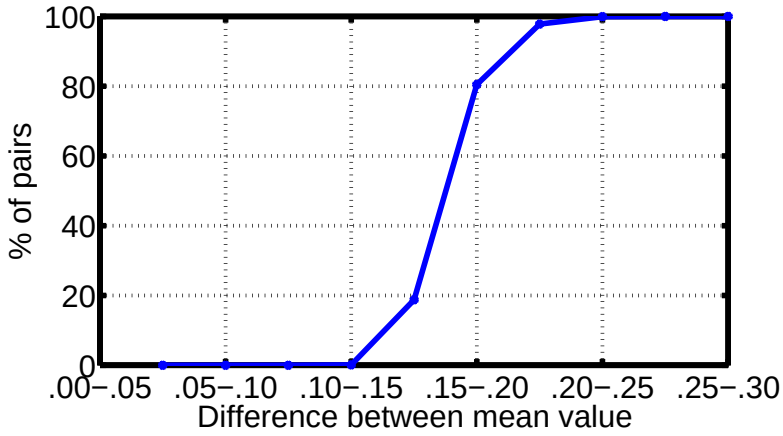


Figure 7.5: % of pairs with statistically significant differences in mean scores as a function of difference in mean score.

be to sort the images using their mean score. Such a ranking would assume that the difference between the mean scores of a pair of images, termed $\Delta_{i,j}$, is statistically significant.

To test the validity of this assumption, we sorted all images in AVA by their mean scores and applied two-sample t-tests to adjacent images. For each pair, the null hypothesis was that the means of the score distributions of the images were equal. We assumed the distributions to be normally distributed, which is a fair assumption as described in [86]. We also assumed that an image's votes are independent of each other, which is also fair as a user is not shown the votes already submitted for an image prior to voting. Lastly, the variances of the distributions were assumed to be unequal. We found that it is not a good option to use ranks derived from sorting mean votes. In fact, none of the $\Delta_{i,j}$ values for adjacent pairs in such a rank are statistically significant at the 10% significance level. As can be seen from Fig. 7.5, $\Delta_{i,j}$ should be set around .20 to generate statistically significant pairs. Therefore, we opted for an annotation strategy involving three labels: $High-quality$, $Medium-quality$, $Low-quality$. A simple thresholding operation is performed on the mean of the original votes to define for each image one of the three labels. A very small amount of image pairs picked around these

thresholds are not statistically significant, but this does not impact the performance of our model. We believe that using three labels to represent aesthetic quality is a good compromise between using the mean scores and using binary labels.

7.3.2 Experimental protocol

We experiment with the images in AVA that are associated with the textual tags listed in Fig. 7.4. These images were split into 5 folds, with images being evenly distributed over the folds according to their semantic tags (training, validation and test lists will be made available on-line for those interested in reproducing our results). Three folds were used for training, one fold was used for validation, and one fold was used for testing. The models were trained 5 times, with folds being switched in a round-robin fashion so that every fold was used as the validation and the test fold exactly once. The results we present are the average over the five folds.

Features. Each image is described using the Fisher Vector (FV) described in section 5.1. Specifically, we extract low-level SIFT descriptors [75] from 32x32 patches on dense grids every 4 pixels at 5 scales. The 128-D SIFT descriptors are reduced with PCA to 64-D. The Gaussian Mixture Model (GMM) is learned using a standard EM algorithm. We experimented with various vocabulary sizes (different numbers of Gaussians, typically between 16 and 256). Note however that the models we will benchmark are independent of the image descriptors.

Measures of performance. We report the normalized Discounted Cumulative Gain (nDCG), Precision and mean Average Precision (mAP). We focus on nDCG and Precision at 10, 20 and 50 as, in a real world application, it is more important to have accurate results among the top ranked images (typically the ones fitting in the first two or three pages of a search engine result). We also plot mAP calculated on the whole image ranking. We report nDCG@K averaged over all semantic tags. nDCG@K was computed as:

$$nDCG@K = \frac{DCG@K}{IDCG@K};\ DCG@K = \sum_{i=1}^{K} \frac{2^{rel_i} - 1}{log_2(1+i)} \tag{7.4}$$

where rel_i is the relevance level of the image at rank position i and IDCG@K is the DCG@K for a perfect ranking. mAP was computed as the mean, over the semantic tags, of the precision averaged over the set of evenly spaced recall levels 0 0 0 1 0 2 1 0 . To compute mAP, images with a relevance level of 3 (semantically relevant images with high aesthetic quality) were considered relevant.

7.3.3 Retrieval Models

We assume that we have a training set of N images $\mathcal{I} = (x_i\ y_i\ z_i)\ i = 1 \quad N$ where $x_i \quad \mathcal{X}$ is an image descriptor, $y_i \quad \mathcal{Y}$ is a semantic label and $z_i \quad \mathcal{Z}$ is an aesthetic label. In what follows, we assume that $\mathcal{X} = \mathcal{R}^D$ is a D-dimensional descriptor space, $\mathcal{Y} = 0\ 1\ ^C$ is the space of C semantic labels (where $y_{i,c} = 1$ indicates the presence of semantic class c in image i), and $\mathcal{Z} = 1 \quad K$ is the set of K aesthetic labels. In our case we have $K = 3$, where 3= $High - quality$, 2= $Medium - quality$ and 1= $Low - quality$. A major difference between spaces $\mathcal{Y}$ and $\mathcal{Z}$ is that there is a natural order on $\mathcal{Z}$. Given a semantic query specified by a class c (e.g. $c =$ "Cat"), a traditional retrieval system would compute and rank the set of image descriptors x according to their relevance $p(y_c = 1\ x)$. The problem we are investigating here is the design of a retrieval mechanism returning *high-quality* images which are also *semantically* relevant. We would also like semantically-relevant but medium-quality images to be ranked before low-quality images, as this ordering will be beneficial for classes with few high-quality images. Hence, we want to estimate $p(y_c = 1\ z > \theta\ x)$, where θ is some threshold on the aesthetic labels. Rather than set θ, we will rank images using ranking functions trained with aesthetic labels.

We first review the approach of [40] which consists of training a single ranker that learns simultaneously the semantics and aesthetics. We outline its limitations and then propose two models which learn separate semantic and aesthetic models.

The joint ranking model (JRM)

Original model. This approach was first proposed in [40]. Because we do not assume the availability of textual features, the approach of [40] translates to training one ranker per class in our case. Each semantic class is treated independently in which case the label set can be simplified to $\mathcal{Y} = \ 0\ 1\ $, i.e. semantically irrelevant or relevant. A new set of labels denoted u_i is then defined as follows: $u_i = y_i z_i$. We have $u_i \quad \mathcal{U} = \ 0\ 1 \qquad K\ $. Hence $u = 0$ means that the image is irrelevant, $u = 1$ means that the image is relevant and that its quality is the poorest possible and $u = K$ means that the image is relevant and has the highest possible quality. [40] proposes to learn a linear classifier which ranks images according to this new label u. For this purpose they train a ranking SVM as proposed for instance in [58]. Let us denote by $(x^+\ u^+)$ and $(x^-\ u^-)$ a pair of images together with their semantic and aesthetic labels in $\mathcal{U}$ such that $u^+ > u^-$. **JRM** learns w such that $w\ \ x^+ > w\ \ x^-$. This can be done by minimizing the following regularized loss function:

$$\min_{(x^+,u^+),(x^-,u^-):u^+>u^-} \max\ 0\ \Delta(u^+\ u^-) - w\ \ (x^+ - x^-)\ \ + \frac{\lambda}{2}\ \ w\ \ ^2 \tag{7.5}$$

where $\Delta(u^+\ u^-)$ encodes the loss of an incorrect ranking, for instance $\Delta(u^+\ u^-) = u^+ - u^-$. One ranker w_c is learned for each class $c = 1 \qquad C$.

Data rebalancing. JRM has an ambitious task: simultaneously learn aesthetics *and* semantics. In this case, the ranker has to deal with 4 relevance levels (the three aesthetic labels, and the semantic irrelevance level). As can be seen in Fig. 7.7, labels are very imbalanced. In particular, for the "Nature" category, the probability of one of the images in a randomly-chosen pair having relevance level 0 is more than 98% (for the other classes we observed similar trends). Therefore, virtually all pairs used to train the **JRM** model encode semantic differences, rather than aesthetic information. Correcting for data imbalances has been explored extensively for multi class categorization but little, if anything, has been done for data imbalances in ranking problems with multiple relevance levels.

We implemented the following rebalancing strategy: first, we randomly draw a pair of images $(i\ \ j)$ subject to $u_i = u_j$. Then we simply multiply the probability $p_i(u)$ of drawing an image i with relevance level u_i by the probability of drawing an image j with relevance u_j. The inverse of this value is the weight:

$$\mathcal{W}_{i,j} = [p_i(u = u_i) \cdot p_j(u = u_j)]^{-1} = \left(\frac{N_{u_i}}{N_T} \cdot \frac{N_{u_j}}{N_T - N_{u_i}}\right)^{-1} \tag{7.6}$$

where N_T is the total amount of *training* images and N_{u_i},N_{u_j} the number of images with relevance level u_i and u_j. At iteration t of the SGD optimization, the $\mathcal{W}_{i,j}$ weight for the sample pair is applied to the update term and suppresses the amount by which the model is updated, for frequently-occuring pairs. With this weighting, highly probable relevance pairs, such as $(0\ \ 2)$, are strongly penalized.

Results. In Table 7.6, we shows precisions at differing ks with and without rebalancing for **JRM**. It is not completely surprising that **JRM** without rebalancing performs similarly to a semantic classifier. In fact, pairs showing the ranker differences between high and low quality images are very rare. Most pairs train the ranker to discriminate between the various semantic classes. With rebalancing we greatly improve the performance since aesthetically relevant pairs are given more importance. These results will serve as a baseline for the two models we introduce in the next subsection.

	nDCG(k)			mAP
METHOD	k=10	k=20	k=5	
Semantic class. only	0.230	0.227	0.224	5.810
JRM	0.234	0.228	0.217	5.602
JRM-rebalanced	**0.253**	**0.244**	**0.227**	**6.980**
	Precision(k)			
	10	20	50	
Semantic class. only	8.538	8.284	8.270	
JRM	8.760	8.254	7.762	
JRM-rebalanced	**14.272**	**13.104**	**11.574**	

Figure 7.6: Results with and without data rebalancing.

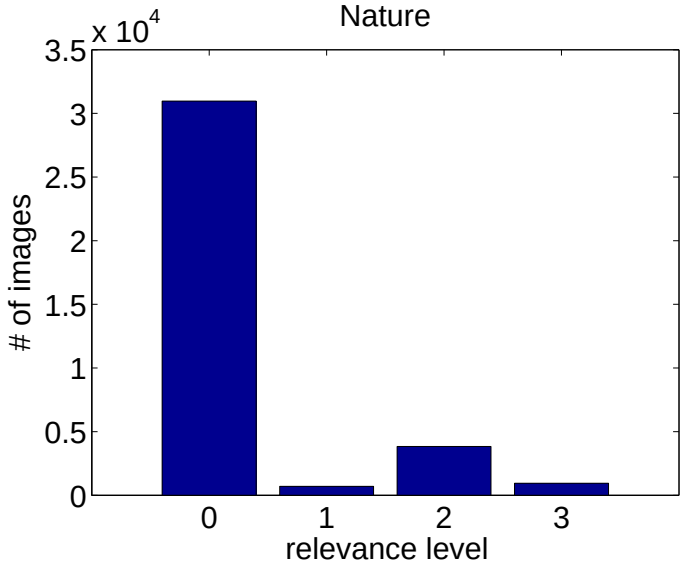


Figure 7.7: Distribution of relevance levels for the "Nature" category.

Separating semantics and aesthetics

We believe that a major weakness of the **JRM** is that it confounds both sources of variability: semantics and aesthetics. This makes the task of the linear SVM ranker more difficult. Instead, we advocate models which treat semantic and aesthetic separately.

Independent Ranking Model (IRM). The simplest strategy one can think of to model aesthetic and semantic information is the **IRM** of Figure 7.8. It consists of training a set of semantic classifiers (one per class) and a single class-independent aesthetic ranker capable of learning differences in quality between pairs of images.

The underlying assumption is to consider these two sets of labels as *independent*:

$$p(y\ z\ x) = p(y\ x)p(z\ x) \tag{7.7}$$

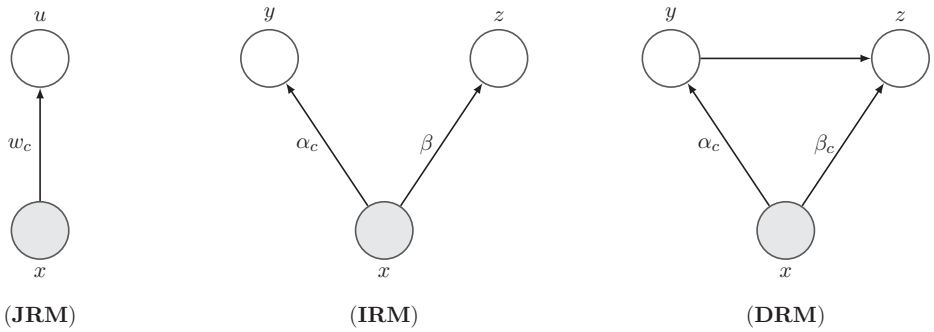


Figure 7.8: The three learning models we evaluate. **JRM** models semantics and aesthetics **jointly**, whereas **IRM** and **DRM** learn two separate models with different dependence assumptions.

For the semantic part, we learn a multi-class classifier. We use the popular strategy which consists of learning a set of one-vs-rest binary classifiers independently. We learn one linear classifier with parameters α_c per class, using the set $(x_i\ y_i)\ i = 1 \quad N$. We use a logistic loss:

$$-\log p(y_c = 1\ x) = \log\Big(1 + \exp(-\alpha_c\ x)\Big) \tag{7.8}$$

The semantic parameters α_c are learned by minimizing the (regularized) negative log-likelihood of the data on the model, which leads to the traditional logistic regression formulation:

$$-\sum_{i=1}^{N} \log p(y_{i,c}\ x) + \frac{\alpha_c\ ^2}{2} \tag{7.9}$$

As a rule of thumb, the logistic loss gives results which are similar to the hinge loss of the SVM but the former option has the advantage that it provides directly a probability estimate.

For the aesthetic part, we learn a class-independent aesthetic ranker on the set $(x_i\ z_i)\ i = 1 \quad N$. Let us denote by $(x^+\ z^+)$ and $(x^-\ z^-)$ a pair of images with their aesthetic labels in $\mathcal{Z}$ such that $z^+ > z^-$. We learn the aesthetic parameters β by minimizing the following regularized loss:

$$\sum_{(x^+,z^+),(x^-,z^-):z^+>z^-} \log[1 + \exp(-\beta\ \ (x^+ - x^-))] + \frac{\lambda}{2}\ \ \beta\ ^2 \tag{7.10}$$

We then use a sigmoid fit to transform the score into a probability estimate $p(z > \theta\ x)$.

Dependent Ranking Model (DRM). In this model, following the lessons of [31, 76] (see also introduction), we introduce an explicit dependence of the aesthetic labels on the semantic labels:

$$p(y\ z\ x) = p(y\ x)p(z\ y\ x) \tag{7.11}$$

We train one-vs-rest binary semantic classifiers independently for each class, as was the case for the **IRM** model. However, as opposed to the **IRM**, to model the dependence of aesthetics on semantics, we train one aesthetic ranker per class independently. The loss we optimize is the same of the **IRM** (see equation 7.10). The only difference is that for class c we learn a ranker with parameters β_c using only the images of this class. As was the case for the **IRM**, we use a sigmoid fit to transform the ranker output score into a probability estimate: $p(z > \theta\ y_c = 1\ x)$.

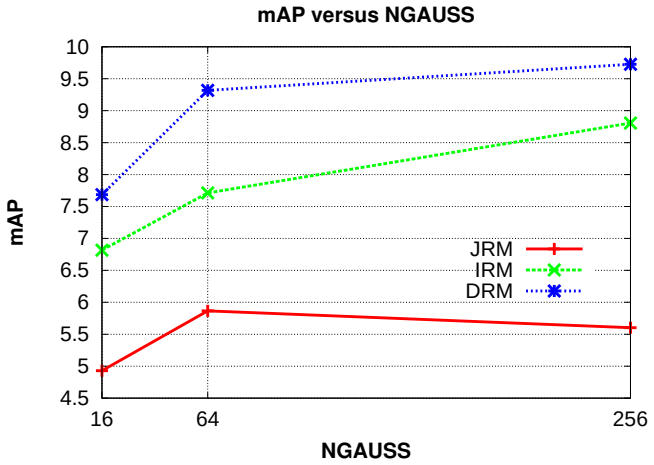


Figure 7.9: Performance with different visual vocabulary sizes.

	mAP	Precision@K			nDCG@K		
METHOD		K=10	K=20	K=50	K=10	K=20	K=50
JRM	5.602	8.760	8.254	7.762	0.234	0.228	0.217
IRM	8.806	18.128	17.000	15.450	0.255	0.247	0.236
DRM	**9.726**	**20.992**	**19.912**	**17.444**	**0.295**	**0.285**	**0.265**

Table 7.3: Comparison between the three learning strategies

Results. Table 7.3 shows a comparison between the three methods we propose. They measure the performance in terms of nDCG, mAP and Precision at K. The best performance is achieved by **DRM**. **IRM** performs slightly better than **JRM**. The advantage of **DRM** is consistent over the three measures. Worth noticing is that on this database, a baseline implemented using a discriminative semantic classifier, already performs rather well in retrieving relevant high-quality images at the top of the rank. This may be due to the fact that good quality images are highly discriminative for their semantic category.

However, as the mAP results show, the difference in performance is more marked if the whole rank of images is taken into account for each semantic tag. We also evaluate the impact of the model complexity by varying the visual vocabulary size (number of Gaussians). As can be seen in Fig. 7.9, a good trade-off between computational complexity (at training time) and performance is achieved by selecting $N = 64$ Gaussians. In fact performances reach a plateau after $N = 64$.

In Fig. 7.10 we present a breakdown of the results (nDCG@20) for each semantic tag in order to understand where content-dependence is most beneficial. From this graph we can draw some conclusions. First, **DRM** provides the best results for 15 semantic tags. For most of the other tags it is outperformed only by a small margin. Second, content dependence seems to help more for the semantic tags that are easier for the semantic classifier to learn. Data-rebalancing experiments were also performed for **IRM** and **DRM** but no significant difference was found. This is expected because for **IRM** and **DRM**, separate aesthetic ranking models are trained using only relevance levels 1,2 and 3 which are much less unbalanced.

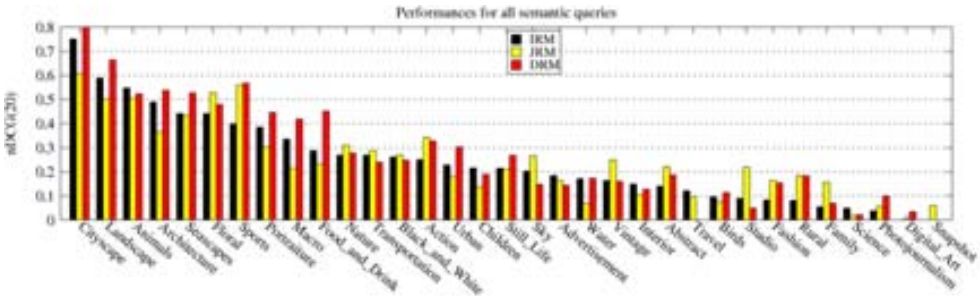


Figure 7.10: Performances measured with nDCG@20 for all semantic tags for the three models.

7.3.4 Qualitative analysis

To have a better understanding of the quantitative results outlined above, we also conducted a qualitative analysis. We inspected the ranking results for several semantic queries based on the performances outlined in Fig. 7.10. In particular, we selected ranks with high, medium and low performance. The retrieved images for some of these queries are shown in Fig. 7.11. For each selected rank we plotted the top K images ranked using a semantic 1-vs-all classifier and **DRM**. The ground truth relevance levels are represented for each image by a colored image border (green="semantically relevant *and* high quality", yellow="semantically relevant *and* medium quality", red="semantically relevant *and* low quality", black ="semantically non relevant").

The first conclusion that we can draw is that, as expected, using **DRM** we improve the retrieval results for those semantic tags that are easy to learn. Next, it can be noticed that no low quality images are retrieved by **DRM**. This is a positive result since we certainly do not want to return low quality images in the top rank. Another observation is that most of the images with black borders ("semantically non relevant images") have a visual content which is indeed representing the semantic tag for which the image was retrieved (aside from some examples in the "Birds" category). This means that the labels in the AVA database contain many false negatives, and that semantic classification is robust at the top of each rank.

Figure 7.11: Ranking results: For each tag, the top row shows results for DRM and the bottom row shows results for the baseline semantic classifier.

Part III

Unified Approach and Conclusions

Chapter 8

Aesthetics Estimation using a Low-level Vision Front-end

As described previously, most work on aesthetic visual analysis in the computer vision community has focused on designing features which explicitly capture photographic rules and techniques used by skilled photographers. These features may attempt, for example, to detect the presence of a "rule-of-thirds" composition, or a shallow depth-of-field. Features have also been designed to capture low-level image data. Datta *et al.* [25] used Daubechies wavelet coefficients to construct a feature representation of local texture. In addition, Marchesotti *et al.* [81] showed that generic low-level features, such as SIFT-based features or features based on color histograms, perform at least as good as "hand-crafted" aesthetic features.

The success of these low-level features, which are based on local texture or gradient information, are unsurprising given that image contrast, color composition, clarity and complexity are known to be important factors in visual aesthetics [82, 102, 103, 121]. Reber *et al.* [103] found then when viewers were asked to rate (on a scale from 1 to 9) the "prettiness" of light circles on a black background, or dark circles on a white background, increasing the contrast between the circle and the background led to a higher average "prettiness" rating for the circle. Wallraven *et al.* [121] found that low-level information such as color distribution was used by observers when evaluating paintings for an aesthetic rating task. Interestingly, they also found that saliency estimations given by two saliency models were fairly well correlated with the eye-fixations of observers when engaged in aesthetic appraisal of the artworks. Massaro *et al.* [82] investigated bottom processes evoked by color and dynamism, finding that for paintings without human subjects, color and dynamism increased the preference ratings they were given by observers. They surmised that this was due to color enhancing an image's dynamism and complexity in nature scenes (without human subjects). In psychology, aesthetic appreciation has come to be viewed as a multi-stage process where both top-down and bottom-up factors come into play [23, 71]. Leder *et al.* [71] proposed a conceptual model of aesthetic appreciation and judgments. In the first stage, the visual input is analysed with respect to bottom-up features such as complexity, contrast, symmetry, order and grouping. However, this model is yet to be experimentally validated.

The success of low-level processes and features in explaining aspects of aesthetic experience also provides supportive evidence for certain theories found in the nascent field of "neuroaesthetics" [19]. This field, pioneered by Zeki [61, 132] and Ramachandran [101], studies the neuro-biological underpinnings of human aesthetic appreciation. In [101], the authors present a theory of aesthetic experience based on neural mechanisms. They proposed eight "laws of aesthetic experience", several of which involved processes occuring in early vision, including perceptual grouping, contrast extraction and feature isolation.

Research into neuroaesthetics suggests that aesthetic appreciation, like visual attention, is the result of interactions between bottom-up perceptual mechanisms and top-down semantic and task-driven cues [24, 82]. Cupchik *et al.* [24] used fMRI to compare brain region activation patterns during aesthetic viewing with the patterns produced while performing an object identification task. They found that lateral prefrontal cortex, an area associated with top-down control of cognition, and left superior parietal lobule, associated with bottom-up feature processing, were activated in both viewing conditions, although to different extents. In [11], Brown *et al.*concluded that aesthetic appraisal was the result of activation in reward circuits in the brain, which in turn receive multisensory inputs, including from vision areas.

Therefore, research in computer vision, psychology, and neuroaesthetics have amassed significant evidence of the influence of bottom-up features, including color, salience, and contrast, and their corresponding neural processing mechanisms, in aesthetic experience. As SIM measures local contrast and feature isolation in conjunction with color, it is fair to entertain the hypothesis that these measures may contain information on the aesthetics of an image. In this chapter, we test this hypothesis by using the induction weights described in chapter 3 to construct feature vectors which we use to represent the aesthetics of an image. In doing this, we make 3 main contributions:

- we propose an image descriptor for aesthetics which achieves a good balance between compactness and discriminative power;
- we introduce a new color space which affords a more detailed representation of the color content present in the image. This detail is crucial as the aesthetics of an image is highly dependent on its color composition;
- we demonstrate that a biologically-inspired model of local saliency, itself derived from a model of color perception, may be used to extract image characteristics that describe image aesthetic quality.

The success of these image features adds to the evidence for common bottom-up mechanisms for different visual tasks.

The rest of this chapter is organized as follows: we describe related feature representations in section 8.1. We then describe our feature representation and its performance in sections 8.2 and 8.3 respectively. Lastly, we analyse qualitative and quantitative results in section 8.4.

8.1 Related Work

Wavelet-based image descriptors have a long history in image processing and computer vision. Kundu & Chen [68] applied a quadrature mirror filter bank to an image, then computed statistical, correlational and other features. These features were then grouped and used to train a texture classifier. Chang & Kuo [17] used a tree-structured wavelet transform to successively decompose image subbands having a certain minimum average energy. This energy was computed as the mean absolute value of the coefficients in the subband. The energies in different subbands constituted features which were then used to train a texture classifier. Liang & Kuo [74] used the number of significant coefficients in an image subband as a feature. Coefficients were significant if they exceeded a pre-determined threshold. They constructed texture, color and shape descriptors for an image using the normalized sum of significant coefficients for each subband. Van de Wouwer *et al.* [119], like Chang & Kuo, used subband energy to characterize texture. In addition, they introduced two feature vectors, one of which is the histogram of wavelet coefficients, and the other of which is the co-occurrence matrix of the coefficients. These three types of feature vectors were extract for subbands in different spatial frequencies and orientations and used, either separately or in combination, for texture discrimination.

Image signatures based on statistics of wavelet coefficients have mostly been supplanted by bags-of-visual-words-based representations. However, wavelet-based feature vectors are still being proposed,

particularly for texture description. Xu *et al.* [129] proprosed a texture descriptor whose features were computed using multifractal analysis of coefficients in the subbands of a multi-resolution and multi-orientation wavelet decomposition. As mentioned previously, Datta *et al.* [25] used Daubechies wavelet coefficients to construct a feature representation of local texture, which they then used for aesthetic classification.

Our image descriptor differs from previous methods in that, rather than using the raw wavelet coefficients, or simple statistics computed from them, we use the *local center-surround contrast* and *spatial scale* of wavelet coefficients to compute our features. Local contrast, together with spatial scale, are input to the $ECSF$ which outputs the induction weights that server as our feature vectors. In addition, we use a much richer color space to represent our image. The feature extraction process is described in the next section.

8.2 Feature extraction

To extract a feature vector representation using induction weights, we follow a procedure little changed from that described in section 4.3. In $Stage(I)$, we represent the image in our proposed color space, described in section 8.2. Then, the following stages are applied separately to each color channel of an image.

Stage (II): Spatial decomposition Each channel is decomposed in two successive steps. The first one uses the wavelet transform in equation 3.1, obtaining $w_{s,o}$. Subsequently, on each wavelet plane the grouplet transform in equation 4.2 is applied:

$$I_c \stackrel{WT}{\longrightarrow} {}_{s,o} \stackrel{GT}{\longrightarrow} d_{s,j,o} \tag{8.1}$$

where $d_{s,j,o}$ denotes the detail plane at scale j. For a wavelet plane whose largest dimension is size D, $J = log_2 D$. To group features, the association field for a wavelet plane is initialized perpendicularly to its orientation o. Thus for a horizontal wavelet plane, the Haar differencing in equation 4.2 is conducted column-wise and vice versa.

Stage (III): Normalized Center Contrast (NCC) We compute the NCC, $z_{s,j,o}(x\ y)$, for every grouplet coefficient $d_{s,j,o}(x\ y)$ using equation 3.3. The number of pixels spanning the center region and the extended region are 17 and 97 respectively. These are the widths that were obtained for SIM when being fit with the Bruce *et al.* eye-fixation dataset, as described in section 3.2. In using these widths we assume that the viewing distance to the images in our evaluation databases would be similar to that of the Bruce & Tsotsos database. As the dimensions of the images in both databases are on average quite similar, this is a fair assumption.

Stage (IV): Induction weights ($ECSF$) The $ECSF$ function is used to compute induction weights $\alpha_{s,j,o}(x\ y)$ for every grouplet coefficient $d_{s,j,o}(x\ y)$:

$$\alpha_{s,j,o}(x\ y) = ECSF(z_{s,j,o}(x\ y)\ s) \tag{8.2}$$

Stage (V): Binning of induction weights For each grouplet plane, a histogram of the induction weights is constructed.

Stage (VI): Histogram concatenation The histograms for all grouplet planes are concatenated.

Figure 8.2 shows a schema of our feature extraction procedure. These histograms contain a wealth of information about the contrast at each location in an input image, for different scales and orientations, for a given image color channel. The color channels we use are described next.

Color representation

In essence, our features are extracted using outputs from SIM, a model of spatio-chromatic features, which is itself based on a color induction model. This model was defined to predict induction considering as input an opponent representation of color, such as the LGN outputs. These LGN outputs are based on the dominant "parallel streams of cardinal-directions sensitive cells" paradigm of Hurvich & Jameson [51], which is itself supported by psychophysical and physiological measures. However, this view contradicts recent predictions that there are V1 neurons which respond maximally to a broad distribution of color space directions, rather than responding only to the three opponent axes found in pre-cortical vision, typically referred to as the bipolar representation of a color basis. Evidence for these predictions come from the spatial clustering of neurons with similar color preferences found using multivoxel fMRI analysis in V1 [95], intrinsic optical imaging of the macaque brain [128] and from recordings of neurons habituated by prolonged exposure to chromatic modulation [114] among others. In addition, Goddard *et al.* [42] found evidence that information from color-opponent pathways are combined in V1.

In accordance with these results we propose to move from a 3-D bipolar representation of the opponent space towards a 10-D representation derived from the three bipolar opponent axes. In this way we do not change color directions, but rather divide the responses of opposite directions into different channels.

First, the opponent color channels are obtained from image I by converting each (RGB) value, after γ correction, to the opponent space as follows:

$$O1 = \frac{R-G}{R+G+B} \qquad O2 = \frac{R+G-2B}{R+G+B} \qquad O3 = R+G+B \tag{8.3}$$

Each image channel $I_c; c \quad O1 \;\; O2$ is then half-wave rectified twice - once for positive values (I_{c^+}) and once for negative values (I_{c^-}) - resulting in 4 color channels. The intensity channel $O3$ is separated into light and dark channels by its median value. In addition to these 6 channels, we created 4 additional channels by projecting the $(O1 \;\; O2)$ values for each pixel onto the 4 vectors 45° from the cardinal axes, as shown in Figure 8.1(b). As a result, there are 10 channels, 8 of which are chromatic and 2 of which are achromatic, as illustrated in Figure 8.1(c).

8.3 Experiments

8.3.1 Experimental protocol

We extract features for images by decomposing each of the 10 channels into 4 wavelet spatial scales, 4 grouplet spatial scales, and 4 orientations. The α weights are binned into a histogram of length 10. This results in an "$\alpha vector$" of length $10x4x4x4x10 = 6400$.

A whitening transformation was performed on the training vectors in order to decorrelate the features [33]. After whitening PCA was performed for dimensionality reduction, ensuring that 99% of the energy was retained in the projected vectors. The whitening parameters and PCA transformation matrix that were computed for the training vectors were also used for the test vectors.

8.3.2 Quantitative evaluation

We evaluated the performance of our model for the problem of classifying images into two classes: "high-quality" and "low-quality", and compared the performance to state-of-the-art methods.

In our first experiment, we followed the experimental procedure and used the dataset described in [76]. This dataset contains 17,690 images divided into 7 semantic categories, with each category having 2,527 images on average. Each image was labeled as either high or low-quality by at least 8 of

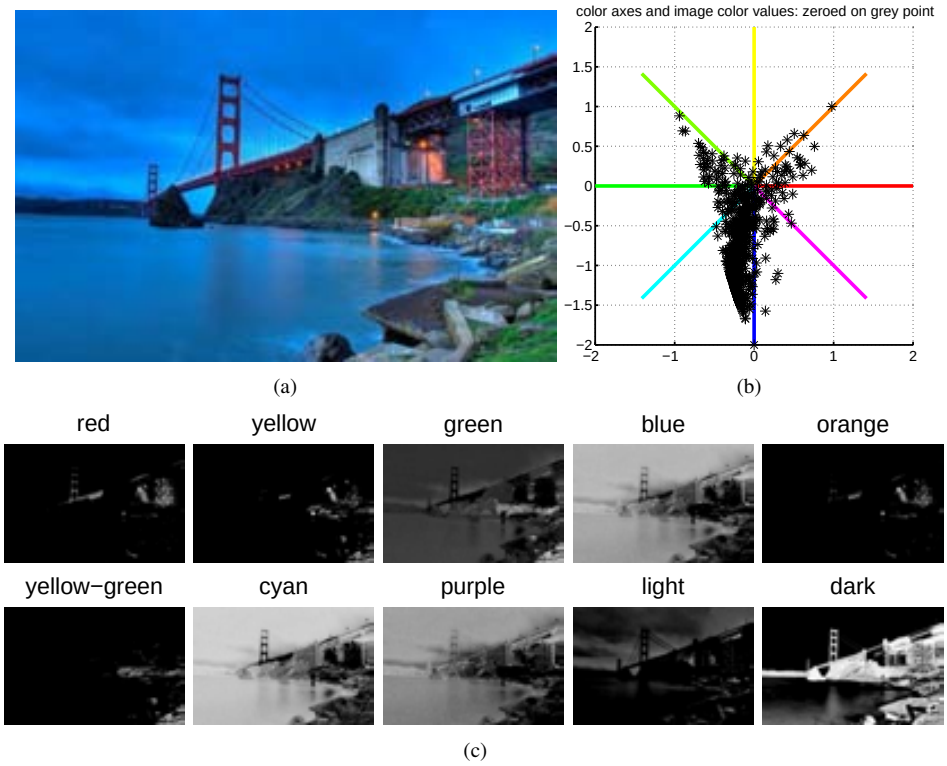


Figure 8.1: Color space representation: (a) Original image. (b) Chromatic 01-02 plane. The image is first represented in color-opponent space. Eight vectors are defined as shown. (c) The 10 resultant channels. Eight channels are chromatic, while two are achromatic.

10 annotators. We trained linear SVMs with our α vectors, using stochastic gradient descent. 50% of the images in a category were randomly selected as training images and the rest were used for testing. We repeated this processing 10 times. The results, which we report in Table 8.1 are the average of the results for these 10 runs. We compare with the results reported by [76] for their proposed features as well as a combination, which we call DKLS, of other state-of-the-art features [25, 62, 77, 78]. As the results show, our features achieve competitive performance.

In our second experiment we created a dataset of 70,000 images from AVA, which we term sAVA, by randomly selecting 30,000 images for training, 10,000 for validation, and 30,000 images for testing. We compare with the aesthetics-specific features of [27] and the generic low-level features of [81]. As Table 8.2 shows, our features achieve competitive performance.

8.4 Discussion

The ability of our α vectors to describe aesthetic characteristics of images may be attributed to several factors. First, the distribution of α weights in each plane informs about the clarity of the image. If there are many salient image regions, i.e. regions with high α values, this may have a negative impact on saliency. Second, the hue composition is captured by the feature vector. This can be seen by comparing

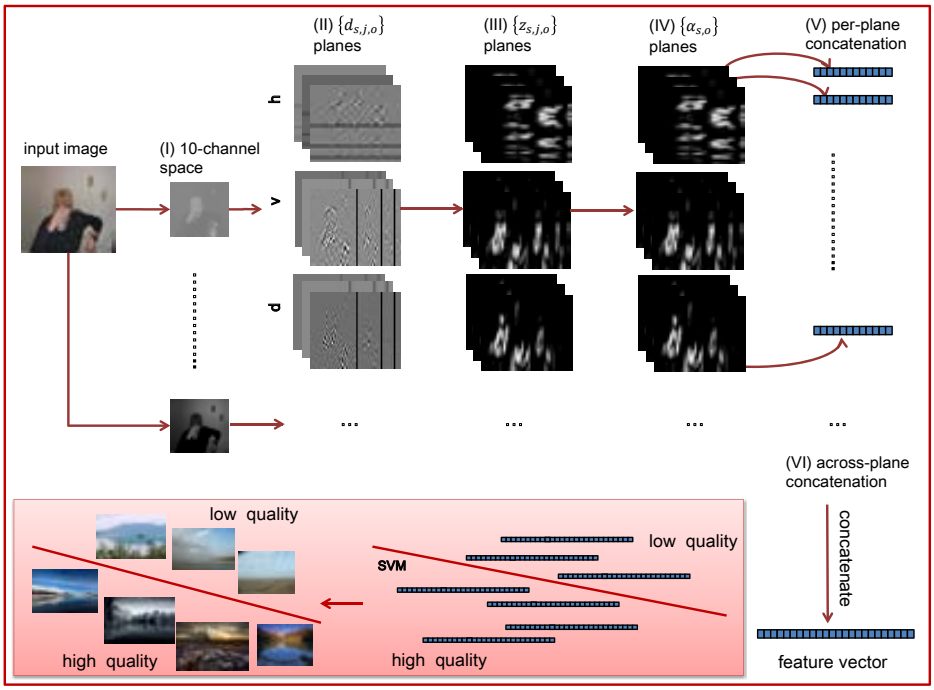


Figure 8.2: Schema of our feature extraction procedure: (I) The image is converted to the 10-D color space. (II) Each channel is decomposed using a wavelet transform. (III) NCC values are calculated. (IV) The $ECSF$ is used to produce the plane of induction weights $\alpha_{s\ o}$. (V) The $\alpha_{s\ o}(x, y)$ values for a given plane are binned into a histogram . (VI) The histograms of each plane are concatenated to produce the feature vector for the image. This feature vector can then be used a train a linear discriminative model of visual aesthetic quality.

Method	Category							all
	animal	architecture	human	landscape	night	plant	static	
DKLS	0.8202	0.8647	0.8915	0.8412	0.7343	0.8762	0.8230	0.8409
Luo *et al.*	0.8712	**0.9004**	**0.9631**	**0.9273**	0.8309	0.9147	0.8890	**0.9044**
α vectors	**0.8851**	0.8615	0.9455	0.9158	**0.8521**	**0.9303**	**0.8917**	0.8665

Table 8.1: Comparison of our proposed feature vectors with the state-of-the-art. The area under the ROC curve is reported for aesthetic models trained only with images in a given category as well as a model trained using all images.

the high-scoring images in Figure 8.3 to the low-scoring ones. Images with many colors are given low scores by our SVM classifier. In addition, color contrast is seen to be a discriminating feature.

In conclusion, our feature vectors, constructed simply from concatenated histograms of a local contrast measure, can achieve state-of-the-art performance. In fact, their performance is only inferior to that of high-dimensional, non-sparse Fisher Vectors, which require significant computational and storage requirements. The proposed vectors, on the other hand, are sparse and quite small, with a

Method	Accuracy
ACQUINE	59.37
BoV+Color+SP	55.23
BoV+SIFT+SP	55.23
BoV+Color+SIFT+SP	60.51
FV+Color+SP	63.05
FV+SIFT+SP	64.00
FV+Color+SIFT+SP	66.05
α vectors	62.57

Table 8.2: Accuracy in predicting binary labels from sAVA dataset.

fixed length of 1600. We believe therefore that these features afford a good balance between high classification performance and efficiency in computation and storage.

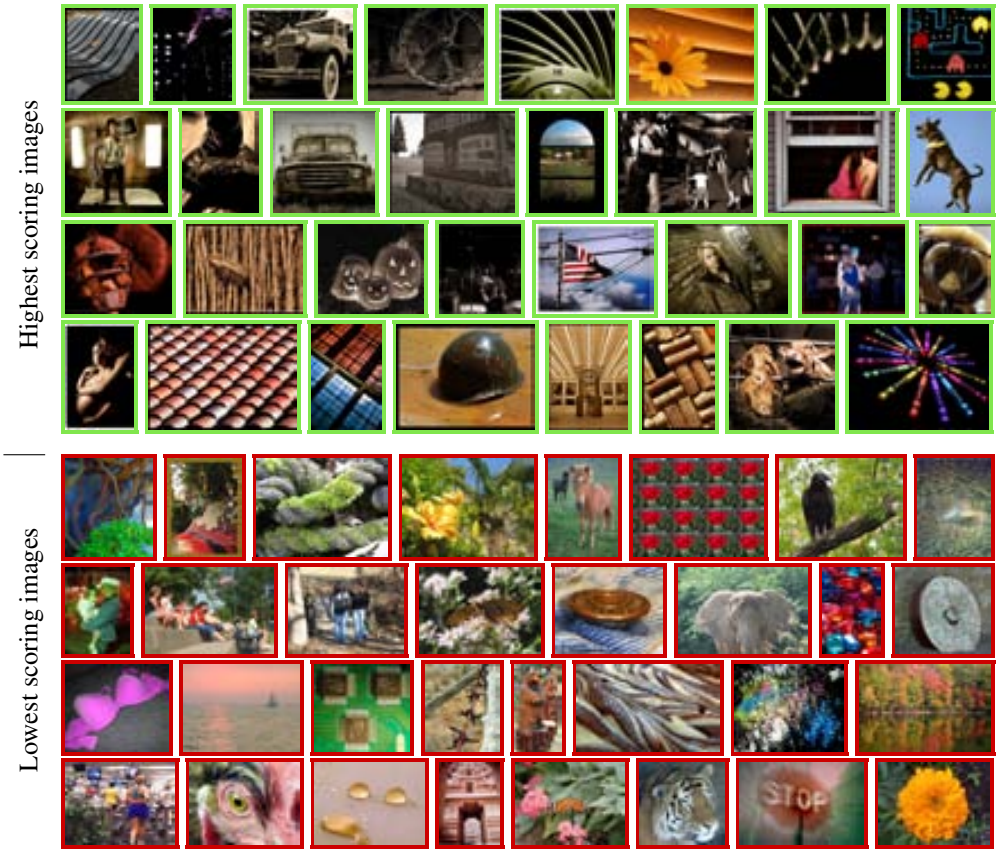


Figure 8.3: Qualitative results on the sAVA dataset: the highest and lowest rank images are shown. The colored frame represents the ground truth (green for "good quality" and red for "bad quality").

Chapter 9

Conclusions and Future Directions

This dissertation was an exploration of the experiences of visual attention and visual aesthetic appreciation. The claim that a bottom-up perspective, afforded by a low-level computational model of color perception, could account in part for behavioral data related to these experiences was advanced and evidence in its favour was presented. In the following sections we summarize the contributions made in support of this claim, and also discuss possible avenues for future research on this topic.

9.1 Summary of Contributions

Fitting the parameters of a low-level vision model using psychophysical data: We fit the parameters of the brightness and color $ECSF$s using data obtained from psychophysical experiments related to brightness and color induction respectively [88]. The visual stimuli used in these experiments were gratings, bars, and concentric circles of alternating colors, and were carefully designed by experimenters.

Estimating saliency using the low-level vision model: We then made several small adaptations to this model. In the first, we changed the spatial extent of the center and surround regions to better conform to known properties of receptive fields in V1. In the second, we performed an inverse wavelet transform on the induction weights themselves in order to produce a saliency map, rather than a perceived image. We then used this map to predict eye-fixations of observers viewing images of natural scenes [88]. Although the visual stimuli used to fit the model parameters are quite different to those typical of natural scenes, the adapted model, which we call SIM (Saliency by Induction Mechanisms), outperforms state-of-the-art saliency models at predicting eye-fixations. Moreover, the psychophysically-tuned parameters are shown to be optimal for both eye-fixation prediction and color perception modeling. This indeed suggests a similar architecture in area V1 for both color perception and saliency. In addition, because the model inherits a principled selection of parameters and an innate spatial pooling mechanism from the color perception model on which it is based, it addresses key criticisms of and unresolved issues with biologically-inspired saliency estimation models. The main criticisms are that (i) such models are difficult to tune owing to their myriad parameters; and (ii) such models do not have a principled manner of pooling information gleaned across different spatial scales.

Improving the image representation of the saliency model: SIM was highly responsive to edges as well as more complex features created by superpositions of edges, such as corners and junctions. However, complex features have been shown to be preferentially fixated upon in comparison to simpler features. Therefore, an image representation for which the response amplitudes of complex features are enhanced relative to simpler features such as edges was desirable. To this end we incorporated an

image decomposition termed the grouplet transform, which was originally used for image de-noising, into our saliency model. To do this, we simply applied a grouplet transform, which was implemented as a Haar transform over a support defined by block matching, to each wavelet plane in the original image decomposition. This operate produces grouplet planes on which the $ECSF$s are applied. The grouplet transform-based image representation essentially extends the extent of the region over which spatial competition occurs for each local feature response. This new representation had the desired effect of enhancing complex features and was able to improve eye-fixation performance [89].

Constructing a large-scale database, AVA, for image aesthetics analysis: After developing the SIM model, we began studying image aesthetics in a computational framework. Computational models of image aesthetics are overwhelmingly trained in a supervised learning framework. Consequently, rich and diverse training images and annotations are critical to the success of such models, moreover because aesthetics itself is a multi-faceted concept without a single interpretation. However, as this is a new area of research, there is a dearth of robust and diverse datasets for training, evaluation and analysis of computational models of aesthetics. To address this issue we made our next contribution: the assembly and in-depth analysis of a large-scale database for image aesthetics analysis, which we call AVA [86, 87]. AVA contains over 200,000 images, with hundreds of score annotations each. These score annotations form score distributions over a rating scale. We have shown that these distributions are largely Gaussian. Their means and variances allow one to gain an idea of the general consensus on the aesthetic quality of an image while the variance informs about the degree of agreement between observers of the image. Many of the images in AVA also have semantic tags given by users, which can aid in understanding the relationship between semantic content and aesthetic judgments. In addition, the images have many associated textual comments given by annotators, providing detailed feedback on an image's aesthetic characteristics and attributes.

Demonstrating the advantages of the large-scale and versatile data in the AVA database: In [85–87] we demonstrated, through several applications, how the large scale and diverse annotations of AVA can be leveraged to improve performance on existing preference tasks and inspire new ones. In particular, we built models to perform binary classification into "high-quality" and "low-quality" aesthetic categories, aesthetic score prediction, and image ranking. We showed that the large scale of training data in AVA enabled significant improvement in model training. We also showed that by judiciously selecting training images from among those in AVA, we could retain model performance even when fewer training images are used. In the case of image re-ranking, we used the semantic labels given to images in AVA to train semantic classifiers. We then used the aesthetic labels in AVA to train both content-dependent and content-independent aesthetic models. We combined the output of semantic and aesthetic models in several ways, which allowed us to rank images according to both their semantic and aesthetic characteristics.

Estimating aesthetic quality using the low-level vision model and large-scale data: At this stage, armed with a suitable dataset and baseline methods, we returned to the central theme of the dissertation: the plausibility of using a common low-level vision model to predict different complex visual experiences. We again made slight adaptations to the color perception model and were able to extract image features which can predict aesthetics labels given to images by human annotators. In this instance, we formed histograms of the alpha weights computed by the $ECSF$s for each plane. We then concatenated these histograms to form the feature vector. In addition, we introduced a new 10-channel color space representation which provides more fine-grained information about the colors present and absent in the image. Our final feature vector was a concatenation of the feature vectors from each color channel. These feature vectors were used to train SVM models for binary aesthetic classification. The features were shown to perform at a state-of-the-art level when compared with features extracted using procedures that have been hand-crafted especially for aesthetics and also when compared with sophisticated generic low-level visual features. We believe that this is because low-level visual features in

our saliency model capture local image characteristics such as feature contrast, grouping and isolation, characteristics thought to be related to universal aesthetic laws.

Thus, our saliency model and aesthetics features, both of which have been directly derived from a model of low-level color perception, achieve state-of-the-art performance on related predictive tasks. Their success adds evidence to the hypothesis that color perception, bottom-up visual attention and visual aesthetics appreciation are driven in significant part by cell responses from a common neural substrate in the early human visual system.

9.2 Future Directions

There are several future directions, described below in which to further develop the work presented in this dissertation.

The Low-level Vision Model

A fine-grained color-space representation was shown to be beneficial for modeling aesthetic quality. Further research is needed to determine whether this sort of hue-map inspired representation [95, 114, 128] is also beneficial for color perception and saliency models. The best color spaces axes used to create these maps must also be determined. Color names may also be explored for determining these axes.

Another area of improvement for the low-level vision model is to more precisely model the center-surround regions. In the current model, the spatial scale at peak contrast sensitivity is processed by a receptive field with a center of 1° of visual angle and an extra-receptive field about 5 times that. These sizes correspond to current estimates found in the literature [14, 18, 112, 120]. In the current model, spatial scales below peak sensitivity have larger center-surround regions while spatial scales above peak sensitivity have smaller center-surround regions. Further research should be done to better determine how these region sizes should change in relation to spatial scale.

Interplay Between Aesthetic Appreciation and Visual Attention

The relationship between aesthetic appreciation and visual attention has been explored in several recent studies [24,82,121], by using fMRI and eye-tracking data. However, because low-level mechanisms are better understood and easier to interpret than such behavior data, it may be beneficial to also investigate the interplay between these two experiences from a bottom-up perspective. This perspective may be afforded by examining modulated low-level contrast ($ECSF$ weights), which was shown in chapters 3 and 8 to account in part for both eye-fixations and aesthetic judgments. In particular, it remains unclear whether visual attention and visual aesthetic appreciation each have a direct relationship with local contrast, or whether this relationship is caused indirectly by a dependence of one on the other.

Extensions of the Low-level Vision Model for Other Vision Problems

We would like to explore the application of grouplet-based representations to other computer vision problems, such as feature detection, which typically involve scale-space decompositions. In addition, because our "aesthetic" feature vector is in fact generic, we would like to investigate its performance when used as an image signature for scene categorization.

Adding top-down cues

Our bottom-up models provide a unified view of different visual experiences, by incorporating low-level mechanisms common to them. However, this view is incomplete and as a result cannot hope to replicate behavior related to visual attention or aesthetic appreciation, which is highly susceptible to top-down task-driven cues as well as to semantic cues. For this reason, we would like to expand the model to include such cues, particularly those which are also a factor in different visual experiences, such as the presence of faces in visual stimuli.

List of Publications

This dissertation has led to the following communications:

Journal Papers

N. Murray, M. Vanrell, X. Otazu, and C.A. Párraga. Low-level spatio-chromatic grouping for saliency estimation. *Submitted to IEEE Transactions on Pattern Analysis and Machine Intelligence*, August 2012.

N. Murray, L. Marchesotti, and F. Perronnin. Robust features and data for image aesthetics analysis. *Submitted to International Journal of Computer Vision*, November 2012.

Conference Contributions

N. Murray, M. Vanrell, X. Otazu, and C.A. Párraga. Saliency estimation using a non-parametric low-level vision model. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 433–440. IEEE, 2011.

M. Vanrell, N. Murray, R. Benavente, C.A. Párraga, X. Otazu, and R. Baldrich. Perception based representations for computational colour. In *Proceedings of the Third international conference on Computational color imaging*, CCIW'11, pages 16–30, Berlin, Heidelberg, 2011. Springer-Verlag.

N. Murray, L. Marchesotti, and F. Perronnin. AVA: A large-scale database for aesthetic visual analysis. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 2408–2415. IEEE, 2012.

N. Murray, L. Marchesotti, and F. Perronnin. Learning to rank images using semantic and aesthetic labels. In *Brit. Mach. Vision Conf*, 2012.

Published abstracts

M. Vanrell, N. Murray, X. Otazu, and C.A. Párraga Computation of saliency maps using psychophysical measurements of colour induction. *AVA/BMVA (to appear in forthcoming issue of Perception)*, 2012.

Patents

N. Murray and L. Marchesotti. Image Selection Based on Photographic Style. *US patent application led, patent pending*.

Final Acknowledgements

We thank Hae Jong Seo for sharing his evaluation code. This work has been supported by Projects TIN2007-64577, TIN2010-21771-C02-1 and Consolider-Ingenio 2010-CSD2007-00018 from the Spanish Ministry of Science. C. Alejandro Parraga was funded by grant RYC-2007-00484. This work has also been supported within a sponsored research agreement between the Universitat Autonoma de Barcelona and Xerox Research Centre Europe on the topic of "Applied Visual Aesthetics".

Bibliography

[1] Entry: "aesthetics". *The American Heritage® Dictionary of the English Language, Fourth Edition*. http://dictionary.reference.com/browse/aesthetics, Oct 2012.

[2] A.N. Akansu and P.R. Haddad. *Multiresolution signal decomposition: transforms, subbands, and wavelets*. Academic Press, 2000.

[3] J. Atick. Could information theory provide an ecological theory of sensory processing? *Network*, 3:213–251, 1991.

[4] F. Attneave. Some informational aspects of visual perception. *Psychol Rev*, 61(3):183–93, 1954.

[5] Erhardt Barth, Christoph Zetzsche, and Ingo Rentschler. Intrinsic two-dimensional features as textons. *J. Opt. Soc. Am. A*, 15(7):1723–1732, Jul 1998.

[6] C. Blakemore and F. W. Campbell. On the existence of neurons in the human visual system selectively sensitive to the orientation and size of retinal images. *The Journal of Physiology*, 203(1):237–260, 1969.

[7] B. Blakeslee and M. E. McCourt. Similar mechanisms underlie simultaneous brightness contrast and grating induction. *Vision Research*, 37(20):2849–2869, 1997.

[8] Ali Borji and Laurent Itti. State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 99(PrePrints), 2012.

[9] Ali Borji, Dicky N. Sihite, , and Laurent Itti. Salient object detection: A benchmark. In *In ECCV*, 2012.

[10] L. Bottou and O. Bousquet. The tradeoffs of large scale learning. In *Advances in neural information processing systems*, 2007.

[11] S. Brown, X. Gao, L. Tisdelle, S.B. Eickhoff, and M. Liotti. Naturalizing aesthetics: brain areas for aesthetic appraisal across sensory modalities. *Neuroimage*, 58(1):250–258, 2011.

[12] N. D. Bruce and J. K. 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, MIT Press, 2006. MIT Press.

[13] F.W. Campbell and JG Robson. Application of fourier analysis to the visibility of gratings. *The Journal of Physiology*, 197(3):551, 1968.

[14] James R. Cavanaugh, Wyeth Bair, J. Anthony Movshon, James R, Wyeth Bair, and J. Anthony Movshon. Nature and interaction of signals from the receptive field center and surround in macaque v1 neurons. *J Neurophysiol*, pages 2530–2546, 2002.

[15] M. Cerf, J. Harel, W. Einhäuser, and C. Koch. Predicting human gaze using low-level saliency combined with face detection. *Advances in neural information processing systems*, 20, 2008.

[16] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[17] T. Chang and C.C.J. Kuo. Texture analysis and classification with tree-structured wavelet transform. *Image Processing, IEEE Transactions on*, 2(4):429–441, 1993.

[18] Li Chao-Yi and Li Wu. Extensive integration field beyond the classical receptive field of cat's striate cortical neurons–classification and tuning properties. *Vision Research*, 34(18):2337 – 2355, 1994.

[19] A. Chatterjee. Neuroaesthetics: a coming of age story. *Journal of Cognitive Neuroscience*, 23(1):53–62, 2011.

[20] X. Chen, G.J. Zelinsky, et al. Real-world visual search is dominated by top-down guidance. *Vision research*, 46(24):4118–4133, 2006.

[21] Duncan Cramer and Dennis Howitt. *The SAGE dictionary of statistics*. SAGE, 1st edition, 2004. p. 21 (entry "ceiling effect"), p. 67 (entry "floor effect").

[22] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *ECCV SLCV Workshop*, 2004.

[23] G.C. Cupchik. From perception to production: A multilevel analysis of the aesthetic process. *Emerging visions of the aesthetic process: Psychology, semiology, and philosophy*, pages 61–81, 1992.

[24] G.C. Cupchik, O. Vartanian, A. Crawley, and D.J. Mikulis. Viewing artworks: Contributions of cognitive control and perceptual facilitation to aesthetic experience. *Brain and cognition*, 70(1):84–91, 2009.

[25] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z. Wang. Studying aesthetics in photographic images using a computational approach. In *ECCV*, pages 7–13, 2006.

[26] Ritendra Datta, Jia Li, and James Z. Wang. Learning the consensus on visual quality for next-generation image management. In *ACM-MM*, 2007.

[27] Ritendra Datta and James Ze Wang. Acquine: aesthetic quality inference engine - real-time automatic rating of photo aesthetics. In *MIR*, 2010.

[28] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2009.

[29] A.M. Derrington, J. Krauskopf, and P. Lennie. Chromatic mechanisms in lateral geniculate nucleus of macaque. *The Journal of Physiology*, 357(1):241–265, 1984.

[30] R. Desimone. Neural mechanisms for visual memory and their role in attention. *Proceedings of the National Academy of Sciences*, 93(24):13494–13499, 1996.

[31] S. Dhar, V. Ordonez, and T.L. Berg. High level describable attributes for predicting aesthetics and interestingness. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 1657–1664. IEEE, 2011.

[32] A. Duchowski. *Eye tracking methodology: Theory and practice*, volume 373. Springer, 2007.

[33] R.O. Duda, P.E. Hart, and D.G. Stork. *Pattern classi cation*. Pattern Classification and Scene Analysis: Pattern Classification. Wiley, 2001.

[34] H.E. Egeth and S. Yantis. Visual attention: Control, representation, and time course. *Annual review of psychology*, 48(1):269–297, 1997.

[35] M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2008 (VOC2008) Results, 2008.

[36] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 2010.

[37] T. Foulsham and G. Underwood. What can saliency models predict about eye movements? spatial and sequential aspects of fixations during encoding and recognition. *Journal of Vision*, 8(2), 2008.

[38] D. Gao, V. Mahadevan, and N. Vasconcelos. On the plausibility of the discriminant center-surround hypothesis for visual saliency. *Journal of Vision*, 8(7:13):1–18, 2008.

[39] D. Gao and N. Vasconcelos. Bottom-up saliency is a discriminant process. In *Proc. IEEE Int l Conf. Computer Vision*, pages 1–6, 2007.

[40] B. Geng, L. Yang, C. Xu, X.S. Hua, and S. Li. The role of attractiveness in web image search. In *Proceedings of the 19th ACM international conference on Multimedia*. ACM, 2011.

[41] Hannah Ginsborg. Kantś aesthetics and teleology. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Fall 2008 edition, 2008. `http://plato.stanford.edu/archives/fall2008/entries/kant-aesthetics/`.

[42] E. Goddard, D.J. Mannion, J.S. McDonald, S.G. Solomon, and C.W.G. Clifford. Combination of subcortical color channels in human visual cortex. *Journal of vision*, 10(5), 2010.

[43] Ted Gracyk. Humeś aesthetics. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Winter 2011 edition, 2011. `http://plato.stanford.edu/archives/win2011/entries/hume-aesthetics/`.

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

[45] K. Hammermeister. *The German aesthetic tradition*. Cambridge University Press, 2002.

[46] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. *Advances in neural information processing systems*, 19:545, 2007.

[47] Thomas Deselaers Henning Muller, Paul Clough and Barbara Caput. Experimental evaluation in visual information retrieval. the information retrieval series. *Springer*, 2010.

[48] X. Hou, J. Harel, and C. Koch. Image signature: Highlighting sparse salient regions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(1):194, 2012.

[49] X. Hou and L. Zhang. Dynamic visual attention: Searching for coding length increments. *Advances in neural information processing systems*, 21:681–688, 2008.

[50] R.S. Hunter. Photoelectric color difference meter. *Josa*, 48(12):985–993, 1958.

[51] L.M. Hurvich and D. Jameson. An opponent-process theory of color vision. *Psychological Review; Psychological Review*, 64(6p1):384, 1957.

[52] L. Itti. Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes. *Visual Cognition*, 12(6):1093–1123, 2005.

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

[54] 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.

[55] 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.

[56] T. Jaakkola and D. Haussler. Exploiting generative models in discriminative classifiers. 1999.

[57] E. Jacobson and W. Ostwald. *Color harmony manual*. Container Corporation of America, 1948.

[58] T. Joachims. Optimizing search engines using clickthrough data. In *SIGKDD*, 2002.

[59] D. Joshi, R. Datta, E. Fedorovskaya, Q.T. Luong, J.Z. Wang, J. Li, and J. Luo. Aesthetics and emotions in images. *Signal Processing Magazine, IEEE*, 28(5):94–115, 2011.

[60] T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. In *Proc. IEEE Int l Conf. Computer Vision*, 2009.

[61] H. Kawabata and S. Zeki. Neural correlates of beauty. *Journal of Neurophysiology*, 91(4):1699–1705, 2004.

[62] Yan Ke, Xiaoou Tang, and Feng Jing. The design of high-level features for photo quality assessment. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2006.

[63] W. Kienzle, F.A. Wichmann, B. Schalkopf, and M. O. Franz. A nonparametric approach to bottom-up visual saliency. In *Advances in neural information processing systems 19*. MIT Press, 2007.

[64] R. Kliegl and R.K. Olson. Reduction and calibration of eye monitor data. *Behavior Research Methods*, 13(2):107–111, 1981.

[65] C. Koch and S. Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. *Human neurobiology*, 4(4):219–227, 1985.

[66] Kodak. *How to take good pictures : a photo guide*. Random House Inc, 1982.

[67] B.P. Krages. *Photography: the art of composition*. Allworth Press, 2005.

[68] A. Kundu and J.L. Chen. Texture classification using qmf bank-based subband decomposition. *CVGIP: Graphical models and image processing*, 54(5):369–384, 1992.

[69] R. Fergus L. Fei-Fei and P. Perona. Learning generative visual models from few training examples: an incremental bayesian approach tested on 101 object categories. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2004.

[70] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2006.

[71] H. Leder, B. Belke, A. Oeberst, and D. Augustin. A model of aesthetic appreciation and aesthetic judgments. *British Journal of Psychology*, 95(4):489–508, 2004.

[72] T.S. Lee. Image representation using 2d gabor wavelets. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 18(10):959–971, 1996.

[73] Z. Li et al. A saliency map in primary visual cortex. *Trends in cognitive sciences*, 6(1):9–16, 2002.

[74] K.C. Liang and C.C.J. Kuo. Waveguide: a joint wavelet-based image representation and description system. *Image Processing, IEEE Transactions on*, 8(11):1619–1629, 1999.

[75] D.G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2), 2004.

[76] Wei Luo, Xiaogang Wang, and Xiaoou Tang. Content-based photo quality assessment. In *Proc. IEEE Int l Conf. Computer Vision*, 2011.

[77] Yiwen Luo and Xiaoou Tang. Photo and video quality evaluation: Focusing on the subject. In *ECCV*, 2008.

[78] Rahul Sukthankar. M. S. Subhabrata Bhattacharya. framework for photo-quality assessment and enhancement based on visual aesthetics. *ACM MM*, 1, Oct. 2011.

[79] D.I.A. MacLeod and R.M. Boynton. Chromaticity diagram showing cone excitation by stimuli of equal luminance. *JOSA*, 69(8):1183–1186, 1979.

[80] Stéphane Mallat. Geometrical grouplets. *Applied and Computational Harmonic Analysis*, 26(2):161 – 180, 2009.

[81] Luca Marchesotti, Florent Perronnin, Diane Larlus, and Gabriela Csurka. Assessing the aesthetic quality of photographs using generic image descriptors. In *Proc. IEEE Int l Conf. Computer Vision*, 2011.

[82] D. Massaro, F. Savazzi, C. Di Dio, D. Freedberg, V. Gallese, G. Gilli, and A. Marchetti. When art moves the eyes: A behavioral and eye-tracking study. *PloS one*, 7(5):e37285, 2012.

[83] O. Le Meur, P. Le Callet, D. Barba, and D. Thoreau. A coherent computational approach to model bottom-up visual attention. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(5):802–817, 2006.

[84] K.T. Mullen. The contrast sensitivity of human color-vision to red green and blue yellow chromatic gratings. *Journal of Physiology*, pages 381–400, 1985.

[85] N. Murray, L. Marchesotti, and F. Perronnin. Learning to rank images using semantic and aesthetic labels. In *Brit. Mach. Vision Conf*, 2012.

[86] N. Murray, L. Marchesotti, and F. Perronnin. AVA: A large-scale database for aesthetic visual analysis. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 2408–2415. IEEE, 2012.

[87] N. Murray, L. Marchesotti, and F. Perronnin. Robust features and data for image aesthetics analysis. *Submitted to International Journal of Computer Vision*, November 2012.

[88] N. Murray, M. Vanrell, X. Otazu, and C.A. Parraga. Saliency estimation using a non-parametric low-level vision model. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 433–440. IEEE, 2011.

[89] N. Murray, M. Vanrell, X. Otazu, and C.A. Parraga. Low-level spatio-chromatic grouping for saliency estimation. *Submitted to IEEE Transactions on Pattern Analysis and Machine Intelligence*, August 2012.

[90] U. Neisser. *Cognition and reality: Principles and implications of cognitive psychology.* WH Freeman/Times Books/Henry Holt & Co, 1976.

[91] E. Niebur and C. Koch. Control of selective visual attention: Modeling the" where" pathway. *Advances in neural information processing systems*, pages 802–808, 1996.

[92] C.F. Nodine, P.J. Locher, and E.A. Krupinski. The role of formal art training on perception and aesthetic judgment of art compositions. *Leonardo*, pages 219–227, 1993.

[93] X. Otazu, M. Vanrell, and C. A. Párraga. Multiresolution wavelet framework models brightness induction effects. *Vision Research*, 48(5):733–751, 2008.

[94] Xavier Otazu, C. Alejandro Parraga, and Maria Vanrell. Toward a unified chromatic induction model. *Journal of Vision*, 10(12), 2010.

[95] L.M. Parkes, J.B.C. Marsman, D.C. Oxley, J.Y. Goulermas, and S.M. Wuerger. Multivoxel fmri analysis of color tuning in human primary visual cortex. *Journal of Vision*, 9(1), 2009.

[96] D. Parkhurst, K. Law, E. Niebur, et al. Modeling the role of salience in the allocation of overt visual attention. *Vision research*, 42(1):107–124, 2002.

[97] F. Perronnin and C. Dance. Fisher kernels on visual vocabularies for image categorization. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2007.

[98] F. Perronnin, J. Sánchez, and Thomas Mensink. Improving the fisher kernel for large-scale image classification. In *ECCV*, 2010.

[99] R. J. Peters, A. Iyer, L. Itti, and C. Koch. Components of bottom-up gaze allocation in natural images. *Vision Research*, 45(8):2397–2416, Aug 2005.

[100] N. Pinto, D. Doukhan, J.J. DiCarlo, and D.D. Cox. A high-throughput screening approach to discovering good forms of biologically inspired visual representation. *PLoS computational biology*, 5(11):e1000579, 2009.

[101] V.S. Ramachandran and W. Hirstein. The science of art: A neurological theory of aesthetic experience. *Journal of Consciousness Studies, 6*, 6(7):15–51, 1999.

[102] R. Reber, N. Schwarz, and P. Winkielman. Processing fluency and aesthetic pleasure: Is beauty in the perceiver's processing experience? *Personality and Social Psychology Review*, 8(4):364–382, 2004.

[103] R. Reber, P. Winkielman, and N. Schwarz. Effects of perceptual fluency on affective judgments. *Psychological science*, 9(1):45–48, 1998.

[104] J.A. Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.

[105] J. San Pedro, T. Yeh, and N. Oliver. Leveraging user comments for aesthetic aware image search reranking. In *WWW*, 2012.

[106] H. J. Seo and P. Milanfar. Static and space-time visual saliency detection by self-resemblance. *Journal of Vision*, 9(12):15.1–27, 2009.

[107] Robert Shapley and Michael J. Hawken. Color in the cortex: single- and double-opponent cells. *Vision Research*, 51(7):701 – 717, 2011. Vision Research 50th Anniversary Special Issue.

[108] James Shelley. 18th century british aesthetics. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Summer 2012 edition, 2012. `http://plato.stanford.edu/archives/sum2012/entries/aesthetics-18th-british/`.

[109] James Shelley. The concept of the aesthetic. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Spring 2012 edition, 2012. `http://plato.stanford.edu/archives/spr2012/entries/aesthetic-concept/`.

[110] Eero P. Simoncelli and Odelia Schwartz. Modeling surround suppression in v1 neurons with a statistically-derived normalization model. In *Advances in neural information processing systems 2*, pages 153–159. MIT Press, 1999.

[111] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *Proc. IEEE Int l Conf. Computer Vision*, 2003.

[112] A.T. Smith, K.D. Singh, A.L. Williams, and M.W. Greenlee. Estimating Receptive Field Size from fMRI Data in Human Striate and Extrastriate Visual Cortex. *Cerebral Cortex*, 11(12):1182–1190, 2001.

[113] S. Suzuki and P. Cavanagh. Facial organization blocks access to low-level features: An object inferiority effect. *Journal of Experimental Psychology: Human Perception and Performance*, 21(4):901, 1995.

[114] C. Tailby, S.G. Solomon, N.T. Dhruv, and P. Lennie. Habituation reveals fundamental chromatic mechanisms in striate cortex of macaque. *The Journal of Neuroscience*, 28(5):1131–1139, 2008.

[115] B.M. ter Haar Romeny. *Front-end vision and multi-scale image analysis*. Kluwer Academic Publishers Dordrecht:, 2003.

[116] A. Treisman. Features and objects in visual processing. *Scienti c American*, 255(5):114–125, 1986.

[117] A.M. Treisman and G. Gelade. A feature-integration theory of attention. *Cognitive psychology*, 12(1):97–136, 1980.

[118] S.K.L.G. Ungerleider. Mechanisms of visual attention in the human cortex. *Annual review of neuroscience*, 23(1):315–341, 2000.

[119] G. Van de Wouwer, P. Scheunders, and D. Van Dyck. Statistical texture characterization from discrete wavelet representations. *Image Processing, IEEE Transactions on*, 8(4):592–598, 1999.

[120] G.A. Walker, I. Ohzawa, R.D. Freeman, et al. Suppression outside the classical cortical receptive field. *Visual neuroscience*, 17(3):369–379, 2000.

[121] C. Wallraven, D. Cunningham, J. Rigau, M. Feixas, and M. Sbert. Aesthetic appraisal of art-from eye movements to computers. *Computational aesthetics*, pages 137–144, 2009.

[122] Annette Werner. The spatial tuning of chromatic adaptation. *Vision Research*, 43(15):1611 – 1623, 2003.

[123] H.R. Wilson and J.R. Bergen. A four mechanism model for threshold spatial vision. *Vision research*, 19(1):19–32, 1979.

[124] A.S. Winston and G.C. Cupchik. The evaluation of high art and popular art by naive and experienced viewers. *Visual Arts Research*, pages 1–14, 1992.

[125] Ou Wu, Weiming Hu, and Jun Gao. Learning to predict the perceived visual quality of photos. In *Proc. IEEE Int l Conf. Computer Vision*, 2011.

[126] www.dpchallenge.com. How do comments work?, September 2012. `http://www.dpchallenge.com/help\_faq.php\#howcomments`.

[127] www.dpchallenge.com. What is the critique club?, September 2012. `http://www.dpchallenge.com/forum.php?action=read\&FORUM\_THREAD\_ID=19842`.

[128] Y. Xiao, A. Casti, J. Xiao, and E. Kaplan. Hue maps in primate striate cortex. *Neuroimage*, 35(2):771–786, 2007.

[129] Y. Xu, X. Yang, H. Ling, and H. Ji. A new texture descriptor using multifractal analysis in multi-orientation wavelet pyramid. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 161–168. IEEE, 2010.

[130] L. Yao, P. Suryanarayan, M. Qiao, J.Z. Wang, and J. Li. Oscar: On-site composition and aesthetics feedback through exemplars for photographers. *International Journal of Computer Vision*, 2012.

[131] C. Yu, S.A. Klein, and D.M. Levi. Facilitation of contrast detection by cross-oriented surround stimuli and its psychophysical mechanisms. *Journal of Vision*, 2(3):243–255, 2002.

[132] S. Zeki. Art and the brain. *Journal of Consciousness Studies*, 6(6-7):6–7, 1999.

[133] S. Zeki, JD Watson, CJ Lueck, K.J. Friston, C. Kennard, and RS Frackowiak. A direct demonstration of functional specialization in human visual cortex. *The Journal of Neuroscience*, 11(3):641–649, 1991.

[134] C. Zetzsche, K. Schill, H. Deubel, G. Krieger, E. Umkehrer, and S. Beinlich. Investigation of a sensorimotor system for saccadic scene analysis: an integrated approach. In *Proceedings of the fth international conference on simulation of adaptive behavior on From animals to animats 5*, pages 120–126, Cambridge, MA, USA, 1998. MIT Press.

[135] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell. SUN: A Bayesian framework for saliency using natural statistics. *Journal of Vision*, 8(7):–, 2008.

[136] Q. Zhao and C. Koch. Learning visual saliency by combining feature maps in a nonlinear manner using adaboost. *Journal of Vision*, 12(6), 2012.