# High Level Describable Attributes for Predicting Aesthetics and Interestingness

A Thesis Presented

by

**Sagnik Dhar**

to

The Graduate School

in Partial Fulfillment of the

Requirements

for the Degree of

**Master of Science**

in

**Computer Science**

Stony Brook University

December 2010

**Stony Brook University**

The Graduate School

**Sagnik Dhar**

We, the thesis committee for the above candidate for the

Master of Science degree,

hereby recommend acceptance of this thesis.

**Professor Tamara Berg - Advisor**
**Assistant Professor, Department of Computer Science**

**Professor Dimitris Samaras - Chairperson of Defense**
**Associate Professor, Department of Computer Science**

**Professor Alexander Berg**
**Assistant Professor, Department of Computer Science**

This thesis is accepted by the Graduate School

Lawrence Martin
Dean of the Graduate School

Abstract of the Thesis

**High Level Describable Attributes for Predicting Aesthetics and Interestingness**

by

**Sagnik Dhar**

**Master of Science**

in

**Computer Science**

**Stony Brook University**

2010

With the rise in popularity of digital cameras, the amount of visual data available on the web is growing exponentially. Some of these pictures are extremely beautiful and aesthetically pleasing. Unfortunately the vast majority are uninteresting or of low quality. This paper demonstrates a simple, yet powerful method to automatically select high aesthetic quality images from large image collections with performance significantly better than the state of the art. We also show significantly better results on predicting the interestingness of Flickr images, and on a novel problem of predicting query specific interestingness. Our aesthetic quality estimation method explicitly predicts some of the possible image cues that a human might use to evaluate an image and then uses them in a discriminative approach. These cues or high level describable image attributes fall into three broad types: 1) compositional attributes related to image layout or configuration, 2) content attributes related to the objects or scene types depicted, and 3) sky-illumination attributes related to the natural lighting conditions. We demonstrate that an aesthetics classifier trained on these describable attributes can provide a significant improvement over state of the art methods for predicting human quality judgments.

*Dedicated to my mother, whose forays into the world of Architecture instilled in me my earliest notions of aesthetics.*

# Contents

# Acknowledgements

I would like to thank my advisor Prof. Tamara Berg for her continuing guidance and advice. Right from the formative stages of this thesis to the final draft, she has helped me in various aspects of my work. Her advice has been invaluable as I was also being introduced to the various nuances of Computer Vision while working on my thesis.

I would also like to thank Vicente Ordoneź who worked very closely with me and assisted me in certain portions of the thesis.

I would also like to thank all my friends for their constant support. I would especially like to thank Rupsa Datta and Debaleena Chattopadhyay for frequent technical discussions during the course of my thesis. I am grateful to Ambuj Thacker and Sumati Priya, who helped me annotate vast portions of the dataset that I used in my thesis. Finally, I would also like to thank my friends and family for their moral support.

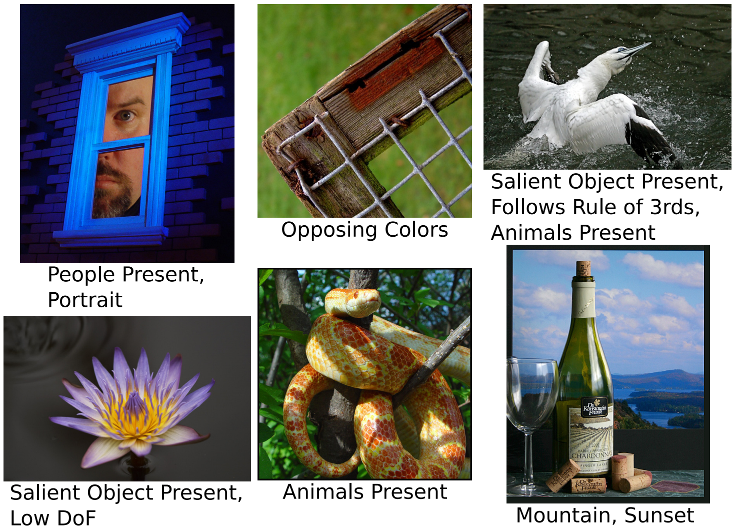
The text of this thesis in part has been sent as a submission to the conference, IEEE Computer Vision and Pattern Recognition (CVPR) 2011. Vicente Ordoneź and Tamara Berg are listed as co-authors in the submission entry. They have assisted, as mentioned above, in the research that forms the basis for this thesis.

# Chapter 1

# Introduction

Automating general image understanding has proven very difficult and is a far from solved problem. Research has progressed in the area of solving the problem of describing what objects are present in an image (including their spatial arrangements and interactions), what general scene type is shown (e.g. a beach, office, street etc.), or the visual qualities of an image (such as whether a picture was captured indoors, or outside on a sunny day). None of these sub-problems solve the grand problem of complete image understanding, but are steps towards solving it. Being able to identify the basic scene type is an analysis of the contents of the image. Obtaining information about the visual qualities of an image aims at being able to identify cues (possibly weak) about the illumination context of the image. In my thesis, I have attempted to use progress made in different aspects of image understanding to extract the information conveyed by it as exhaustively as possible.

Very few web scale systems exist which attempt to do image understanding. This is because of the high computational requirements to do this in an unsupervised environment. As a result, the web primarily relies on using human annotated tags or ratings to associate information with an image. This dependence on the 'wisdom of the crowd' might not be very effective for a system which infers the aesthetic quality of an image. To correctly ascertain the aesthetic quality of an image, a human being would have to judge a photograph with multiple perspectives. He would have to verify whether the golden rules of photography have been followed or not. He would have to also carefully observe the content of the photograph. Do the subjects and their interactions in the photograph arouse a positive or negative opinion in the mind ? These are two completely different thought processes and it is not easy for a person to deploy both simultaneously to judge an image. As a result, an aesthetics rating system controlled by humans would have a high degree of ambiguity.


People Present, Portrait

Opposing Colors

Salient Object Present, Follows Rule of 3rds, Animals Present

Salient Object Present, Low DoF

Animals Present

Mountain, Sunset

Figure 1.1: High Level describable attributes automatically predicted by our system.

In this thesis, we build on the progress made by previous research to develop techniques for automatically estimating high level describable visual attributes of images and then demonstrate the utility of these estimates for predicting perceived aesthetic quality of images. In particular we demonstrate that useful information can be extracted about the following attributes of images:

1. Compositional Attributes - characteristics related to the layout or configuration

of an image that indicate how closely the image follows photographic rules of composition.

2. Content Attributes - characteristics related to the presence of specific objects or categories of objects including faces, animals, and scene types.

3. Sky-Illumination Attributes - characteristics of the natural illumination present in a photograph.

We use the phrase *high level describable attributes* to indicate that these are the kinds of characteristics that a human might use to describe an image. Describability is key here so that we can ask people to label images according to the presence or absence of an attribute. In our data-centric approach, more the number of images the algorithm could learn from, the better trained it is. One sees the possibility of using systems similar to Amazon's Mechanical Turk to label hundreds of images. We could then use this labeled data to train classifiers for recognizing images displaying the attribute. Again, because these are describable attributes, they may be useful as additional constraints to be specified in a search query, or as features for effectively organizing image collections.

In addition, we propose that these describable visual attributes could be used as informative image features when training classifiers for a variety of tasks. Kumar et al [14; 15], have shown that for face verification, describable facial attributes could produce better performance than purely low level features. While our focus is on attributes of images, not of faces, we pursue a similar direction to show that our high level describable attributes can be used to produce powerful classifiers for: estimation of aesthetic quality (Chapter 3.1), estimation of general interestingness (Chapter 3.2), and estimation of query specific interestingness (Chapter 3.3).

We demonstrate that classifiers trained on high level attribute predictions are more powerful than those trained on purely low level features for aesthetics and interestingness classification, and can be made even more accurate when trained on a combination of low level features and high level attributes (fig 2.1). This indicates that these attributes may capture some high level information about images that could be useful for a variety of general image related tasks. However, results of the tasks often depend on the dataset being used.

Our main contributions include a focus on extracting high level visual attributes of images (as opposed to objects), and novel attributes related to image layout. We explicitly train classifiers to estimate attributes, and evaluate the accuracy of these estimates. Much previous work contains related intuition about important high level attributes for aesthetics, but uses this intuition to design low level features instead of explicitly estimating the high level attributes. We also show that the estimated high level attributes significantly improve accuracy on predicting perceived aesthetic quality of DPChallenge photographs, as well as enabling[1] image based prediction of interestingness (a measure of social interaction) of Flickr images. Our final contribution is a method for estimating query specific interestingness.

## 1.1 Previous Work

Our work is related to three main current areas of research: estimating visual attributes, algorithms for estimating the aesthetics of photographs, and human judgments of aesthetics. We briefly outline work in each area.

**Attributes:** Recognizing attributes of objects in images can improve object recognition and classification as well as provide useful information for organizing collections of images. As an example, recent work on face recognition has shown that the output of classifiers trained to recognize attributes of faces – gender, race, age, etc. – can improve the process of face verification [15; 14]. The system was

[1] This is the first time, to our knowledge.

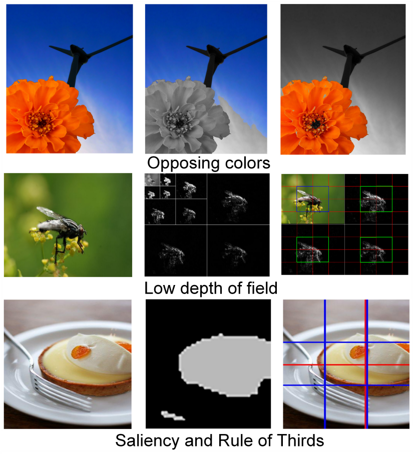


Figure 1.2: **Top** – "presence of opposing colors" attribute (left: original image, center and right: colors in the largest 2 color bins). **Center** "Low DoF" attribute (left: original image, center: wavelet transform, right: wavelet coefficients on 4x4 grid, and center surround computation). **Bottom** "salient object presence" and "rule of 3rds" attributes (left: original image, center: detected salient object region, right: centroid and conformity to rule of 3rds in red).

further used to design an image search engine based completely on faces. One can enter facial descriptions and the system would search over a space of 3.1 million face entries, which have been classified on the basis of discovered attributes, and return matching results. Other work has shown that learning to recognize attributes can allow recognition of unseen categories of objects from their description in terms of attributes, even with *no* training images of the new categories [16; 5; 8]. Our work is related to these methods for extracting attributes, but while they focus on attributes used to describe objects (e.g. "blond" for a person, or "red" for a car), we look at the problem of extracting high level describable attributes for general image content, lighting, and composition (e.g. "presence of animals" or "containing a salient object").

**Aesthetics:** Ideas of estimating the aesthetic quality of images have been explored in a few previous papers to produce classification engines that can differentiate

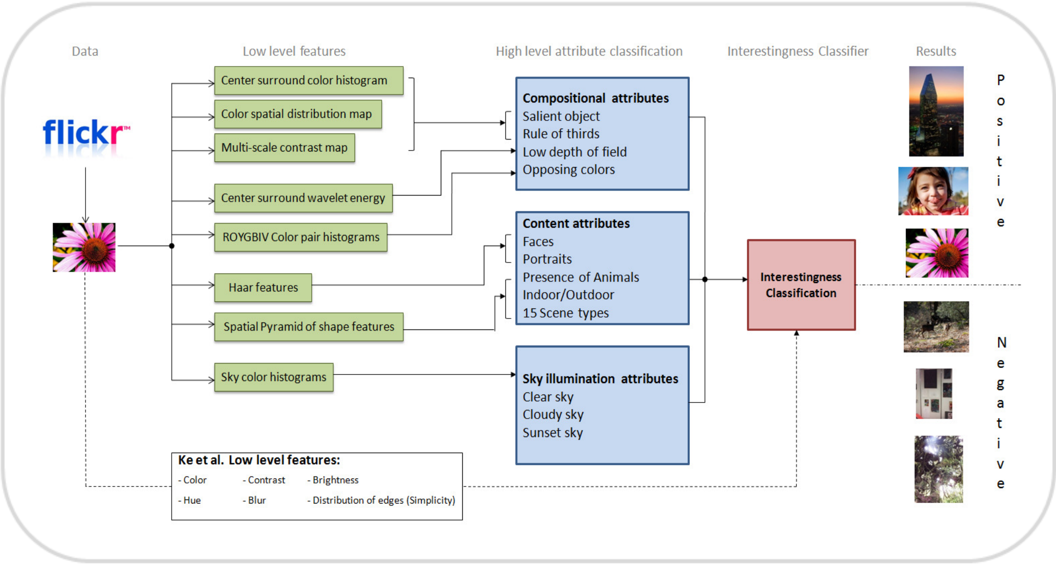
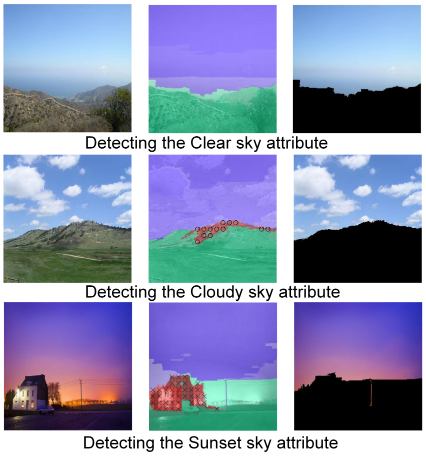


Figure 1.3: Overview of our method for estimating interestingness (estimating aesthetic quality follows a similar path). From left to right we show our system pipeline: a) an example input image b) low level features are estimated c) high level attributes are automatically predicted by our describable attribute classifiers, d) interestingness is predicted based on the high level attribute predictions (or optionally in combination with Ke et al low level classification is predicted.

between images captured by a professional photographer versus an amateur [29; 12; 30; 3; 28]. Previous work has utilized some nice intuition about how people judge aesthetic quality of photographs to design low level features that might be related to human measures. Experiments have also been carried to do attribute selection (using techniques like Boosting) to ascertain computational methods for predicting aesthetic measures. Datta et al select visual features based on artistic intuition and use these to train classifiers for predicting aesthetic quality [3]. Another approach to measure aesthetics in images is by measuring the emotional response evoked by an image in human beings. There has been work which measures aesthetics in terms of an image's emotional quality [4]. Tong et al use measures related to the distortion in photographs [30]. Ke et al select low level features such as average hue, or distribution of edges within an image, that may be related to high level attributes like color preferences or simplicity [12]. Sun et al [28] explore methods for incorporating a computational model of attention in quality assessment. They use face-sensitive saliency maps and a measure known as 'Rate of focused attention' for this purpose.

The main difference between these approaches and ours is that we explicitly train and evaluate classifiers to recognize useful high level describable attributes and use these predictions as input to train a second level classifier that estimates aesthetic quality. We also include some additional tasks: estimating general, and query specific interestingness. Research also shows other measures of image quality include harmony [6], which the authors describe as 'the pleasing or congruent arrangement of parts producing internal calm', image appeal [26], and value to the user [18].

**Human Judgment of Aesthetics:** The existence of preferred views of objects has long been known by Psychologists. In their seminal work, Rosch and Palmer found that humans agree on canonical views of objects and that recognition is faster for these views [23]. Photographers have also proposed a set of composition rules for improving the aesthetic quality of photos, but for the most part these rules are



Detecting the Clear sky attribute

Detecting the Cloudy sky attribute

Detecting the Sunset sky attribute

Figure 1.4: **Top** “clear sky” attribute, **Center** “cloudy sky” attribute, **Bottom** “sunset sky” attribute. For each sky-illumination attribute we show original photo (left), geometric context (center), and extracted sky region (right). ROYGBIV binned color histograms are used to train classifiers for each attribute.

only a set of guidelines passed down over time without any quantitative evaluation of their validity. More recently however, there have been some studies that expand the idea of view preferences to more general notions of human perception and judgment of aesthetics. These experiments include evaluating the role of color preferences [27; 21; 24] and spatial composition [9; 1] on human aesthetic judgment. Other work in computational neuroscience has looked at developing models of visual attention including ideas related to saliency [13; 11]. Some of our attributes are directly related to these ideas, including predicting the presence of opposing colors in images, and attributes related to the presence of salient objects, and arrangement of those objects at compositionally preferred locations.

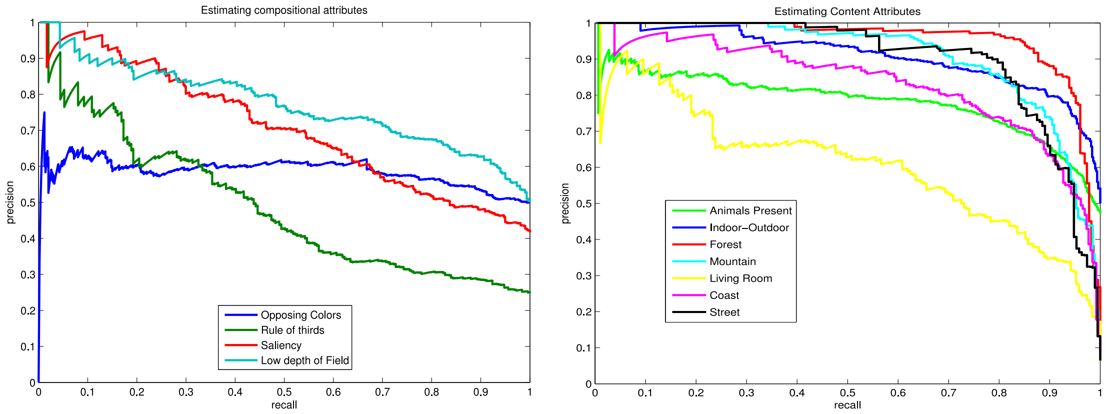


Figure 1.5: **Left:** Precision-Recall curves for Composition attributes including "Salient Object Present", "Follows Rule of 3rds", "Displays Opposing Colors". **Right:** Precision-Recall curves for Content attributes including "presence of animals", "indoor-outdoor" and a variety of "scene type" attributes.

## 1.2 Overview of Approach

The first phase of our work consists of developing high level attribute classifiers. We do this by collecting positive and negative example images for each attribute, picking an appropriate set of low level features, and training classifiers to predict the presence or absence of the attribute in images. This process is described in Chapters 2.1, 2.2, 2.3.

Next we demonstrate that these high level attribute classifiers provide useful information for predicting rankings of images based on perceived aesthetic quality (for DPChallenge) and "interestingness" (for Flickr images). For each application, a set of training images is collected consisting of highly ranked images as positive examples and low ranked images as negative examples. A ranking classifier is trained using the output of the high level attribute classifiers we developed as features. The ranking classifiers are then evaluated on held out data. We also show results for training the ranking classifiers using only low level features, and using a combination of low level features and our high level attributes. For most of our datasets, we observe that the classifier with only low-level features perform the worst, while the classifiers with both low-level features and high-level attributes perform the best. This aligns with our expectations as we expect the high-level attributes will be able to encapsulate considerably more information than the low-level features, while adding the high-level attributes to the low-level features adds more useful information to it. Results on aesthetics for DPChallenge are in Chapter. 3.1 and interestingness for Flickr are in Chapter. 3.2. Finally we show results on ranking for interestingness of specific queries in Chapter. 3.3. The specific query based classifier tests the genericness of our set of attributes. Not all of the attributes are useful for all categories of images. However, we show that the error percentages while using a generic set of attributes is not too high as compared to classification using query specific classifier.

# Chapter 2

# Describable attributes

We have developed high level describable attributes to measure three types of image information: attributes related to the layout or composition of an image (Chapter 2.1), attributes related to image content (Chapter 2.2), and attributes related to the sky-illumination present in an image (Chapter 2.3). Some images with automatically predicted attributes are shown in figure 1.1.

## 2.1 Compositional Attributes

The first kind of attribute we develop are attributes related to image composition. Composition is the positional arrangements of objects according to the way they are placed in the image. Composition also takes into account the color tones used in the image. The common questions it answers are, are there many objects present or a single salient subject? Does the image contain many different colors or mainly display a few highly contrasting colors? These attributes also correspond to several well known photographic rules of composition.

We design 4 describable attributes related to composition:

- Presence of a salient object – a photo depicting a large salient object, well separated from the background.
- Rule of Thirds – a photo where the main subject is located near one of the dividing third-lines.
- Low Depth of Field – the region of interest is in sharp focus and the background is blurred.
- Opposing Colors – a photo that displays color pairs of opposing hues.

**Presence of a salient object:** We define images containing a salient object as those that depict some large object, well separated from its background. For this attribute we would like to predict whether an image contains some highly distinctive salient object. To do this we take advantage of recent developments in automatic top down methods for predicting locations of salient objects in images [20]. In addition to the original evaluation presented in this paper – they evaluate localization accuracy only on images already known to contain salient objects – we demonstrate that the outputs of this predictor can be effectively used to classify images according to whether they display a salient object.

For this attribute, we have implemented 3 features related to saliency: a multi-scale contrast map, a center surround histogram map, and a center weighted color spatial distribution map. Though for a detection task there are many possible windows where a salient object could be located, the otherwise computationally intensive center surround histogram map can be computed efficiently for all windows using integral histograms [25]. All three of these feature maps are supplied to a conditional random field (CRF) algorithm to train a salient object detector. (an example output saliency map is shown in figure 1.2, bottom row center image). Finally, we use the free energy output of the CRF to predict presence or absence of a salient object in images.

We evaluate classification accuracy on this task using a set of 1000 images that have been manually labeled as to whether they contain a salient object. Precision-recall curves for predicting the presence of a salient object are shown in figure 1.5 (left plot, red curve), showing that our salient object classifier is quite accurate at predicting the presence of a salient object.

**Rule of thirds:** The rule of thirds is a common compositional rule in photography. If you consider two vertical lines dividing the image horizontally into 3 equal parts, and two horizontal lines dividing the image vertically into 3 equal parts, (blue lines shown in figure 1.2, bottom row right image), then the rule of thirds suggests that it will be more aesthetically pleasing to place the main subject of the picture on one of these lines or on one of their intersections.

For our rule of thirds attribute, we again make use of the salient object detector. We calculate the minimum distance between the center of mass of the predicted saliency mask and the 4 intersections of third-lines. We also calculate the minimum distance to any of the third-lines. We use the product of these two numbers (scaled to the range [0,1]) to predict whether an image follows the rule of thirds and evaluate this attribute on images manually labeled as positive examples if they conform to the rule of thirds and as negative examples otherwise. Precision-recall curves are shown in figure 1.5 - left plot, green curve.

**Low depth of field:** An image displaying a low depth of field (DoF) is one where objects within a small range of depths in the world are captured in sharp focus, while objects at other depths are blurred. This effect is often used in photography to emphasize a region or object of interest in an image, and is especially common in macro photos. For our low depth of field attribute we train a classifier to differentiate between images displaying a low depth of field from those not showing this effect. We utilize Daubechies wavelet based features indicative of the amount of blurring present [3].

The wavelet transform is applied to the image and then we consider the third level coefficients of the transformation in all directions. (fig 1.2, middle row center image). If you consider a 4x4 grid over the image, we divide the sum of the coefficients in the four center regions by the sum of coefficients over all regions. (fig 1.2, middle row right image). This gives us a vector of 3 numbers, one for each direction of the transformation. An image with a low DoF in its center region will produce larger values than one not displaying a low DoF. We use these values to train an SVM classifier for predicting whether an image has a low DoF.

A manually labeled dataset of 2000 images from Flickr and Photo.net is used to train and test our classification algorithm – where positive examples display low DoF, and negative examples do not. Precision-recall curves for the low DoF attribute are shown in figure 1.5 - left plot, cyan curve. We can reliably detect whether an image displays a low DoF.

**Opposing colors:** Color plays an important role in perception [27]. Some color singles, pairs, or triples are more pleasing to the eye than others [24], giving rise to the opposing colors rule in photography which says that images displaying contrasting colors (those from opposite sides of the color spectrum) will be aesthetically pleasing. For this attribute we train classifiers to recognize the presence of opposing colors in images using an image representation based on the presence of color pairs. We first discretize pixel values into 7 possible values corresponding to the ROYGBIV spectrum. Figure 1.2, top row right two images shows the two largest bins for an example image. We then build a 7x7 histogram based on the percentage of each color pair present in an image.

We train an SVM classifier on these features to detect opposing colors. For training and testing data we use 1000 hand-labeled images from Flickr – here positive examples are images displaying strong opposing colors, and negative examples are images not

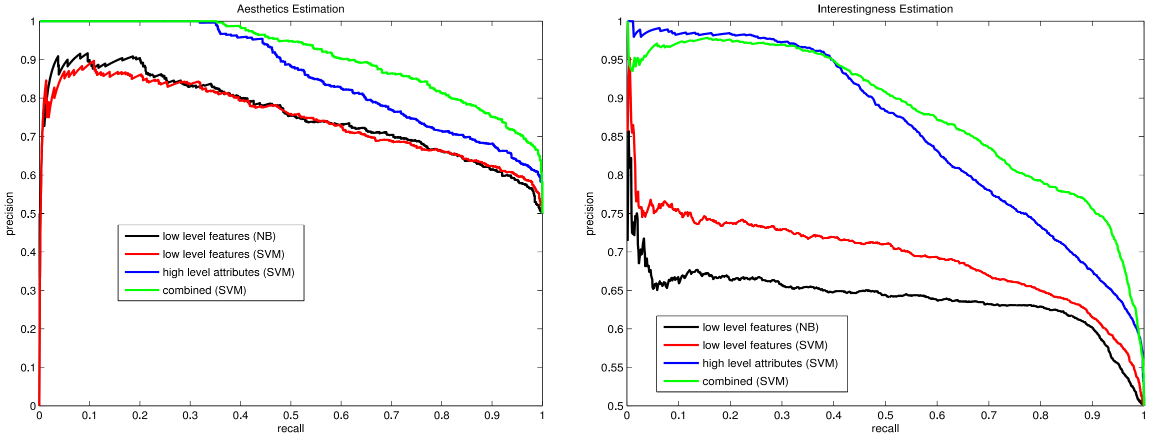


Figure 2.1: Left shows: Precision-Recall curves for aesthetics estimation on the DPChallenge dataset using low level image features with Naive Bayes classification, corresponding to the previous method by Ke et al [12] (black), low level image features with SVM classification (red), high level describable attributes with SVM classification (blue), and a combination of low level features and high level attributes with SVM classification (green). Right shows: Precision-Recall curves for interestingness estimation on Flickr images (curves are averaged over 6 specific search queries, and one general set of photos. For both tasks our high level attributes (blue) produce more powerful classifiers than the previous state of the art method (black), and can provide complimentary information when used in combination with low level features (green).

displaying strong opposing colors. Classification accuracy is shown in figure 1.5 - left plot, blue curve. Our classifier for this attribute is not particularly strong because even images containing opposing colors that are obvious to the viewer may not present with large peaks in their color pair histogram, but it still provides some useful signal for our aesthetics and interestingness classifiers (Secs 3.1, 3.2).

## 2.2 Content Attributes

Ideally, given a photograph, we would like image analysis algorithms to tell us exactly what is depicted within the picture. This might include descriptions of what objects are present, what kind of general scene is shown, or specific descriptive qualities of the image. We present a set of high level content attributes that are at the forefront of recognition technologies and that can be reliably recognized in images.

The content attributes we demonstrate are:

- Presence of people – a photo where faces are present.
- Portrait depiction – a photo where the main subject is a single large face.
- Presence of animals – whether the photo has animals.
- Indoor-Outdoor classification – whether the photo was captured in an indoor setting,
- Scene type – 15 attributes corresponding to depiction of various general scene types (e.g. city, or mountain).

**Presence of people:** The "presence of people" attribute provides a rough measure about whether people are present in a photograph. We use the OpenCV face detector [31] to estimate whether any faces are present in an image. This detector has been trained on a large amount of hand labeled face and non-face regions and can reliably detect faces in novel images. For this attribute we output a binary value (1, if faces have been detected, and 0 otherwise). This will be a rather modest estimate for the presence of people in images since there will be many photographs containing people but no visible faces. We label a test dataset of 2000 images from Photo.net for the presence or absence of faces to evaluate our classifier and obtain an accuracy of 78.9%.

**Portrait depiction:** Our "portrait depiction" attribute predicts whether the main theme of a photograph is a human face. This is also estimated through face detection, but images are only classified as portraits given the presence of a large detection. Here we use the face detector with a search window of one fourth the minimum size dimension of the image so that we capture only large faces. We evaluate this feature on 5000 images from Photo.net hand labeled as portrait or non-portrait images and obtain an accuracy of 93.4%.

**Object and scene attributes:** For the attributes denoting the "presence of animals", "indoor-outdoor classification" and "scene type" we train 17 SVM classifiers to recognize each individual attribute (1 classifier for "presence of animals", 1 for "indoor-outdoor", and 15 classifiers to recognize various scene categories). Since spatial pyramid matching (SPM) has been shown to produce good recognition results for both scene classification and object category recognition [17], we utilize this image representation using code provided by Svetlana Lazebnik [17].

The spatial pyramid representation partitions the image into a hierarchy of increasingly fine windows and computes a histogram of the local features found inside each window. This resulting spatial pyramid provides a way to quickly approximate the all-pairs correspondence between the bags-of-features present in two images by building the spatial matching cost into a hierarchical representation. Here matching features at the global level is weighted less than matching at finer levels of the pyramid. For our particular implementation the histograms are computed on visual dictionaries of local shape features, specifically SIFT features [19] captured on a uniform grid across the image with region size 16x16 and spacing of 8 pixels. The SIFT features for 100 random images are clustered to form a single visual dictionary which is used for all of the content attribute types. Using the spatial pyramid representation, we then train SVM classifiers for each attribute type using an intersection kernel (the bin-wise minimum between the spatial pyramid for each image).

For the "presence of animals" attribute we train a classifier on images from the Animals on the Web dataset [2]. This dataset contains images collected from the web depicting 10 animal categories from "alligators" to "dolphins" to "penguins". Images are hand labeled as depicting the category or not depicting the category. For our task we select positive training images randomly from each of the categories so that we have up to 200 images from each category (less may be selected if the category has fewer than 200 positive examples). For the negative training set we randomly select 200 images labeled as negative examples for each of the animal categories. We repeat this process for the test set. Precision-recall curves for predicting the presence of animals are shown in figure 1.5 - right plot, green curve. Though it is well known that recognition of specific animal categories is a very challenging problem [2], we do quite well at predicting whether *some* animal is present or not in an image.

For the "indoor-outdoor classification" attribute we collect 1000 images for training and 1000 images for testing from Flickr, half showing indoor scenes, and half showing outdoor scenes. Precision-recall curves for indoor-outdoor classification are shown in figure 1.5 - right plot, blue curve. The indoor-outdoor classifier is quite accurate for most images.

For the 15 "scene type" attributes we use a labeled scene category dataset, originally collected by Oliva & Torralba [22] and later expanded on by Fei-Fei and Perona [7], and Lazebnik et al [17]. This dataset contains images depicting various scene types ranging from kitchen, or living room, to mountain, or city. For each scene category we randomly sample 100 images for the training set, and use the rest for testing. Precision-recall curves for 5 of our scene types are shown in figure 1.5 - right plot (the remaining scene curves are similar, but removed for clarity of presentation). Outdoor scene types such as forest, mountain, or street tend to be more accurate than indoor scenes such as living room, or kitchen.

## 2.3 Sky-Illumination Attributes

Another kind of information we might like to extract from images are attributes related to illumination. The lighting present in a scene can greatly effect perception of an image. In addition, images captured with interesting lighting conditions such as indirect lighting can be more aesthetically pleasing. Because estimating indoor illumination is still a very difficult open research problem, we focus on estimating natural outdoor illumination.

We provide sky-illumination attributes of 3 broad types:

- Clear skies – photos taken in sunny clear conditions.
- Cloudy skies – photos taken in cloudy conditions.
- Sunset skies – photos taken when the sun is low in the sky.

To train our sky attribute classifiers we first extract rough sky regions from images using Hoeim et al's work on geometric context [10]. This work automatically divides image regions into sky, horizontal, and vertical geometric classes using adaboost on a variety of low level image features. On the portion of an image labeled as sky (shown in fig 1.4) we compute 3d color histograms in HSV color space, with 10 bins per channel. These histograms are used to train an SVM classifier to recognize each sky attribute. Each classifier is trained and tested on 1000 hand labeled images from Flickr, where positive examples display the attribute class, and negative examples are sampled from the other sky-illumination attribute classes.

The accuracies of the three sky-illumination attribute classifiers are: "clear skies" 99%, "cloudy skies" 91.5%, "sunset skies" 96.7%. These classifiers produce extremely accurate classifications because skies of these 3 types tend to look quite different, producing a large separation between the descriptors computed for each attribute type.

# Chapter 3

# Estimating Aesthetics & Interestingness

## 3.1 Aesthetics

The first task we apply our describable attributes to is the problem of estimating the aesthetic quality of an image. Aesthetic quality could be defined as the opinion a photograph generates in the mind of the viewer. A positive opinion is often generated when the photograph pertains to the golden rules of photography or uses the right combination of colors. Sometimes even the subject of the image could solely change the aesthetic quality of the image. A negative opinion is often generated when the image has a very poor quality. Here the task is to design a classifier to differentiate between images of high photographic quality from images of (low) snapshot quality. To estimate this classification we train an SVM using the outputs of our high level describable attribute classifiers as our input image feature representation. This is a 26 dimensional feature vector where each value represents our prediction for a particular high level describable attribute. Figure 1.3 shows our method pipeline.

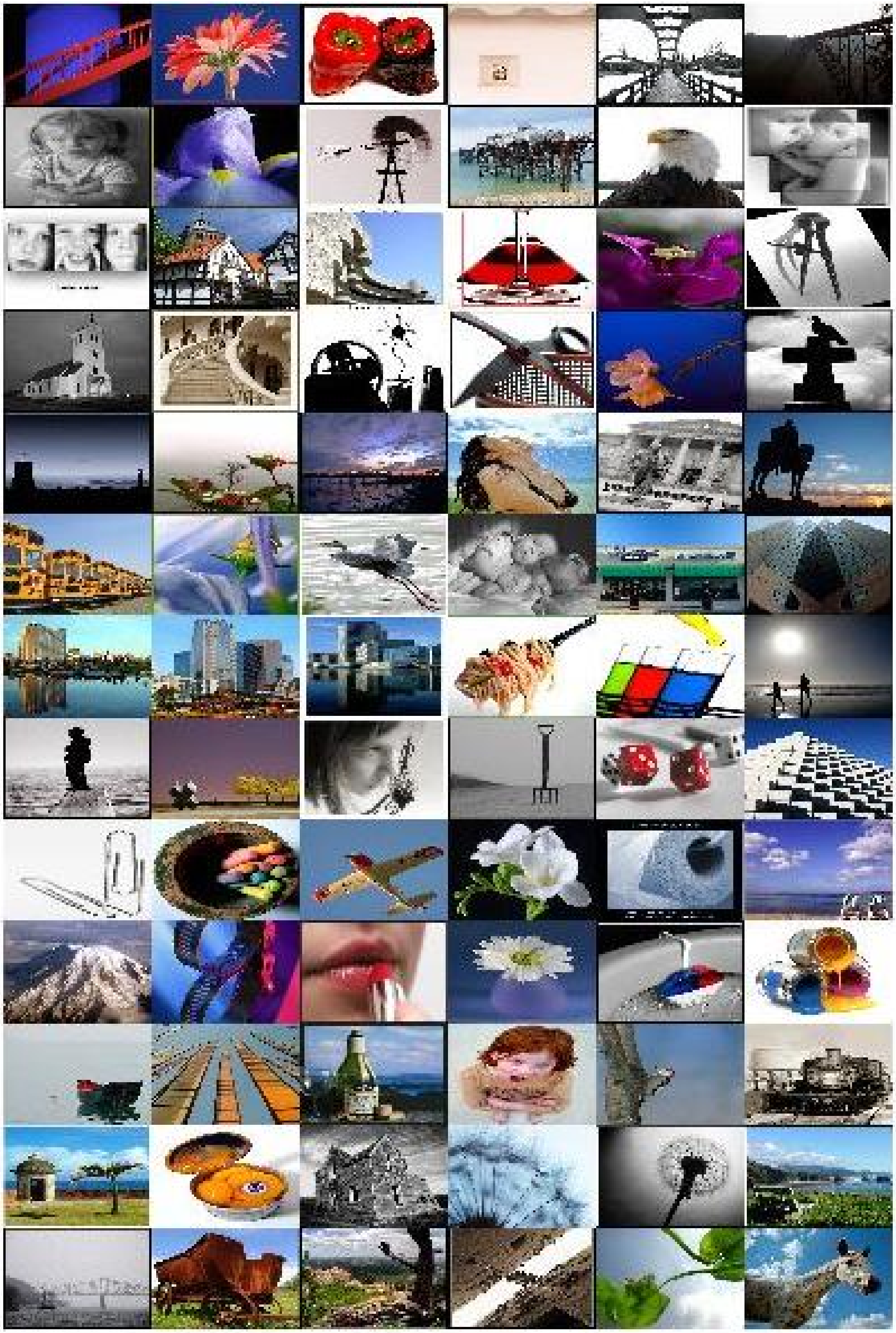


Figure 3.1: DPChallenge photos ranked by aesthetics. This figure shows the top 78 images as ranked by our aesthetics classifier.

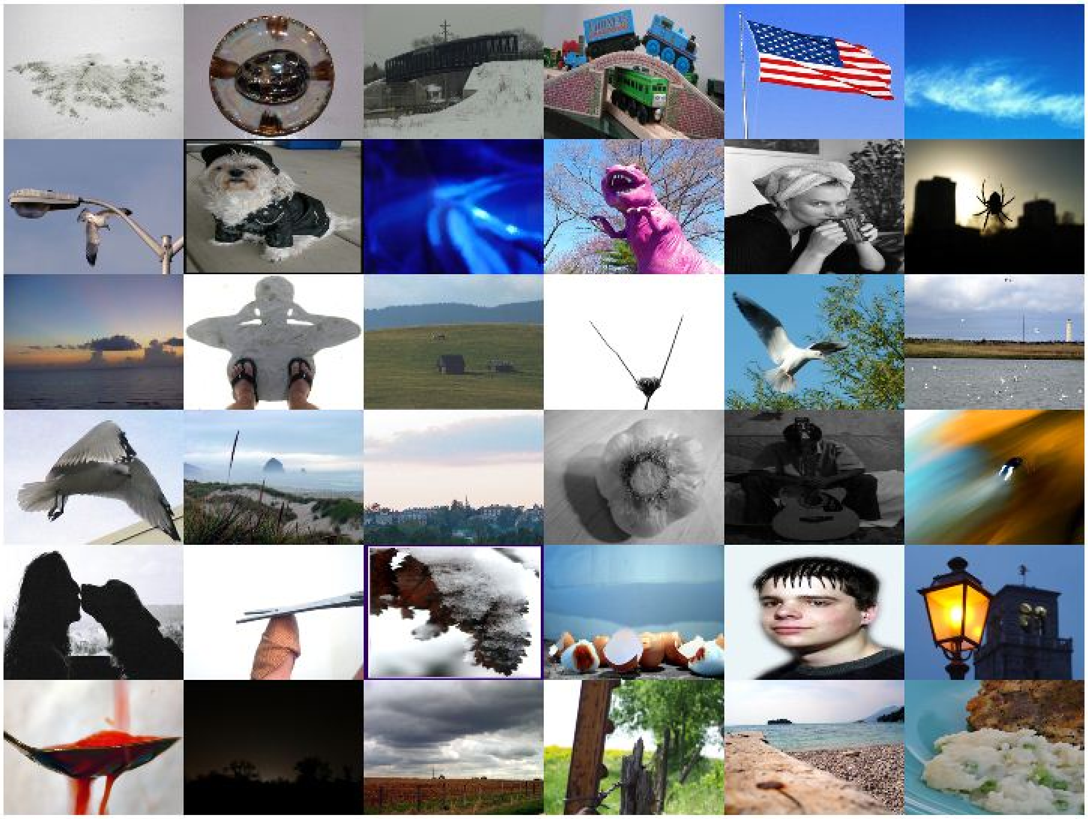


Figure 3.2: DPChallenge photos ranked by aesthetics. This figure shows the bottom 36 images as ranked by our aesthetics classifier.

**Experiments:** In our approach, we evaluated images for training and testing this classifier. We collect a large dataset of 16,000 images from the DPChallenge website[1]. These images have been quantitatively rated by a large set of human participants (many of whom are photographers). We label the top 10% rated photos as high aesthetic quality, and the bottom 10% as low aesthetic quality photos. This is to prevent the noise due to ambiguity in human consensus affecting the training process, Half of each of these sets is used for training the classifier, while the remaining half is used for evaluation.

[1] http://www.dpchallenge.com/

Though the attributes we develop are general measures of visual information and could be used for many different problems, we find that our high level attributes produce powerful classifiers for this task (fig 2.1 left in blue).

For comparison we also re-implement the state of the art aesthetics classifier used in Ke et al [12]. We show results of their original Naive Bayes classification method (figure 2.1, left plot black curve) and also train an SVM on their low level features (fig 2.1, left plot red curve). Our high level attributes produce a more accurate ranking than the previous approach, and when used in combination with these low level features can produce an even stronger classifier (fig 2.1 left plot, green curve). This suggests that our high level attributes are providing a source of useful complimentary information to purely feature based approaches.

## 3.2 General interestingness

We also apply our describable attributes to a related, but deceptively different problem of estimating interestingness in photos. While DPChallenge directly measures aesthetic quality through user ratings of photos, Flickr's "interestingness" measure[2] is computed more indirectly. Here interestingness translates to a measure of whether a photo is of interest to many different users. This value is computed through analysis of social interactions with that photo (viewing patterns, popularity of the content owner, favoring behavior, etc). Nowadays many online social networking systems exist which involve the user into judging photographs belonging to people in their social networks. This judgement could be a measure of the image's aesthetic quality or could be a decision based on the content of the photograph (I would have a high chance of liking a group photograph of my best friends or a photograph of my favorite celebrity). Computationally predicting this value is a tougher problem as it is

[2] http://www.flickr.com/explore/interesting/7days/

not obvious what attributes the human mind uses to come up with such a decision. In our system, we again train an SVM classifier to predict interestingness using our describable attribute classifications as input (fig 1.3 shows our method pipeline).

**Experiments:** For our general interestingness classifier we collect a dataset from Flickr using interestingness-enabled Flickr searches that return images ranked by interestingness score. Using time limited querying, we obtain 40,000 Flickr images sorted by interestingness. The top 10% of these images are used as positive examples for our interestingness classifier, while the bottom 10% are used as negative examples. We split this set into half for training, and testing.

We train our interestingness classifier on the high level attributes predicted by our describable attribute classifiers, a 26 dimensional input feature vector (fig 2.1, right plot blue curve). For comparison, we also train an interestingness classifier on the low level features used in Ke et al [12] using their original Naive Bayes approach (fig 2.1, right plot black curve), and using an SVM classifier (fig 2.1, right plot red curve). Lastly, we train a combined classifier on their low level features and our high level attribute classifications, a 32 dimension input feature vector (fig 2.1, right plot green).

In figure 3.3 we show images ranked by automatically predicted interestingness score. The top 5 rows show the 50 highest ranked images, and the bottom 2 rows show the 20 lowest ranked images. Lower ranked images show lower quality than higher ranked images.

We also evaluate our method quantitatively. Precision-recall curves are shown in figure 2.1 - right plot. Our method performs quite well at estimating interestingness. The high level attributes produce a powerful classifier for predicting interestingness (fig 2.1 blue curve), and improve slightly with the addition of low level features (fig 2.1 green curve). Compared to aesthetics classification, for interestingness classification our method shows an even larger increase in performance over the previous approach

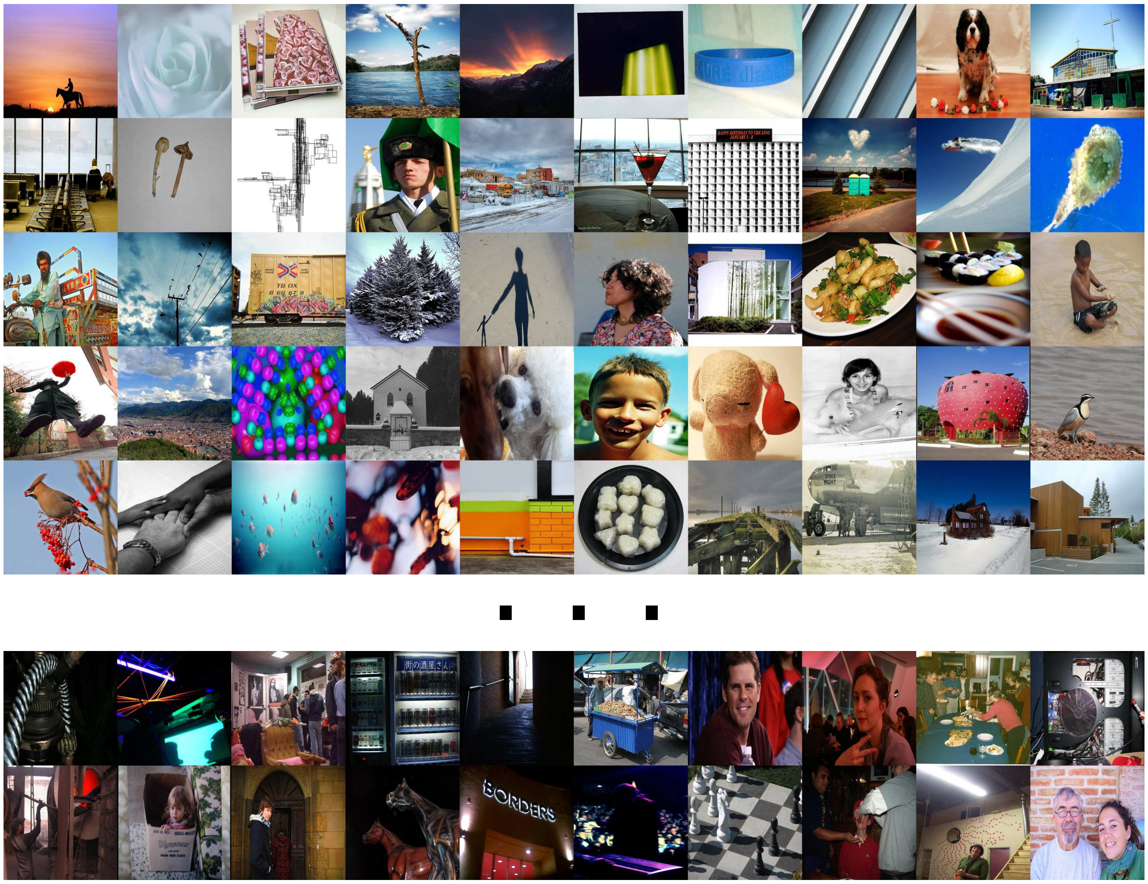


Figure 3.3: General Flickr photos ranked by interestingness. Top 5 rows show the first 50 images ranked by our interestingness classifier. Bottom two rows show the last 20 images ranked by our interestingness classifier.

(fig 2.1, black vs blue curves). We posit that the larger increase for interestingness as compared to aesthetics may be due to differences between Flickr and DPChallenge images. The images on DPChallenge are often posted by photographers who might apply a great deal of post-processing to their pictures to adjust color levels etc, while those on Flickr tend to be the results of more amateur photographers taking nice pictures. Perhaps the low level features are able to utilize these post-processing variations for aesthetics prediction.

## 3.3 Query specific interestingness

Lastly, we introduce a method to produce query specific classifiers to predict interestingness. In general we expect some of our attributes to be more useful for predicting interestingness than others. We also expect that the usefulness of an attribute might vary according to the specific search query used to collect images – e.g. low DoF may be more useful for predicting interestingness of images returned for the query "insect" than for the query "beach".

**Experiments:** We collect a dataset of images from Flickr using 6 different query terms: "beach", "building", "car", "horse", "insect", and "person", retrieving 20,000 images for each query ranked by interestingness. Images in the top 10% are labeled as positive examples, and images in the bottom 10% are labeled as negative examples. Again, half of the images are used for training and half for testing.

For each query we train an interestingness predictor on images returned for the query. We then evaluate the accuracies of using our general interestingness classifier and using our query specific classifiers to rank images from the held out test set. For some queries, the query specific classifiers outperform the general interestingness classifier (classification errors are shown in fig 3.4), indicating that the importance of individual attributes may differ by query.

Ranked results for some of our query specific interestingness classifiers are shown in figure 3.5. The top 3 rows for each query show the 30 most highly ranked images for that query. The bottom row for each query shows the 10 least highly ranked images for that query. For "beach" at the top of the ranking we observe very beautiful, clear depictions, often with pleasant sky illuminations. At the bottom of the ranking we see more cluttered images often displaying groups of people. These are much less "interesting" than those at the top of the ranking. For the insect query, the top of the ranking shows pretty images where the insect is the main subject of the photograph,

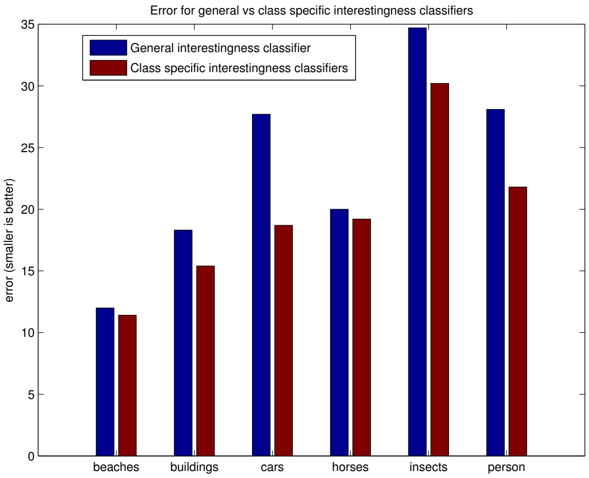


Figure 3.4: Error rates for interestingness prediction of images returned for various query terms. Error rates for our general interestingness classifier are in blue, while error rates for query specific interestingness classifiers are in red. For some categories, the query specific classifier has a significantly lower error rate than the general one, indicating that attribute importance may vary by search query.

and a low DoF is often used to emphasize the importance of the subject. The bottom of the ranking shows depictions that are not as attractive. For car, top rankings show much cleaner depictions than those at the bottom of the ranking.

This experiment has direct applications in 'Image Search' systems where the user's search query could be classified to fall into a particular category. Once the broader category is known, the most appropriate interestingness classifier could be used to re-rank the search results to bring to top the best possible results for that particular category. This would be a better estimation of interestingness as the attributes which are most useful in being able to differentiate between interesting and non-interesting photos of that particular category has been used.

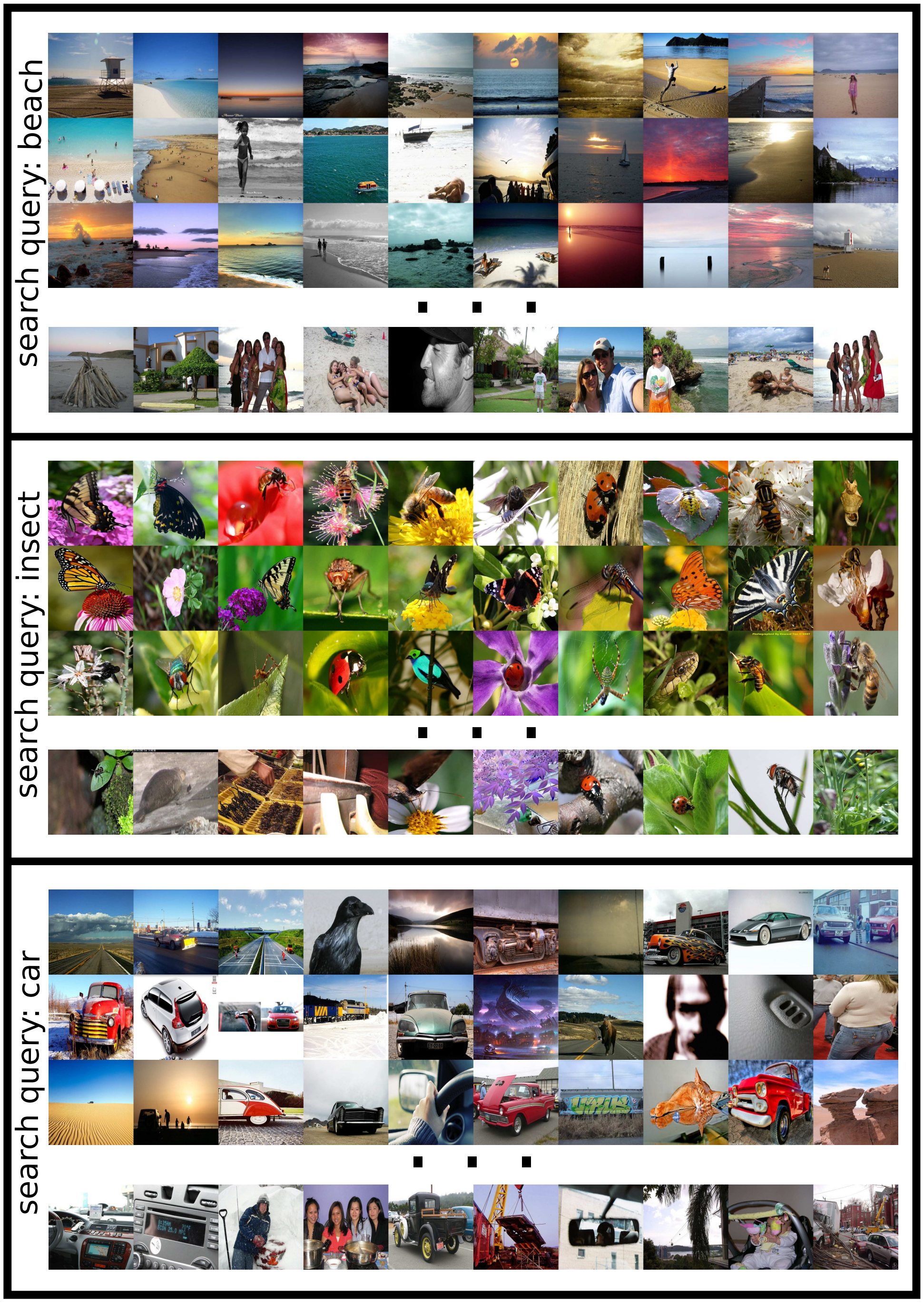


Figure 3.5: Flickr photos ranked by automatically predicted interestingness for search terms: “beach” (top), “insect” (middle), “car” (bottom). Top three rows for each query show the most highly ranked images by our interestingness classier. Bottom rows for each query show the least highly ranked images by our interestingness classifier.

# Chapter 4

# Classifier Analysis

In this chapter, we analyze how a Decision-tree based SVM classifier would have performed with respect to a native Support Vector Machine(SVM). We also go on to study aspects of the SVM kernel to be able to infer which attribute or attribute-pair(s) play an important role in the final classification of aesthetics or interestingness.

## 4.1 Decision-Tree based SVMs

We have two binary features in our 26 attribute-long score vector, the 'presence of people' and the 'presence of Portraits'. Initially we treat these features as similar to the non-binary attributes, but we could use the binary features as prior information to the classification process. To do this, we adopt a Decision Tree based approach, where we use the 4 possible states of the 2 binary attributes to cluster our training set into 4 sets. These 4 sets are individually trained to get 4 models for each of the 4 possible states. Once we have 4 models, we can feed test samples to the models according to which state they fall in. In this way, we are training our SVMs on more specific knowledge, and in the ideal case it is expected to perform better. In our case, the performance is almost similar to the non-Decision Tree based approach, as seen in Fig. 4.1. In some cases, there is a possibility of the model giving lesser accuracies. This could happen when the information used to train the model is too specific and the model suffers from the phenomenon called ‘Overfitting’.

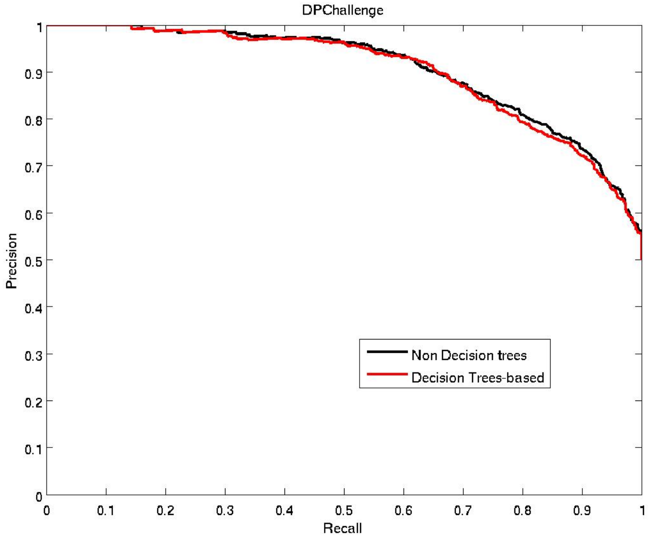


Figure 4.1: Comparative Analysis of Decision-tree based SVM and non-Decision-tree based SVMs for classification. For the Decision-tree based approach, we consider 2 binary attributes, ‘Presence of people’ and ‘Presence of Faces/Portraits’ which lead to 4 possible states and hence 4 possible SVMs.

## 4.2 Radial-Basis-Function Kernel & Polynomial Kernel results

In this section, we analyze the Aesthetics classification process for the DPChallenge.com dataset with various Support Vector Machine kernels and also try to infer a comparison of the contribution of attribute or attribute-pair for the classification task. Fig. 4.2 shows the previously presented results using a Radial Basis Function

Kernel with the parameter Gamma equal to 1 and regularization parameter equal to 1.

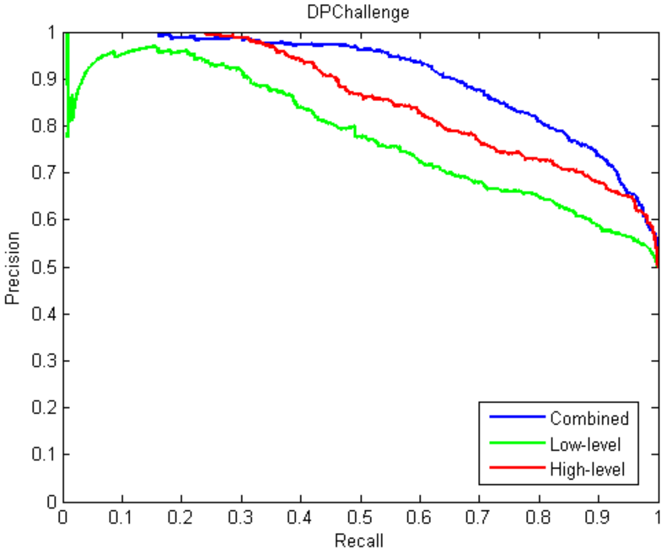


Figure 4.2: Classification using the RBF kernel, $K(X, Z) = e^{-|X-Z|^2}$, Accuracy (Combined features) = 80.375% (1286/1600), Accuracy (Low Level features) = 68.5625 % (1097/1600), Accuracy (High Level features) = 74.6875% (1195/1600)

By using the fact that the RBF function can be expanded using Taylor series, we know there should be a polynomial kernel that can approximate this RBF kernel. In particular for the given parameter of the RBF kernel we are using,

$$e^{-|U-V|^2} = \sum_{k=0}^{\infty} (-1)^k * \frac{|U-V|^{2k}}{k!}$$

We obtain an equivalent Polynomial kernel for the Radial Basis Function kernel we are using for the task of inferring Aesthetics on a DPChallenge.com dataset. This is done so as to obtain the weights of the coefficients of the SVM, so that we can rank the features according to their importance in the classification problem.

As shown in Fig. 4.3 using a polynomial of degree two gives us a classifier that performs almost as good as the one using the RBF Kernel. We are still using the same regularization parameter and the same training and testing sets.

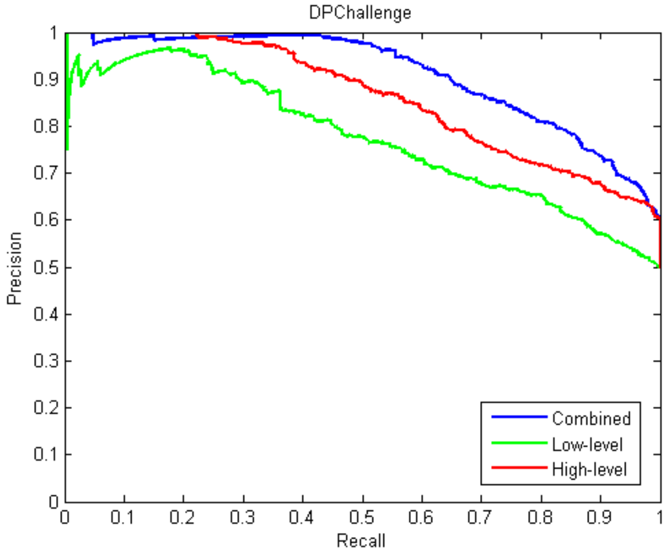


Figure 4.3: Classification using the Polynomial kernel, $K(X, Z) = (1 + XZ)^2$, Accuracy (Combined features) = 80.4375% (1286/1600), Accuracy (Low Level features) = 68.4375 % (1097/1600), Accuracy (High Level features) = 74.25% (1195/1600)

Fig. 4.4 shows the results of the two kernels for the high level attributes only on the DPChallenge Dataset.

## 4.3 Attribute contribution for classification

The contribution of each attribute or a pair of attributes to the process of classification can be obtained from the coefficients of the terms in the kernel function. The polynomial kernel,

$$K(x, z) = (1 + x.z)^d = \sum_{t=0}^{a} \binom{d}{t} (x.z)^t = \sum_{t=0}^{a} \binom{d}{t} (\sum_{j=1}^{n} x_j z_j) i$$

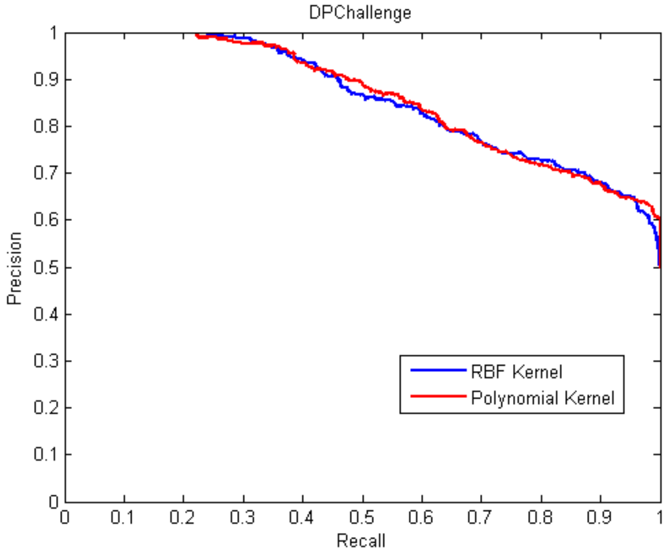


Figure 4.4: Comparison of classification using the Polynomial kernel, $K(X,Z) = (1 + XZ)^2$ and the RBF Kernel, $K(X,Z) = e^{-|X-Z|^2}$

$$
\begin{aligned}
&= \sum_{t=0}^{d} \binom{d}{t} \sum_{i_1,i_2,\ldots i_n} \binom{t}{i_1, i_2, i_3 \ldots i_n} \prod_{j=1}^{n} x_j^{i_j} z_j^{i_j} \\
&= \sum_{t=0}^{d} \sum_{i_1,i_2,\ldots i_n} \binom{d}{l} \binom{t}{i_1, i_2, i_3 \ldots i_n} \prod_{j=1}^{n} x_j^{i_j} z_j^{i_j} \\
&= \sum_{t=0}^{d} \sum_{i_1,i_2,\ldots i_n} (\sqrt{\binom{d}{t} \binom{t}{i_1, i_2, i_3 \ldots i_n}} \prod_{j=1}^{n} x_j^{i_j}) (\sqrt{\binom{d}{t} \binom{t}{i_1, i_2, i_3 \ldots i_n}} \prod_{j=1}^{n} z_j^{i_j}) \\
&= \sum_{t=0}^{d} \sum_{i_1,i_2,i_3 \ldots i_n} \phi_{l,i}(x) \phi_{l,i}(z) = \phi(x).\phi(z)
\end{aligned}
$$

In our particular case for our polynomial kernel the degree is 2, so d=2 and the number of high level features is 26, so n=26. The weights per polynomial term, per degree of the term are given by the formula below.

$$w_{t,i} = \sum_{k \in SVs} \alpha_k y_k \phi_{t,i}(x_k)$$

Fig. 4.5 is a plot of the weights for every pair term $(t, i)$ using guidelines that separate the range of terms depending on the form of the term. The first segment corresponds to all the linear terms, the second are all the quadratic terms and the following terms are the multiplication of two features, so the third segment is the first feature multiplied by every other feature, the forth term is the second feature multiplied by every other feature except the first feature, and so on.

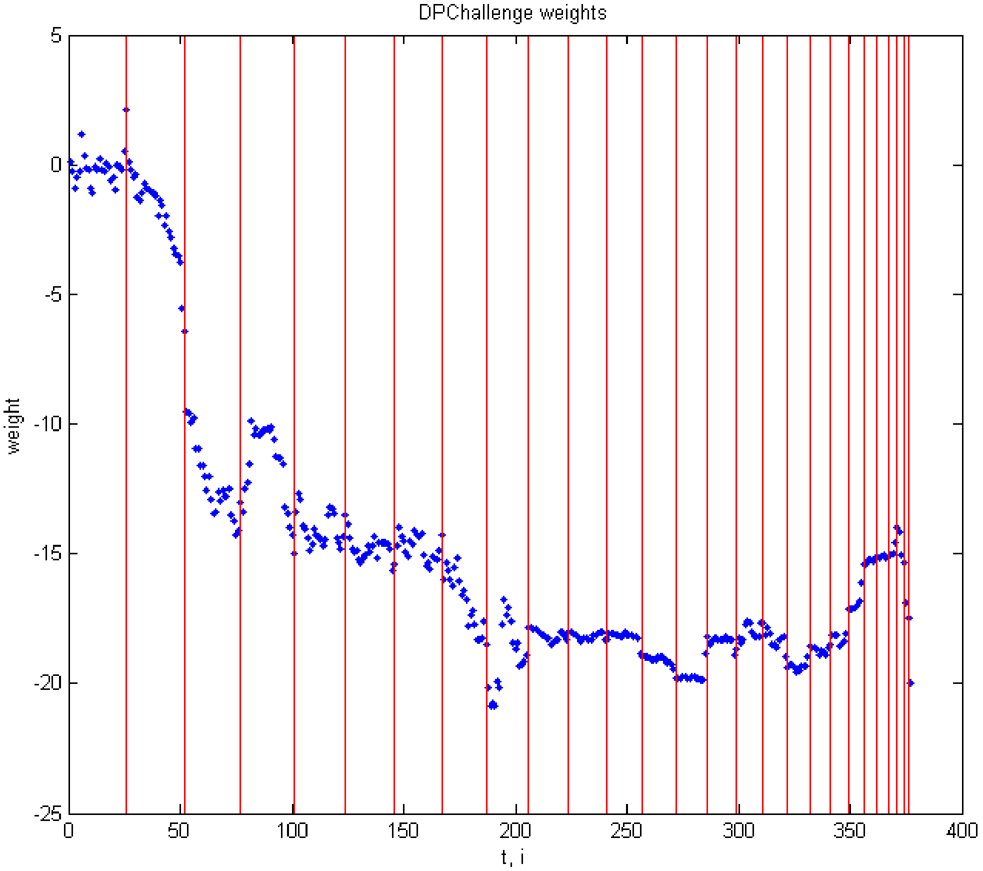


Figure 4.5: Weights $w(t, i)$ for every term $(t, i)$ in the polynomial transformation

Fig. 4.6 shows a plot of the weights calculated above but in sorted order, so that we can see how the contribution of each of the weights decreases.

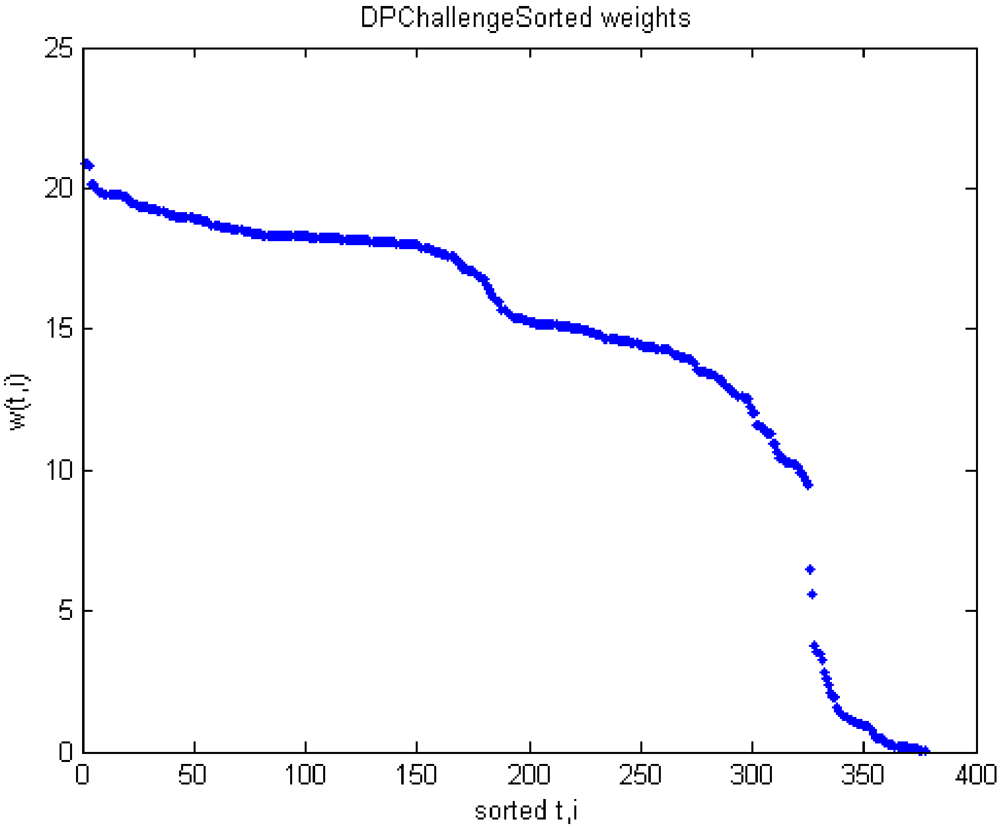
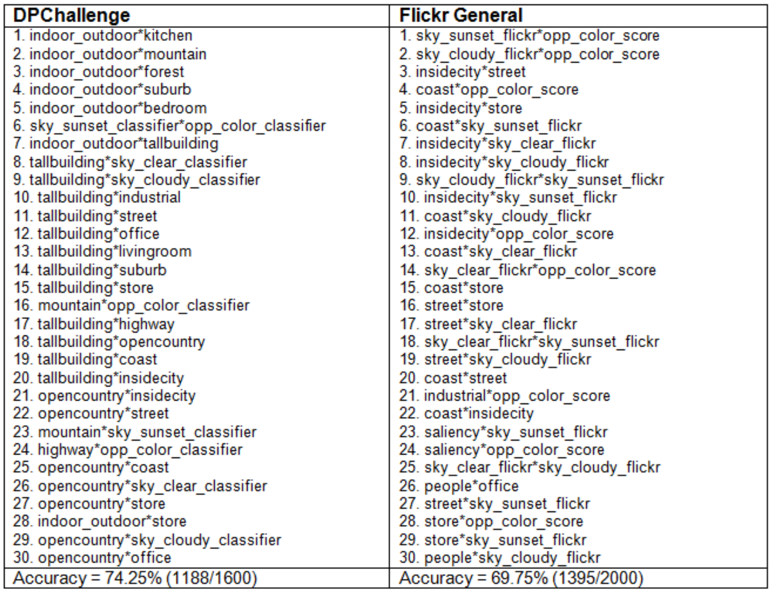


Figure 4.6: Weights $w(t,i)$ for every term $(t,i)$ sorted in descending order to show the decay behavior of the weights.

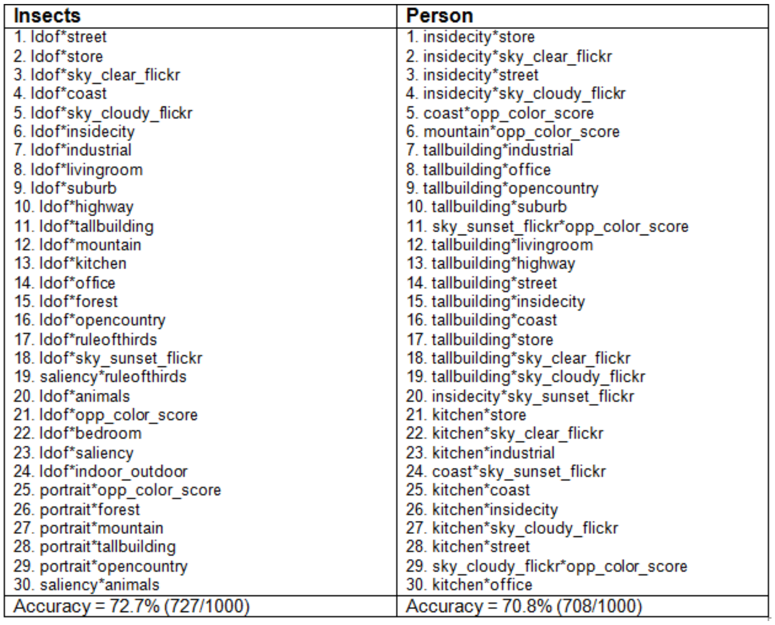
And using those weights we find the attributes that correspond to each element $(t,i)$ to create table Fig. 4.7, Fig. 4.8 of the terms with the highest weights for the DPChallenge dataset, Flickr General Dataset and the query-specific Flickr datasets insects and person.

Fig. 4.9 and Fig. 4.10 show plots of the $\phi_{t,i}(x_k)$ values and the same values weighted using the values calculated before and hence it is a plot of $w_{t,i} * \phi_{t,i}(x_k)$.



| DPChallenge | Flickr General |
|---|---|
| 1. indoor_outdoor*kitchen | 1. sky_sunset_flickr*opp_color_score |
| 2. indoor_outdoor*mountain | 2. sky_cloudy_flickr*opp_color_score |
| 3. indoor_outdoor*forest | 3. insidecity*street |
| 4. indoor_outdoor*suburb | 4. coast*opp_color_score |
| 5. indoor_outdoor*bedroom | 5. insidecity*store |
| 6. sky_sunset_classifier*opp_color_classifier | 6. coast*sky_sunset_flickr |
| 7. indoor_outdoor*tallbuilding | 7. insidecity*sky_clear_flickr |
| 8. tallbuilding*sky_clear_classifier | 8. insidecity*sky_cloudy_flickr |
| 9. tallbuilding*sky_cloudy_classifier | 9. sky_cloudy_flickr*sky_sunset_flickr |
| 10. tallbuilding*industrial | 10. insidecity*sky_sunset_flickr |
| 11. tallbuilding*street | 11. coast*sky_cloudy_flickr |
| 12. tallbuilding*office | 12. insidecity*opp_color_score |
| 13. tallbuilding*livingroom | 13. coast*sky_clear_flickr |
| 14. tallbuilding*suburb | 14. sky_clear_flickr*opp_color_score |
| 15. tallbuilding*store | 15. coast*store |
| 16. mountain*opp_color_classifier | 16. street*store |
| 17. tallbuilding*highway | 17. street*sky_clear_flickr |
| 18. tallbuilding*opencountry | 18. sky_clear_flickr*sky_sunset_flickr |
| 19. tallbuilding*coast | 19. street*sky_cloudy_flickr |
| 20. tallbuilding*insidecity | 20. coast*street |
| 21. opencountry*insidecity | 21. industrial*opp_color_score |
| 22. opencountry*street | 22. coast*insidecity |
| 23. mountain*sky_sunset_classifier | 23. saliency*sky_sunset_flickr |
| 24. highway*opp_color_classifier | 24. saliency*opp_color_score |
| 25. opencountry*coast | 25. sky_clear_flickr*sky_cloudy_flickr |
| 26. opencountry*sky_clear_classifier | 26. people*office |
| 27. opencountry*store | 27. street*sky_sunset_flickr |
| 28. indoor_outdoor*store | 28. store*opp_color_score |
| 29. opencountry*sky_cloudy_classifier | 29. store*sky_sunset_flickr |
| 30. opencountry*office | 30. people*sky_cloudy_flickr |
| Accuracy = 74.25% (1188/1600) | Accuracy = 69.75% (1395/2000) |

Figure 4.7: Attributes that correspond to each element $(t, i)$ of the terms with the highest weights for the DPChallenge, Flickr General Dataset



| Insects | Person |
|---|---|
| 1. ldof*street | 1. insidecity*store |
| 2. ldof*store | 2. insidecity*sky_clear_flickr |
| 3. ldof*sky_clear_flickr | 3. insidecity*street |
| 4. ldof*coast | 4. insidecity*sky_cloudy_flickr |
| 5. ldof*sky_cloudy_flickr | 5. coast*opp_color_score |
| 6. ldof*insidecity | 6. mountain*opp_color_score |
| 7. ldof*industrial | 7. tallbuilding*industrial |
| 8. ldof*livingroom | 8. tallbuilding*office |
| 9. ldof*suburb | 9. tallbuilding*opencountry |
| 10. ldof*highway | 10. tallbuilding*suburb |
| 11. ldof*tallbuilding | 11. sky_sunset_flickr*opp_color_score |
| 12. ldof*mountain | 12. tallbuilding*livingroom |
| 13. ldof*kitchen | 13. tallbuilding*highway |
| 14. ldof*office | 14. tallbuilding*street |
| 15. ldof*forest | 15. tallbuilding*insidecity |
| 16. ldof*opencountry | 16. tallbuilding*coast |
| 17. ldof*ruleofthirds | 17. tallbuilding*store |
| 18. ldof*sky_sunset_flickr | 18. tallbuilding*sky_clear_flickr |
| 19. saliency*ruleofthirds | 19. tallbuilding*sky_cloudy_flickr |
| 20. ldof*animals | 20. insidecity*sky_sunset_flickr |
| 21. ldof*opp_color_score | 21. kitchen*store |
| 22. ldof*bedroom | 22. kitchen*sky_clear_flickr |
| 23. ldof*saliency | 23. kitchen*industrial |
| 24. ldof*indoor_outdoor | 24. coast*sky_sunset_flickr |
| 25. portrait*opp_color_score | 25. kitchen*coast |
| 26. portrait*forest | 26. kitchen*insidecity |
| 27. portrait*mountain | 27. kitchen*sky_cloudy_flickr |
| 28. portrait*tallbuilding | 28. kitchen*street |
| 29. portrait*opencountry | 29. sky_cloudy_flickr*opp_color_score |
| 30. saliency*animals | 30. kitchen*office |
| Accuracy = 72.7% (727/1000) | Accuracy = 70.8% (708/1000) |

Figure 4.8: Attributes that correspond to each element $(t, i)$ of the terms with the highest weights for the category Datasets insects and person.

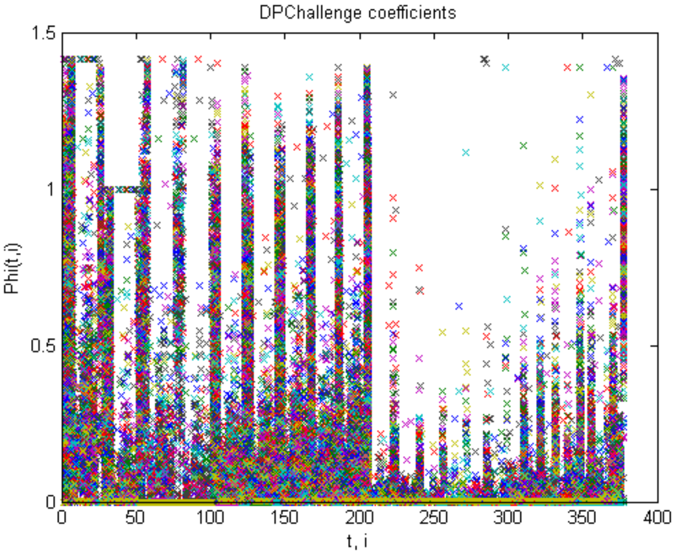


Figure 4.9: Values $\phi_{i,t}$ of the transformation function for every pair $(t, i)$ applied to the support vectors

Now using the weighted values obtained before, we calculate the mean for each $(t, i)$ term and use the highest values to create the tables with the corresponding attributes or pair of attributes that were weighted more. Fig. 4.11 is a plot of the mean values.

Finally Fig. 4.12, Fig. 4.13, Fig. 4.14, Fig. 4.15 contain the attributes corresponding to the highest absolute values of the means from Fig. 4.11.

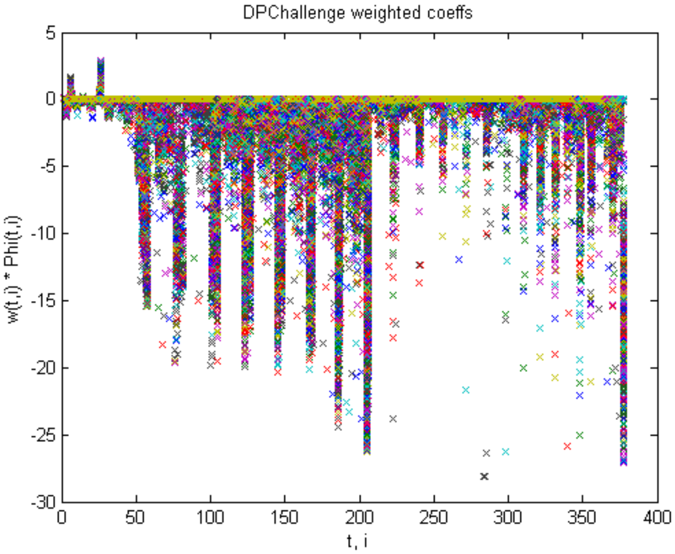


Figure 4.10: Values $\phi_{i,t}$ of the transformation function for every pair $(t, i)$, weighted using the weights $w_{t,i}$ obtained before

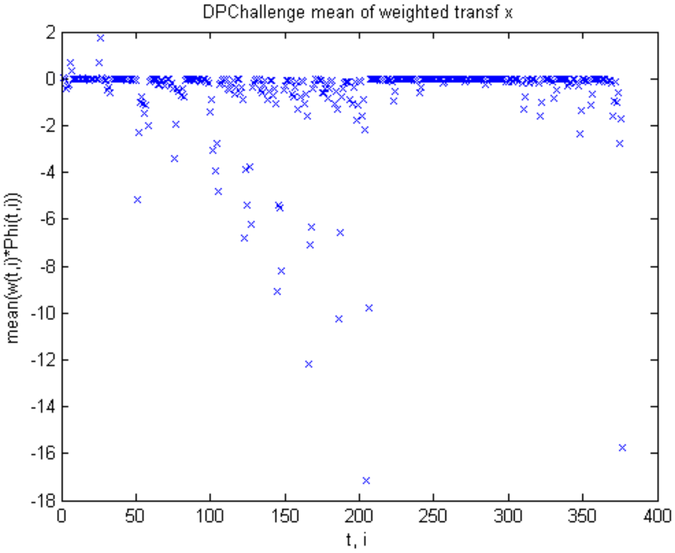
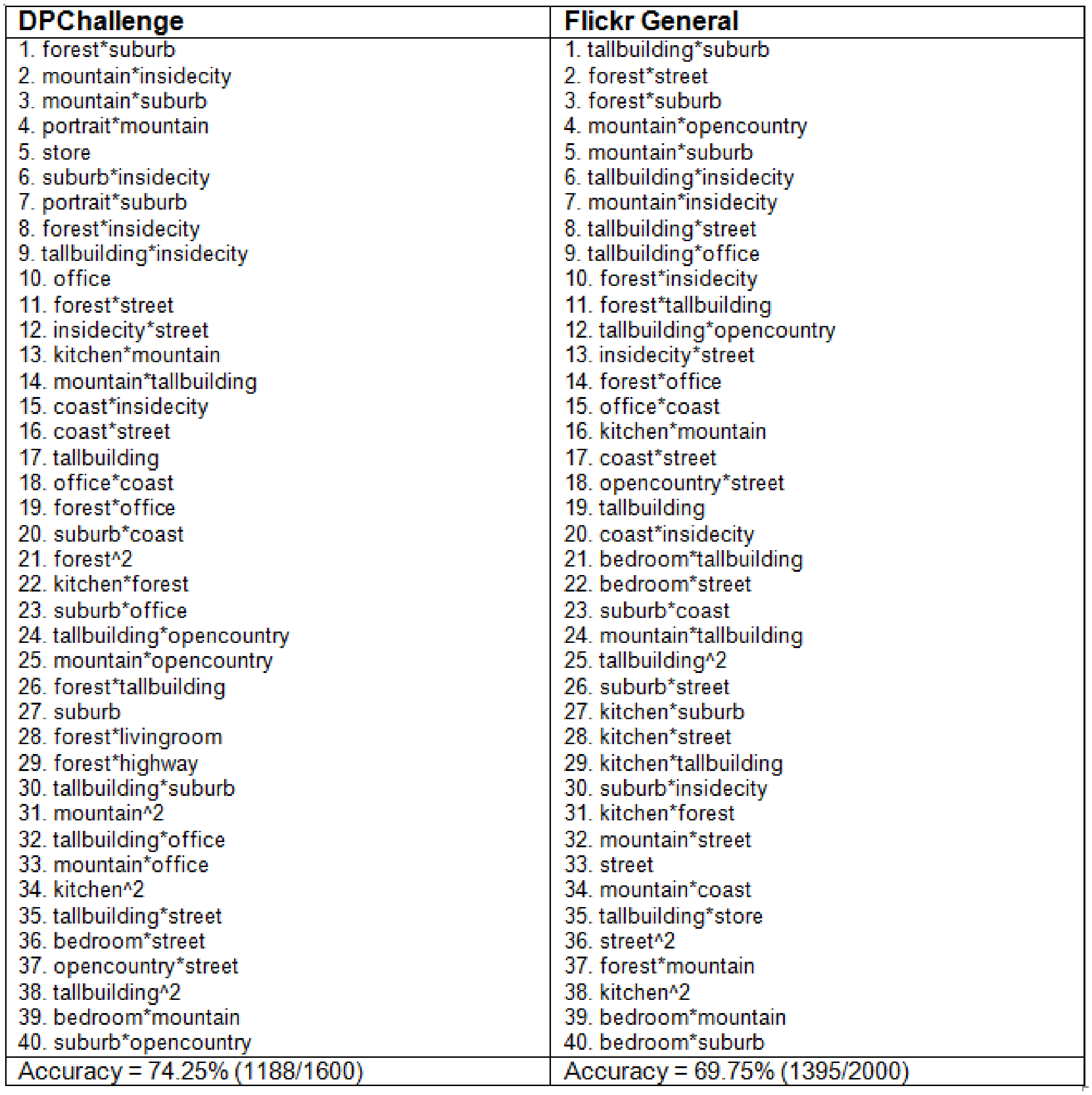
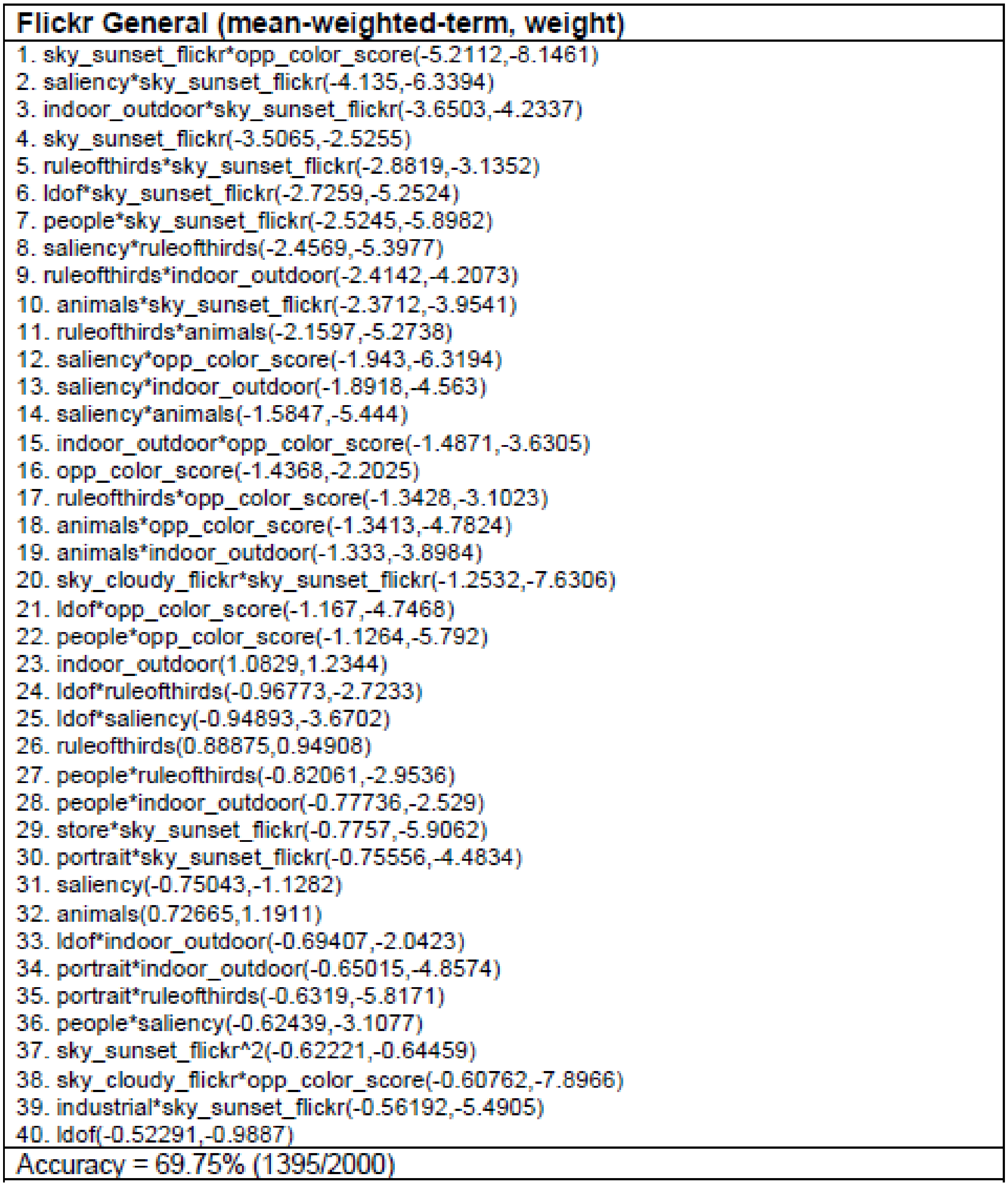


Figure 4.11: Mean values of the weighted transformation coefficients from the values shown in the above figure.



| DPChallenge | Flickr General |
|---|---|
| 1. forest*suburb | 1. tallbuilding*suburb |
| 2. mountain*insidecity | 2. forest*street |
| 3. mountain*suburb | 3. forest*suburb |
| 4. portrait*mountain | 4. mountain*opencountry |
| 5. store | 5. mountain*suburb |
| 6. suburb*insidecity | 6. tallbuilding*insidecity |
| 7. portrait*suburb | 7. mountain*insidecity |
| 8. forest*insidecity | 8. tallbuilding*street |
| 9. tallbuilding*insidecity | 9. tallbuilding*office |
| 10. office | 10. forest*insidecity |
| 11. forest*street | 11. forest*tallbuilding |
| 12. insidecity*street | 12. tallbuilding*opencountry |
| 13. kitchen*mountain | 13. insidecity*street |
| 14. mountain*tallbuilding | 14. forest*office |
| 15. coast*insidecity | 15. office*coast |
| 16. coast*street | 16. kitchen*mountain |
| 17. tallbuilding | 17. coast*street |
| 18. office*coast | 18. opencountry*street |
| 19. forest*office | 19. tallbuilding |
| 20. suburb*coast | 20. coast*insidecity |
| 21. forest^2 | 21. bedroom*tallbuilding |
| 22. kitchen*forest | 22. bedroom*street |
| 23. suburb*office | 23. suburb*coast |
| 24. tallbuilding*opencountry | 24. mountain*tallbuilding |
| 25. mountain*opencountry | 25. tallbuilding^2 |
| 26. forest*tallbuilding | 26. suburb*street |
| 27. suburb | 27. kitchen*suburb |
| 28. forest*livingroom | 28. kitchen*street |
| 29. forest*highway | 29. kitchen*tallbuilding |
| 30. tallbuilding*suburb | 30. suburb*insidecity |
| 31. mountain^2 | 31. kitchen*forest |
| 32. tallbuilding*office | 32. mountain*street |
| 33. mountain*office | 33. street |
| 34. kitchen^2 | 34. mountain*coast |
| 35. tallbuilding*street | 35. tallbuilding*store |
| 36. bedroom*street | 36. street^2 |
| 37. opencountry*street | 37. forest*mountain |
| 38. tallbuilding^2 | 38. kitchen^2 |
| 39. bedroom*mountain | 39. bedroom*mountain |
| 40. suburb*opencountry | 40. bedroom*suburb |
| Accuracy = 74.25% (1188/1600) | Accuracy = 69.75% (1395/2000) |

Figure 4.12: Attributes corresponding to highest absolute values of the means for the DPChallenge Dataset

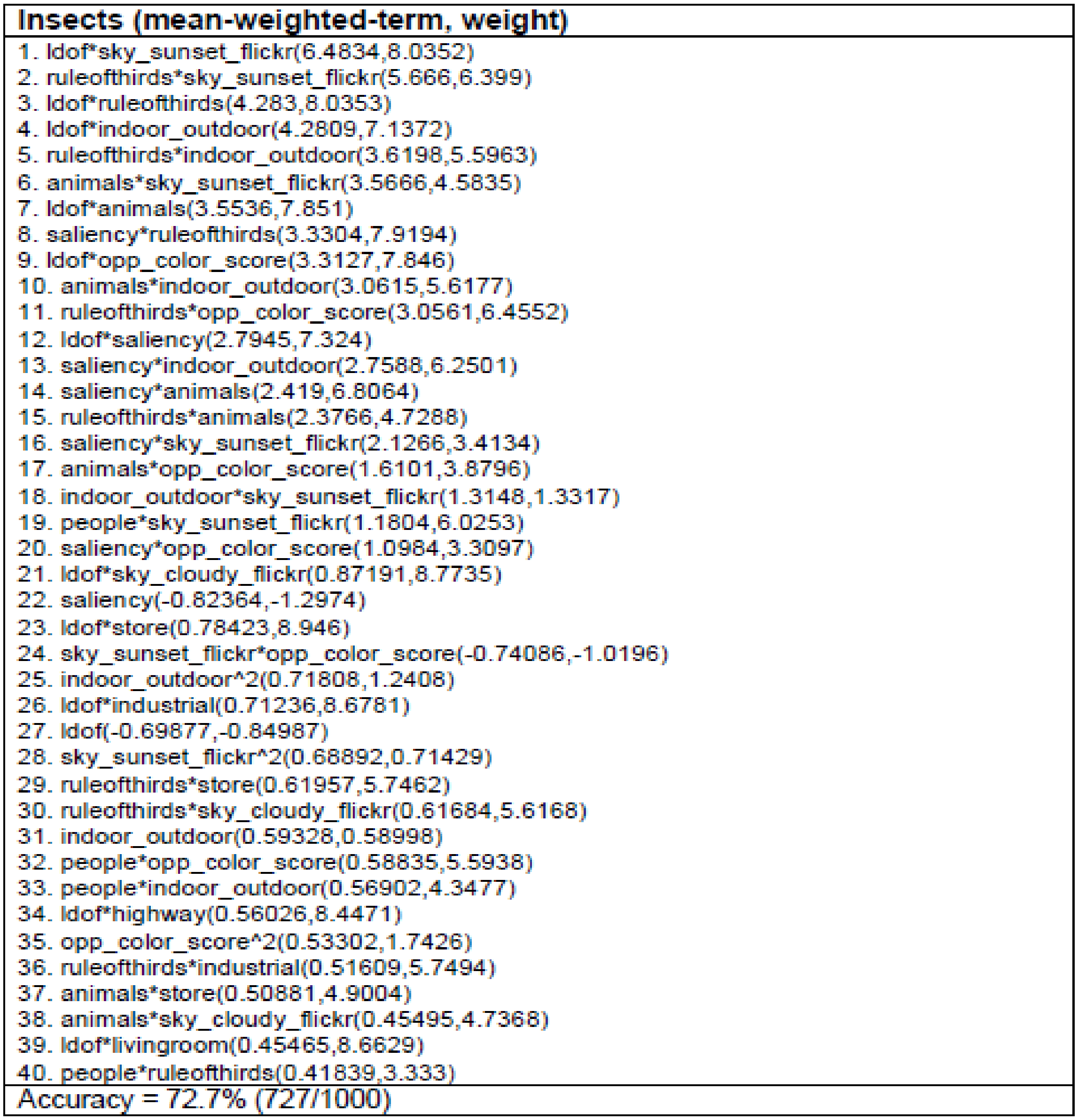


**Flickr General (mean-weighted-term, weight)**

1. sky_sunset_flickr*opp_color_score(-5.2112,-8.1461)
2. saliency*sky_sunset_flickr(-4.135,-6.3394)
3. indoor_outdoor*sky_sunset_flickr(-3.6503,-4.2337)
4. sky_sunset_flickr(-3.5065,-2.5255)
5. ruleofthirds*sky_sunset_flickr(-2.8819,-3.1352)
6. ldof*sky_sunset_flickr(-2.7259,-5.2524)
7. people*sky_sunset_flickr(-2.5245,-5.8982)
8. saliency*ruleofthirds(-2.4569,-5.3977)
9. ruleofthirds*indoor_outdoor(-2.4142,-4.2073)
10. animals*sky_sunset_flickr(-2.3712,-3.9541)
11. ruleofthirds*animals(-2.1597,-5.2738)
12. saliency*opp_color_score(-1.943,-6.3194)
13. saliency*indoor_outdoor(-1.8918,-4.563)
14. saliency*animals(-1.5847,-5.444)
15. indoor_outdoor*opp_color_score(-1.4871,-3.6305)
16. opp_color_score(-1.4368,-2.2025)
17. ruleofthirds*opp_color_score(-1.3428,-3.1023)
18. animals*opp_color_score(-1.3413,-4.7824)
19. animals*indoor_outdoor(-1.333,-3.8984)
20. sky_cloudy_flickr*sky_sunset_flickr(-1.2532,-7.6306)
21. ldof*opp_color_score(-1.167,-4.7468)
22. people*opp_color_score(-1.1264,-5.792)
23. indoor_outdoor(1.0829,1.2344)
24. ldof*ruleofthirds(-0.96773,-2.7233)
25. ldof*saliency(-0.94893,-3.6702)
26. ruleofthirds(0.88875,0.94908)
27. people*ruleofthirds(-0.82061,-2.9536)
28. people*indoor_outdoor(-0.77736,-2.529)
29. store*sky_sunset_flickr(-0.7757,-5.9062)
30. portrait*sky_sunset_flickr(-0.75556,-4.4834)
31. saliency(-0.75043,-1.1282)
32. animals(0.72665,1.1911)
33. ldof*indoor_outdoor(-0.69407,-2.0423)
34. portrait*indoor_outdoor(-0.65015,-4.8574)
35. portrait*ruleofthirds(-0.6319,-5.8171)
36. people*saliency(-0.62439,-3.1077)
37. sky_sunset_flickr^2(-0.62221,-0.64459)
38. sky_cloudy_flickr*opp_color_score(-0.60762,-7.8966)
39. industrial*sky_sunset_flickr(-0.56192,-5.4905)
40. ldof(-0.52291,-0.9887)

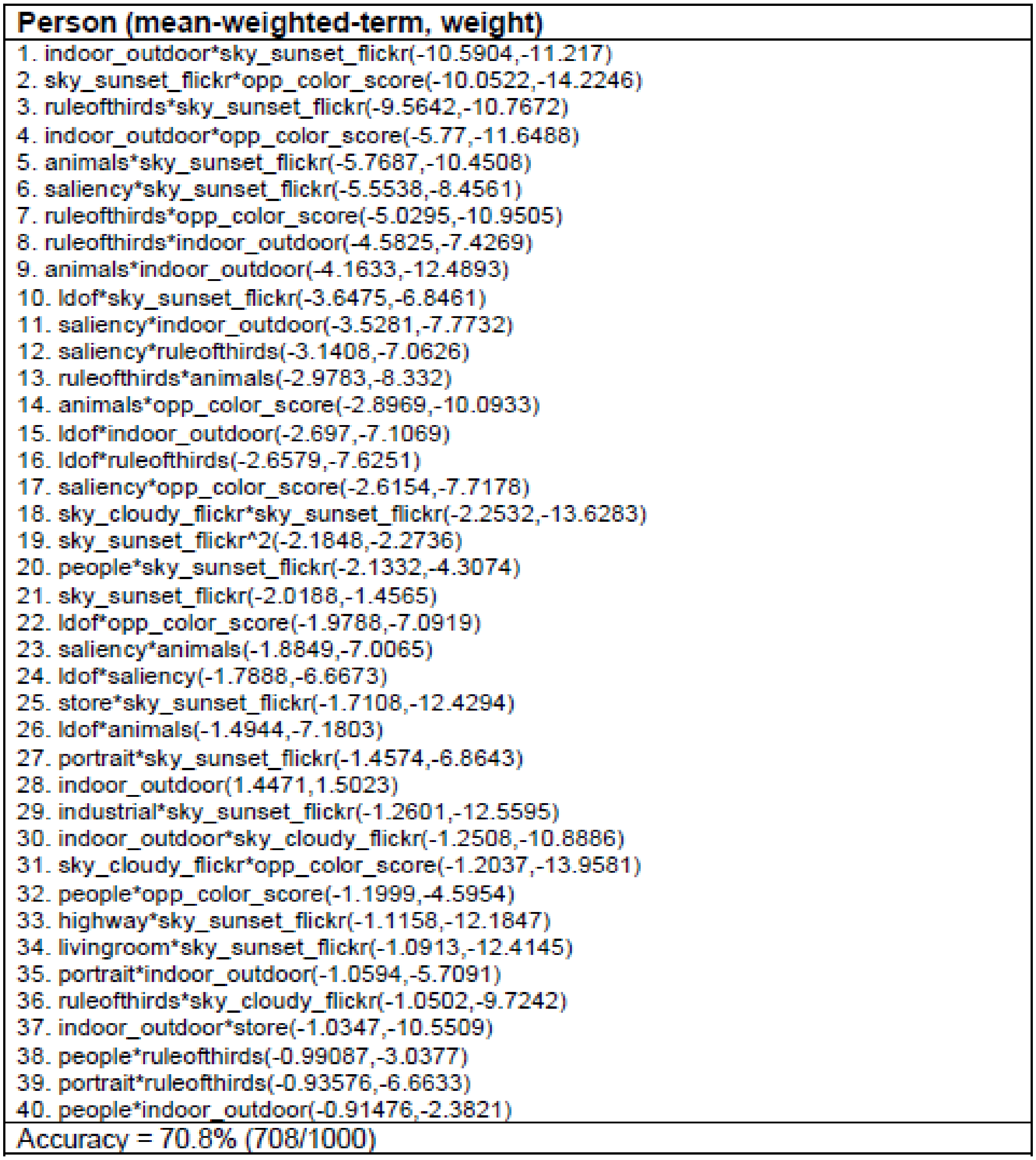
Accuracy = 69.75% (1395/2000)

Figure 4.13: Attributes corresponding to highest absolute values of the means for the Flickr General Dataset



| Insects (mean-weighted-term, weight) |
|---|
| 1. ldof*sky_sunset_flickr(6.4834,8.0352) |
| 2. ruleofthirds*sky_sunset_flickr(5.666,6.399) |
| 3. ldof*ruleofthirds(4.283,8.0353) |
| 4. ldof*indoor_outdoor(4.2809,7.1372) |
| 5. ruleofthirds*indoor_outdoor(3.6198,5.5963) |
| 6. animals*sky_sunset_flickr(3.5666,4.5835) |
| 7. ldof*animals(3.5536,7.851) |
| 8. saliency*ruleofthirds(3.3304,7.9194) |
| 9. ldof*opp_color_score(3.3127,7.846) |
| 10. animals*indoor_outdoor(3.0615,5.6177) |
| 11. ruleofthirds*opp_color_score(3.0561,6.4552) |
| 12. ldof*saliency(2.7945,7.324) |
| 13. saliency*indoor_outdoor(2.7588,6.2501) |
| 14. saliency*animals(2.419,6.8064) |
| 15. ruleofthirds*animals(2.3766,4.7288) |
| 16. saliency*sky_sunset_flickr(2.1266,3.4134) |
| 17. animals*opp_color_score(1.6101,3.8796) |
| 18. indoor_outdoor*sky_sunset_flickr(1.3148,1.3317) |
| 19. people*sky_sunset_flickr(1.1804,6.0253) |
| 20. saliency*opp_color_score(1.0984,3.3097) |
| 21. ldof*sky_cloudy_flickr(0.87191,8.7735) |
| 22. saliency(-0.82364,-1.2974) |
| 23. ldof*store(0.78423,8.946) |
| 24. sky_sunset_flickr*opp_color_score(-0.74086,-1.0196) |
| 25. indoor_outdoor^2(0.71808,1.2408) |
| 26. ldof*industrial(0.71236,8.6781) |
| 27. ldof(-0.69877,-0.84987) |
| 28. sky_sunset_flickr^2(0.68892,0.71429) |
| 29. ruleofthirds*store(0.61957,5.7462) |
| 30. ruleofthirds*sky_cloudy_flickr(0.61684,5.6168) |
| 31. indoor_outdoor(0.59328,0.58998) |
| 32. people*opp_color_score(0.58835,5.5938) |
| 33. people*indoor_outdoor(0.56902,4.3477) |
| 34. ldof*highway(0.56026,8.4471) |
| 35. opp_color_score^2(0.53302,1.7426) |
| 36. ruleofthirds*industrial(0.51609,5.7494) |
| 37. animals*store(0.50881,4.9004) |
| 38. animals*sky_cloudy_flickr(0.45495,4.7368) |
| 39. ldof*livingroom(0.45465,8.6629) |
| 40. people*ruleofthirds(0.41839,3.333) |
| Accuracy = 72.7% (727/1000) |

Figure 4.14: Attributes corresponding to highest absolute values of the means for the category Datasets ‘insects’.



| Person (mean-weighted-term, weight) |
|---|
| 1. indoor_outdoor*sky_sunset_flickr(-10.5904,-11.217) |
| 2. sky_sunset_flickr*opp_color_score(-10.0522,-14.2246) |
| 3. ruleofthirds*sky_sunset_flickr(-9.5642,-10.7672) |
| 4. indoor_outdoor*opp_color_score(-5.77,-11.6488) |
| 5. animals*sky_sunset_flickr(-5.7687,-10.4508) |
| 6. saliency*sky_sunset_flickr(-5.5538,-8.4561) |
| 7. ruleofthirds*opp_color_score(-5.0295,-10.9505) |
| 8. ruleofthirds*indoor_outdoor(-4.5825,-7.4269) |
| 9. animals*indoor_outdoor(-4.1633,-12.4893) |
| 10. ldof*sky_sunset_flickr(-3.6475,-6.8461) |
| 11. saliency*indoor_outdoor(-3.5281,-7.7732) |
| 12. saliency*ruleofthirds(-3.1408,-7.0626) |
| 13. ruleofthirds*animals(-2.9783,-8.332) |
| 14. animals*opp_color_score(-2.8969,-10.0933) |
| 15. ldof*indoor_outdoor(-2.697,-7.1069) |
| 16. ldof*ruleofthirds(-2.6579,-7.6251) |
| 17. saliency*opp_color_score(-2.6154,-7.7178) |
| 18. sky_cloudy_flickr*sky_sunset_flickr(-2.2532,-13.6283) |
| 19. sky_sunset_flickr^2(-2.1848,-2.2736) |
| 20. people*sky_sunset_flickr(-2.1332,-4.3074) |
| 21. sky_sunset_flickr(-2.0188,-1.4565) |
| 22. ldof*opp_color_score(-1.9788,-7.0919) |
| 23. saliency*animals(-1.8849,-7.0065) |
| 24. ldof*saliency(-1.7888,-6.6673) |
| 25. store*sky_sunset_flickr(-1.7108,-12.4294) |
| 26. ldof*animals(-1.4944,-7.1803) |
| 27. portrait*sky_sunset_flickr(-1.4574,-6.8643) |
| 28. indoor_outdoor(1.4471,1.5023) |
| 29. industrial*sky_sunset_flickr(-1.2601,-12.5595) |
| 30. indoor_outdoor*sky_cloudy_flickr(-1.2508,-10.8886) |
| 31. sky_cloudy_flickr*opp_color_score(-1.2037,-13.9581) |
| 32. people*opp_color_score(-1.1999,-4.5954) |
| 33. highway*sky_sunset_flickr(-1.1158,-12.1847) |
| 34. livingroom*sky_sunset_flickr(-1.0913,-12.4145) |
| 35. portrait*indoor_outdoor(-1.0594,-5.7091) |
| 36. ruleofthirds*sky_cloudy_flickr(-1.0502,-9.7242) |
| 37. indoor_outdoor*store(-1.0347,-10.5509) |
| 38. people*ruleofthirds(-0.99087,-3.0377) |
| 39. portrait*ruleofthirds(-0.93576,-6.6633) |
| 40. people*indoor_outdoor(-0.91476,-2.3821) |
| Accuracy = 70.8% (708/1000) |

Figure 4.15: Attributes corresponding to highest absolute values of the means for the category Datasets ‘person’.

# Chapter 5

# Conclusion

We provide a set of high level describable attributes that predict various kinds of useful information about images. These attributes can be used for constraining search results, collection organization, or browsing. We also demonstrate that our describable attributes can be used to produce powerful classifiers for estimating aesthetic quality, general interestingness, and query specific interestingness. In the future, we plan to expand our set of attributes to extract other describable image features, and to apply these attributes to related tasks such as image emotion estimation. We also plan to more thoroughly explore ideas of query specific interestingness, including methods for query specific attribute selection, and methods for interestingness transfer.

The quality estimation method in this system predicts some of the possible image cues that a human might use to evaluate an image and then uses them to classify the image as ‘positive’ or ‘negative’ by measuring how much it deviates from ideal sample of either categories. This could be instrumental in post-processing tools (like Photoshop) to not only improve the quality of an image but also to help a photographer take better photos. The following applications could potentially be very useful:

1. Classification of aesthetic quality is done using a measure of how much a test image deviates from the ideal image. The ideal image(s) are usually generated during the training process of the classifier. The measure of deviation of the test image from the ideal image could be used to suggest potential changes to the user/photographer to help improve the aesthetic quality of his photograph. Eg: One of the attributes used, Rule of Thirds is directly measured by how much it deviates from the thirds lines in the image. The ideal position could be shown as an overlay to the user, as a suggested post-processing step.

2. Our algorithm also takes into account composition of the photograph while doing analysis of aesthetics in the image. Compositional features include Saliency detection, Low Depth of Field.

   - Saliency improvements could be suggested as possible crops of the photograph so as to make the primary object in the photograph most salient.
   - Depth of field measurements could help suggest to the user, modifications in the focus level of various parts of the image. This would make the primary object in the image highly focused while defocussing the surroundings, which is the aim of photographers taking advantage of the low depth-of-field feature or macro mode.

3. Digital media designers strive to achieve perfection when creating digital content (both images and video). During this process, our aesthetics inference algorithm could be used to suggest possible color pairs to enhance the aesthetic value of the image/video being created. Our algorithm results show us a high occurrence of blue+yellow, or black+red combinations. The high number of beach photographs and sunset photographs liked by users on the web validate this observation. The opposing colors attribute computes color histograms of

the test image. The shape of these histograms could be compared to ideal color composition histograms generated during the training process. Our study of research in Psychology literature of color-pair and color-triplet combinations best perceived by the human mind also can be used to enhance the digital media creation process.

4. Digital media creators working in the area of developing animation could use cues from our Content attributes to get an intuition of possible object combinations that they could add to increase the aesthetic quality of a scene. For instance, in outdoor scenes like a beach or a park, we observed that the presence of a pet (like a dog) frequently increases the visual acceptance of the photograph. The sky-illumination attributes could be used to render life-like sky color and texture for various sky types in outdoor scene videos/photographs.

# Bibliography

[1] O. Axelsson. Towards a psychology of photography: dimensions underlying aesthetic appeal of photographs. In *Perceptual and Motor Skills*, 2007.

[2] T. L. Berg and D. A. Forsyth. Animals on the web. In *CVPR*, 2006.

[3] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Studying aesthetics in photographic images using a computational approach. In *ECCV*, 2006.

[4] R. Datta, J. Li, and J. Z. Wang. Algorithmic inferencing of aesthetics and emotion in natural images: An exposition. In *ICIP*, 2008.

[5] A. Farhadi, I. Endres, D. Hoiem, and D. A. Forsyth. Describing objects by their attributes. In *CVPR*, 2009.

[6] E. Fedorovskaya, C. Neustaedter, and W. Hao. Image harmony for consumer images. In *ICIP*, 2008.

[7] L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *CVPR*, 2005.

[8] V. Ferrari and A. Zisserman. Learning visual attributes. *NIPS*, 2007.

[9] J. Gardner, C. Nothelfer, and S. Palmer. Exploring aesthetic principles of spatial composition through stock photography. In *Vision Sciences Society*, 2008.

[10] D. Hoiem, A. A. Efros, and M. Hebert. Geometric context from a single image. In *ICCV*, pages 654–661, 2005.

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

[12] Y. Ke, X. Tang, and F. Jing. The design of high-level features for photo quality assessment. In *CVPR*, 2006.

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

[14] N. Kumar, P. Belhumeur, and S. K. Nayar. FaceTracer: A search engine for large collections of images with faces. In *ECCV*, 2008.

[15] N. Kumar, A. Berg, P. Belhumeur, and S. K. Nayar. Attribute and simile classifiers for face verification. In *ICCV*, 2009.

[16] C. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 2009.

[17] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching. In *CVPR*, June 2006.

[18] A. Loui, M.Wood, A. Scalise, and J. Birkelund. Multidimensional image value assessment and rating for automated albuming and retrieval. In *ICIP*, 2008.

[19] D. G. Lowe. Distinctive image features from scale invariant keypoints. *IJCV*, 2004.

[20] T. Lui, J. Sun, N.-K. Zheng, X. Tang, and H.-Y. Shum. Learning to detect a salient object. In *CVPR*, 2007.

[21] C. Nothelfer, K. B. Schloss, and S. E. Palmer. The role of spatial composition in preference for color pairs. In *Vision Sciences Society*, 2009.

[22] A. Oliva and A. Torralba. Modeling the shape of the scene: a holistic representation of the spatial envelope. *IJCV*, 2001.

[23] S. Palmer, E. Rosch, and P. Chase. Canonical perspective and the perception of objects. In *Attention and Performance*, 1981.

[24] R. M. Poggesi, K. B. Schloss, and S. E. Palmer. Preference for three-color combinations in varying proportions. In *Vision Sciences Society*, 2009.

[25] F. Porikli. Integral histogram: A fast way to extract histograms in cartesian spaces. In *CVPR*, 2005.

[26] A. Savakis, S. Etz, and A. Loui. Evaluation of image appeal in consumer photography. In *SPIE*, 2000.

[27] K. B. Schloss and S. E. Palmer. Color preferences. In *Vision Sciences Society*, 2007.

[28] X. Sun, H. Yao, R. Ji, and S. Liu. Photo assessment based on computational visual attention model. In *ACM MM*, 2009.

[29] H. Tong, M. Li, H. Zhang, J. He, and C. Zhang. Classification of digital photos taken by photographers or home users. In *PCM*, 2004.

[30] H. Tong, M. Li, H. Zhang, C. Zhang, J. He, and W.-Y. Ma. Learning no-reference quality metric by examples. In *ICMM*, 2005.

[31] P. Viola and M. Jones. Robust real-time object detection. In *IJCV*, 2001.