CRANFIELD UNIVERSITY

JOÃO MARQUES PINTO DE FARIA

WHAT MAKES A GOOD PICTURE?

SCHOOL OF ENGINEERING
Msc Computational Software Techniques In Engineering
Digital Signal And Image Processing

MSc Thesis
Academic Year: 2011 - 2012

Supervisor: Stefan Rüger
August 2012

CRANFIELD UNIVERSITY

SCHOOL OF ENGINEERING
Msc Computational Software Techniques In Engineering
Digital Signal And Image Processing

MSc Thesis

Academic Year 2011 - 2012

JOÃO MARQUES PINTO DE FARIA

What makes a good picture?

Supervisor: Stefan Rüger
August 2012

This thesis is submitted in partial fulfilment of the requirements for the degree of Masters of Science

ABSTRACT

Aesthetic principles are inerent to the human being. However, it is difficult to imagine a machine which can simulate human intuition and opinions. In this project, I address the problem of automatic distinction between high quality professional photographs and low quality amateur snapshots. Although quality may be interpreted in different ways, there are some data patterns which can be explored in order to achieve a correct classification. My system uses a set of features to teach a machine learning algorithm how to differentiate between the two considered classes of photographs. Grid search technique is used to explore possible classifiers' parameters combinations and forward/backward selection is performed for feature selection. The system is tested on a large and widely used dataset, so it is possible to compare the results with other works. The reported precision and recall values are much higher than the ones previously observed.

Keywords:

Aesthetics, classification, digital photography, picture assessment, saliency

ACKNOWLEDGEMENTS

I would like to gratefully acknowledge my supervisor, Professor Stefan Rüger , for all the help provided during my thesis period. His guidance allowed me to keep motivated and focused on the crucial aspects of my work.

My special thanks to all my friends from all around the world. In Cranfield, I have met great people who I will not forget. They made my last year much more enjoyable than I could even expect. I left many good fellows in Porto. I truly appreciate all the good times we had together. I would like to thank Hugo and Luís, my all-time friends. I could not forget Grace, who is a great inspiration for me.

I owe my deepest gratitude to all my family, my unconditional supporters. I specially thank my Mom, Dad, and Sister, for always making me a priority on their lives.

At last but not less important, I am grateful to all the teachers who were part of my education. I specially thank them for their patience, which had a crucial contribution on the person I am. Thank you for making me believe I can reach the infinite and beyond.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF EQUATIONS

LIST OF ABBREVIATIONS

WWW	World Wide Web
SVM	Support Vector Machine
RBF	Radial Basis Function
CV	Cross Validation
HSV	Hue, saturation, and value
RGB	Red, green, and blue
RMS	Root Mean Square
EXIF	Exchange Image File
FFT	Fast Fourier Transform
GBVS	Graph-based Visual Saliency
RBCS	Region-based Contrast Saliency

1 INTRODUCTION

The introduction of consumer digital cameras in 1988 [1] had a huge impact on the way people face photography. Although having several limitations, the first models captured people's attention by allowing users to store images in the internal memory and seeing them without the necessity of using any other physical device. Moreover, personal computers were having a great success and the Internet was experiencing an unbelievable growth. The WWW was the engine behind the evolution of digital contents. Millions of people could now share contents in a matter of seconds from their homes. Industry started developing more and more complex digital cameras. They quickly became the standard on photography.

Nowadays, digital images are so proliferated which became important to automatically distinguish pictures by defined characteristics. Search engines, for example, crawl the web collecting results from a given query and presenting the information according to certain rules, like the dominant colour, often decided by the user. However, there are some concepts which are not trivial to evaluate but could reveal relevant on the user perspective.

In this project I address the problem of distinguishing images by their aesthetical appeal. Automatically assessing the quality of photos has become important on several applications. As mentioned before, search engines are always trying to improve the presented results. A high quality picture is usually preferable to a low quality one. Thus, the more efficient the detection of aesthetic value is, the more meaningful results will be reported to the user. This kind of algorithms can also be used to automatically select pictures from a personal data base or, in a more extreme way, allow a digital camera to pick the best photos from a photographic session. Another possible use of the automatic picture quality detection is on computer-aided photo editing [2] [3], where the software changes the characteristics of the images in order to improve their perceived appeal. In some recent research [4], robots with the ability of photographing scenes using some aesthetical principles were developed. There

are several areas which can take advantage of the developing of efficient solutions to the described problem.

Since aesthetics is not an absolute concept, it cannot be dealt in the same way as writing a mathematical formula. In fact, the introduction of subjectivity increases the complexity of this problem. It is not possible to clearly define an image as aesthetically appealing as it is impossible to quantify a person's beauty. Different people have different opinions. However, humans tend to share the same basic principles, making room for a possible solution to the problem.

In this project I will try to develop a set of features which correlate with the way humans perceive the quality of a photo. Although using elements from existing works, I will try to introduce some novel measurements in order to expand the knowledge acquired by the research taken by the scientific community. The values of the extracted features will be then applied to machine learning algorithms, so the model can learn data patterns and classify new examples without previous knowledge of their classes. Similarly to the features part, I propose a mix of previously developed solutions with novel techniques. The researchers usually choose the same machine learning algorithms. This project will incorporate some different ones, so it is possible to test the effectiveness of novel techniques. A comparison of the different methods will be performed in order to select the techniques which suit the best to the facing problem. It will be tested the possibility of combining some of them expecting that the "whole is more than the sum of its parts" as Aristotle once suggested. Furthermore, I will try to create a robust method of results comparison, since there are no standard practices used by the scientific community.

2 LITERATURE REVIEW

In this chapter I am going to describe some of the work developed so far by the scientific community in the domain of image aesthetics evaluation.

The problem of classifying images based on aesthetic properties is a relatively recent research subject. It was first attempted by Tong et al. [5], in 2004, and was inspired in the previous works of Serrano et al. [6], who addressed the challenge of classifying indoor/outdoor scenes, Oliveira et al. [7], who developed a system able to distinguish photographs from graphics, and Vailaya et al. [8], who created an algorithm to differentiate city and landscape pictures.

In [5], the authors mentioned the reasons why they were facing a much more complex problem: 1) although it is trivial for someone to tell the difference between professional photos and snapshots, it is not completely known what kind of high-level factors make them different; 2) even if we know those high-level factors, it is difficult to express them as appropriate low-level features. To address these issues they took a black-box approach, by developing and combining a huge 846-dimensional low-level feature set with different types of learning algorithms for classification.

The most important used features were connected with blurring, by applying the wavelet transform technique described in [9], contrast, colourfulness and saliency, by calculating the first three order moments of a fairly simple saliency map of the image [10].

In terms of classification, the authors of this work decided to use different forms of boosting (AdaBoost and Real-AdaBoost), SVM and Bayesian classifier. In order to evaluate the performance of the algorithm, a data set of 29540 images was assembled. The professional photo examples were formed by images from COREL and Microsoft Office Online, while the snapshots were collected from private collections of the authors' working staff.

The best results were achieved with the Bayesian classifier, with a testing error of 4.9%. However, the used data set is fairly homogeneous and therefore it is easy to separate the two classes. This work was the first step towards a series

of different approaches. Although showing a possible direction to take, it offers little insight into how to design and choose better features. The big number of collected features were simply exhaustively combined and tested in order to achieve the maximum possible score. The results cannot be compared with this work, since it is used a completely different dataset.

(a) (b)

Figure 2-1. Simplicity property. In (a) the subject is much clearer than in (b).

In 2006, Ke et al. [11] tried to address the described limitations by implementing a principled approach. Instead of implementing a big number of different low-level features, they decided to first determine the perceptual factors that distinguish between professional photos and snapshots. They studied the criteria people use to judge a photo by interviewing professional and amateur photographers and non-photographers. In addition, they researched books [12] [13] to learn what techniques photographers often use to improve the quality of their work.

A top-down approach to create high-level features was then presented. The authors found three distinguishing factors: simplicity, realism, and basic photographic techniques. Pictures are simple when it is obvious what one should be looking at. Figure 2-1 shows the importance of such characteristic. In (a), the subject is very clear while in (b) it is almost indistinguishable from the background. The other mentioned characteristic, realism, was considered equally important. Snapshots usually capture what the eye sees. On the other hand, professional photographers use a wide range of conditions, like lighting,

to make photos look surreal. The last identified factors were two basic photographic techniques: blurriness and contrast. In the end, just six features were used: spatial distribution of edges, colour distribution, count of hues, blur calculation, width of 98% of the mass in the brightness histogram and the average brightness.

The authors' primary aim was not on the learning method, so they simply used the one which achieved the best results in the Tong et al. [5] paper, the naïve Bayes classifier. For the first time, it was given a fairly big importance to the creation of a publicly available data set. By crawling a photo contest website, DPChallenge.com, it was possible to select the top and bottom rated pictures labelling them as professional photos and snapshots, respectively. From an initial set of 60000 photos, 12000 were chosen, 6000 for each category. For each set, half of the photos were used for training and the other half for testing. This was a huge contribution for future papers, since authors could finally have a fairly good comparison of results with previous work.

The developed algorithm achieved an error rate of 27.8%. In order to have a basis of comparison the Tong et al. [5] algorithm was tested on the same data set, getting almost the same result, 27.7%, with much more used features. Thus, The Ke et al. [11] paper, showed the importance of a thorough feature selection. This reported result assumes a great importance and will be used as a benchmark for my work, since the same dataset is employed.

In a similar way, Datta et al. [14] computed a set of 56 features based on rules of thumb in photography, common intuition and observed trends in ratings. They introduced some novel measurements, like the aspect ratio of the picture and the rule of thirds, where the main elements of the image are close to the points which divide the image in three parts, both vertically and horizontally.

The initial number of features was reduced to 15, by combining filter-based and wrapper-based methods. This approach introduces an interesting way of reducing the problem complexity and getting the most effective feature set.

Figure 2-2. The rule of thirds.

In this work, a new data set was also created. From the Photo.net website, another online photo sharing community, 20278 pictures were downloaded and splitted into two categories depending on the classification given by the users. Using a SVM (Support Vector Machine) with a RBF (Radial Basis Function) kernel and 5-CV, it was achieved an accuracy of 70.12%. This result is merely illustrative, since the used dataset is too different from the one used in my work.

All the algorithms described so far compute features from the whole image. As Ke et al. [11] mentioned, one of the key points of a good picture is a clear distinction between the subject and the background. Thus, it would make sense identifying and treating these two areas differently.

Luo et al. [15] introduced a new approach: by first differentiating the subject from the background, they were able to calculate more effective features taking into account different characteristics for both areas. An example of the subject extraction result is shown in Figure 2-3. From the original image (b), a binary map was created (a), where the white pixels correspond to the subject area. Luo et al. used a technique exclusively based on blur, proposed in [16], to perform this extraction. The idea was blurring the image with different kernels and determining which areas had large variations on their derivatives. For that, a probabilistic model was created, indicating which blurring kernel had a higher likelihood of being the present one at each pixel location. After ending up with an image similar to the one shown in Figure 2-3 (b), a box with 90% of the energy would define the final subject area.

(a) (b)

Figure 2-3. Subject mask extraction.

The subject/foreground differentiation was evaluated based on four identified characteristics for photo quality assessment: focus controlling, lighting, colour, and composition. Professional photographers usually keep the subject in focus and make the background out of focus. The first extracted feature used the blur feature from Ke et al. [11], independently applied to each of the detected regions. The ratio between both areas reported the blurring relation. Professional photos usually have their subject enhanced by lighting. By calculating the ratio between subject and foreground average brightness values, Luo et al. introduced a way to compare the lighting component on both areas. Other technique used by photographers to reduce the distraction from the subject is keeping the background simple. Thus, the number of relevant colours from the background was calculated. In order to determine the colour harmony, Luo et al. computed a histogram which reports the average occurrence of each pair of colours in the training dataset. In terms of composition, the authors decided to implement the already described rule of thirds, by using the subject position.

The performance of the system was measured by using three different learning algorithms: Bayesian classifier, SVM and Gentle AdaBoost. The system was tested on the same dataset used by Ke et al. The results were higher than any other experiment tried so far. Gentle AdaBoost was reported to achieve an error rate of 6%. However, no details on the selection of parameters were given. This is other important work which will be compared with my solution, since the same

database is used. This was the first work considering the subject/background separation. After that, a large number of researchers decided to use the same approach.

The subject extraction performed by Luo et al. just considered one characteristic which influences human attention, the blurring. However, there are many other aspects which play an important role on this process. The calculation of salient regions within an image is a whole independent research area.

From the end of the 90's, there has been the effort to develop models which are able to accurately predict the contribution of each pixel of an image to the attention of the observers. The work of Itti et al [17] was the first one to achieve a massive recognition from the scientific community. In its conception is the idea of combining information relative to colours, intensity and local orientation. The collected features provide bottom-up input to the saliency map, modelled as a dynamical neural network.

Harel et al. [18] took advantage of the computational power, topographical structure, and parallel nature of graph algorithms to achieve natural and efficient saliency computations. The graphs are interpreted as Markov chains. Besides the improved speed, the Harel et al. algorithm was able to get better saliency maps in terms of resolution. An example of the resulting saliency map of this algorithm is presented in Figure 4-1 (b). Although reporting the most salient locations of an image, the algorithms analysed so far have the disadvantage of considering each pixel independently. However, images are usually perceived as a set of regions.

Cheng et al. [19] addressed the identified problem and presented a new approach based on colour features and image segmentation. The image is first segmented using the algorithm purposed by Felzenszwalb et al. [20]. Each region takes the average value of the pixels that compose it. After converting the values to the CIE L*a*b* colour space, the saliency map is calculated using the linear colour distances between regions. A region which has a very distinctive colour has certainly a higher saliency value than another one which is

similar to most existing patches. The spatial distance between regions is also taken into consideration, by using the Euclidian distance between centroids. If two regions are far from each other, they will have a low mutual contribution to their saliency levels. An example of a saliency map created by the region based contrast algorithm proposed by Cheng et al. is presented in Figure 4-1 (c).

Several approaches tried to take advantage of the parallel studies on the saliency field. Wong et al. [21] used Itti's visual saliency model [17], choosing the three most salient locations as seeds to the subject mask determination. As mentioned before, Itti's algorithm does not provide a region-based saliency. In order to extract all the possible subject areas, the authors performed image segmentation. All the segmented regions that contained the computed seeds were labelled as being part of the subject.

The authors of this work decided to compute 44 features divided into three different types: global image features, features of salient regions, and features that depict the subject-background relationship.

Concerning to the global features, it was used a simplified version of the algorithm proposed in [11], three sharpness measurements and the calculation of the middle 98% mass of the luminance histogram, also based on a feature presented in [11]. The average pixel intensity [14] was calculated as measure of exposure. Some photographic rules were also included, particularly, the rule of thirds and the low depth of field, a measurement of the distances in which an object appears acceptably sharp. Both features were based on the works of Datta et al. in [14].

For the subject regions, the same features were computed, excepting, obviously, the rule of thirds. In addition, there were included other measurements. In terms of texture, it was calculated the sum of the average wavelet coefficients over all levels, applied to each channel on the HSV colour space. In order to test the "fill the frame" principle, which suggests that the subject should occupy a large portion of the image, it was computed the dimension of the salient regions. The position, distribution, and total number of salient locations were also included in this work. A professional photo usually

has a strong focus on the subject, which means the number of salient regions is low and their locations are densely distributed. Such distribution is calculated using the standard deviation of all the salient locations. The mean and standard deviation of the saliency map were also computed, so the experiment includes the salience strength information.

The last set of features, the subject-background comparison, was mostly based on the principle which states that the subject should be emphasized in comparison with the background. Wong et al. applied the same methods used for the global image features to compute exposure, saturation, hue, blurriness, and texture details, for both the salient and background regions. The differences of the respective features of the subject and the background were computed using the squared difference. These measurements are a way of determining in which way the subject is similar to the background. The edge spatial distribution described in [11] was also computed for both the subject and the background areas. The comparison was again calculated by means of the squared difference.

To allow a direct comparison with the results of Data et al. [14], the authors tried to download the same set of photos, from Photo.net. However, some of the users have removed their pictures. Thus, just a subset of the original dataset was retrieved. The features were used to train one-dimensional SVM. Using 5-CV, the algorithm achieved an accuracy of 78.8%, significantly higher than 70.12%, the score obtained by Datta et al. By applying Datta's algorithm on the exact same dataset, the accuracy raised to 73%, still below the performance of the results of Wong et al. Once again, an algorithm using differentiation between subject and background clearly outperformed previous works where such characteristic was not contemplated. The obtained results were calculated by means of a different dataset comparing with the one used in this project. Thus, the reported accuracies cannot be directly matched.

Sun et al. [22] tried a different approach by calculating separated saliency maps for features connected with intensity, colour, orientations, and face locations, and comparing them with the saliency mask. This solution addressed the

problem of personalized photo assessment, taking user preference into quality evaluation. Although being a slightly different problem compared with the one proposed in this project, it is interesting to explore ideas which can contribute to improvements on a potential solution.

Bhattacharya et al. [2] used an aesthetically driven approach to create an interactive application that enables users to improve visual appeal of digital photographs using spatial recomposition. The process was designed in such way that the user is guided by the system which presents recommendations for possible improvements.

The authors included just two aesthetical features. The first one was the already described rule of thirds, which is applied to images containing a single and distinctive foreground object. The second feature was the visual weight ratio for pictures of landscapes or seascapes. In this case, the image has two distinctive parts, divided by the horizon. The visual weight ratio suggests that the ratio of the areas corresponding to each part should approximate to the golden number.

In order to test the importance of the selected features, Bhattacharya et al. built a dataset for each of the identified categories: single-subject compositions and images of landscapes/seascapes. A total of 632 photographs were downloaded from free image sharing portals, like Flickr. A subset of 150 random images for each category was used to train a SVM-based model. Performing grid search on the RBF kernel parameters, it was possible to achieve an accuracy of 87.4% for photographs with single foreground and 96.1% for landscapes/seascapes. There are no details on the techniques used to perform the grid search. The results can be easily over-fitted to data, which means the system would probably perform significantly worse with a new set of pictures. Again, the usage of a different dataset does not allow a direct result comparison.

Cheng et al. [23] developed a completely autonomous system of photo enhancement. The learning framework is quite different from previous works. The objective was developing an algorithm which learns some aesthetical features by example. After that, the user could introduce wide-field camera

inputs, *e.g.*, zoomed-out or panoramic photos, and receive the output picture corresponding to a region of interest of the original image.

In the training phase, professional photographs were segmented using a modified version of [20]. From each extracted patch, the first three colour moments of the three colour channels were calculated. Moreover, texture information was retrieved by calculating the histogram of 8 oriented gradients for each of the 4x4 spatial regions of the patches. These two features allowed the creation of a 137-dimensional feature vector per image. By using K-means algorithm, with K=1000, and determining each clustering centre properties, a visual vocabulary was constructed in order to identify representative patches. A face detection algorithm was also applied, and each identified face is associated with one of the visual words: frontal face region, left profile, or right profile. Due to the large variation of the training photos, the authors thought it would be interesting to group them in sub-topics by, again, applying the K-means algorithm, with K=100, and counting the occurrence of each visual word. Each group of images had its own model, instead of a general one. Each model was composed by a spatial probabilistic distribution for the visual words and a probabilistic distribution of pairs of words.

The enhancing process used the calculated models to select a region of interest in images which would make them more aesthetically appealing. Again, after segmentation, the patch characteristics, colour moments, histogram of oriented gradients, and face occurrence, were computed. A visual word, from the ones calculated during the training phase, was associated with each patch by using nearest neighbour algorithm. Such an algorithm was again applied to determine to which of the sub-topics the image belongs. After knowing the sub-topic of a photo, a probabilistic model was built. The algorithm performed an intensive search within the image to find the region of interest which maximizes the value of the probabilistic model.

The pictures used to train the model were obtained by crawling the online sharing website, Flickr. Just pictures about landscapes were considered, being downloaded 80000 of those. The system was tested by computing the results of

76 wide-field photos and asking 50 subjects for their opinions. Most of them agreed the output was better than the input.

Desnoyer et al. [24] addressed the problem of developing a system which would allow autonomous agents to create pleasing imagery. They stated the previously used datasets from large online communities were not appropriate for learning how to take a good picture, since they represented results late in the media generation process, as the photographers had already framed the shot and picked the best results.

Similarly to the previous works, a set of features was developed. Colour, complexity, contrast, texture/blur, and uniqueness, were considered. Almost all the used features were based on previous work. However, some novel measurements were included. Some examples are the Michelson contrast measure and the RMS of the intensity values.

For classification, the authors decided to use AdaBoost. The parameters of the classifier were selected by applying 5-fold cross validation. The algorithm was tested on two datasets: one downloaded from DPChallenge, similar to the ones used in previous works, and the other constituted by pictures taken by the Apollo astronauts while on the Moon. The latter did not suffer any kind of filtering, providing the required examples of the early stages of the creation process. Such images were labelled using a survey. The system achieved an error rate of 18.1% on the Apollo dataset and 38.2% on the DPChallenge one. The most successful features had similar results for both datasets, meaning they are components of universal beauty. Desnoyer et al. concluded this kind of algorithms could be used in autonomous agents. They mentioned their next step would be developing an autonomous robotic cameraman that can create great imagery. Although some of the photos were retrieved from DPChallenge, the same source used in this work, it is a different dataset, which means the results are not comparable.

Li et al. [25] addressed a problem which had not been much explored. They focused their efforts on pictures with faces. Firstly, a face detection algorithm was applied. The identified faces were considered the subject while the

remaining area was labelled as background. The extracted features were divided in three sets: technical, social, and perceptual. In terms of technical features, the authors used some of the measurements previously described in this document. Colour, brightness and clarity contrast between the subject and the background were among the computed components. The social relationship features included the relative position of faces, measurements of face expression, like mouth and eye openness, and face pose, by calculating a 3-dimensional vector representing the face orientation. The perceptual features are mostly connected with rules of composition. Concepts like the rule of thirds or symmetry on faces distribution were included. As this was a recent research area, there was no public datasets available. From Flickr, it was downloaded a set of 500 images containing faces and, through a survey, the photographs were rated. The system achieved an accuracy of 68% on the categorization test, where a SVM classifier placed the images into one of five categories, depending on their score. The reported accuracy was obtained using a leave-two-out cross validation type, but cannot be compared with the results of the current project, since the used dataset was not the same.

Yeh et al. [26] developed a personalized ranking system for personal use. The system ranks pictures by manual introduction of the weights of the features and by manual selection of examples of good images. Such approach, allows users to tell the algorithm their preferences.

The used features were mostly based on the existing works. The rules were divided in three categories: photograph composition, colour and intensity distribution, and personalized features.

Figure 2-4. Simplicity example. (a): high simplicity; (b): low simplicity.

In terms of photograph composition, the authors included the rule of thirds and two simplicity features. The first one calculates the ratio between the subject and the total image area and the second one is the simplicity measurement used in [15]. The simplicity principle is represented in Figure 2-4.

Concerning to the colour and intensity features, texture, clarity, colour harmonization, intensity balance, and contrast were considered. Texture was extracted from the MPEG-7 homogeneous texture descriptor, which uses Gabor filters to evaluate responses of the image under different scale and orientations. Clarity is the characteristic that measures the blurring levels. Two features were used. The first one calculated the ratio of clear pixels in the image using the technique described in [11]. The other feature attempted to detect the bokeh effect in an image by partitioning it in grids and applying blur detection on them. This effect is known by creating out-of-focus light points. If the number of clear grids is inside a pre-determined interval, the image is said to have bokeh. Colour harmonization feature was calculated by applying a set of hue templates, which are believed to describe the human colour perception, and reporting the value which minimizes the distance between the hue histogram and each of the different templates. To compute the intensity balance, the sum of the absolute differences between symmetric pixels was calculated. If the image is completely symmetric, the final value is zero, otherwise it is as high as the difference between left and right areas. The final considered characteristic was the contrast. For such purpose, the authors included the Weber contrast, which measures the disparity between pixels in terms of intensity. A colour contrast feature was also used, in this case, the colour distance between each pair of segments. For that, the image was segmented and each of the calculated segments assumed the colour of its pixels. The feature value corresponded to the distance-weighted sum of the colour-distance of each pair of segments.

The personalized features included average values of the brightness and saturation channels and the ratio of each channel of the RGB colour space. Further used features include the aspect ratio of the image and two Boolean

values which report the presence of human faces and if the image is in grey scale.

For the ranking test, Yeh et al. used ListNet as classifier. The results were presented to the user ordered by the achieved classification. In order to compare the system with previously developed ones, the authors used the algorithm on the same dataset used by Ke et al. [11] and Luo et al. [15]. The overall result was the same as the one achieved by Luo et al., 93% of accuracy, but the individual features performance was significantly higher. Similarly to the work developed by Luo et al., there are no details on parameters selection. Thus, is difficult to understand to what extent the results are over-fitted to the testing data. However, this work uses the same dataset as in [11] and [15], which will also be part of this work. Thus, the results can be used to comparison.

In [27], Huang et al. created a framework to personalized ranking of portraits. The overall system is very similar to [26]. However, some features, like the ratio of the body part appearing in the image, were originally introduced.

In 2011, Bhattacharya et al. [3] continued their previous work [2]. Among other improvements, the authors included a tilt correction tool and optimal object replacement, in which case requires the previous object place to be repainted. A patch-based filling algorithm was used to perform that task. Under the same conditions as in [2], it was reported an accuracy of 86%.

Gadde et al. [4] developed a photographer robot able to take pictures using aesthetical principles. The whole system was based on the work developed in [15]. The main difference was on using a different algorithm to determine the subject region. By performing the exact same experience as in [15] the authors reported an accuracy of 79%, using the same dataset as in this work. The learned principles were used to guide the robot on the photographic process.

Dhar et al. [28] tried to build a feature set of higher level than the previous developed work. The system itself is similar to [15], and achieved comparable results. However, some of the implemented features are different. In this case,

depth of field measurements, opposing hues calculation, indoor/outdoor classification, presence of people or animals, and sky-illumination attributes detection were performed. It was done a much more intensive work on the attribute testing. For each feature, the authors calculated the precision recall curves in order to detect the most effective ones.

Geng et al. [29] addressed the problem of web image search. Although in theory some of the prior work could be used for such purpose, they included some context information on image evaluation process, like EXIF data, webpage structure and content. The aesthetic features were the same as in previously developed systems. It was reported a 12.9% precision improvement, using the contextual information.

Meur et al. [30] proposed a method for predicting the visual congruency of an image. Visual congruency reflects the variation among different subjects looking at the same picture. If the image has a very distinctive foreground, it will have a high congruency. The results were tested on a specific database to eye-tracking problems. Some interesting features were calculated. The first one was the inclusion of a face detector, since the inclusion of human faces significantly impact the visual deployment of the observer. The same colour harmony feature used in [26] was also included. Furthermore, the authors used a probabilistic model to compute the depth of field for each pixel in the image. An estimate of scene complexity was included by calculating the entropy of the wavelet coefficients. The last two used features are the number of regions produced by meanshift segmentation and the amount of contours, detected using Sobel edge detectors on vertical, horizontal and diagonal directions. Although the features of this work are similar to the ones used on aesthetics evaluation, the problem itself is quite different, so it is not relevant to report them.

The work developed by Su et al. [31] was very different from what had been done so far. They addressed images with no clear distinction between subject and background. The applied features were not innovative. However, it would be quite difficult to apply a subject detection algorithm, so an alternative approach was developed. The images were divided in patches, which represent

areas seen as subject, in black, and background, in white. Features were calculated for the multi-size image patches and a bag of aesthetics features was generated by applying patch-wise operations on the feature vectors. Such bag of aesthetics is used to feed an AdaBoost classifier. In the same dataset used in [23], which is different from the one used in this project, the system achieved an accuracy of 92.06%, significantly higher than the comparing work. The authors also noticed that contrast features perform better than absolute measurements.

Recently, there have been developments on image memorability estimation. Isola et al. [32] [33] developed a way of predicting such characteristic. This is a slightly different problem comparing with aesthetics assessment, but it shows possible uses of the techniques developed during the last decade on aesthetics evaluation applied to other domains.

Moorthy et al. [34] wrote a paper analysing the state of the art on visual quality assessment and pointing possible research directions. The authors pointed out the lack of explicit databases for image appeal. Current available datasets were built by downloading images from photo sharing websites like Flickr, Photo.net and DPChallenge. However, users rate pictures under different conditions. Different monitors, for example, can lead to a very different evaluation, since they're not calibrated in the same way. Thus, current datasets are full of noise making the task of developing and evaluating algorithms rather difficult. Furthermore, Moorthy et al. reported there are no standardized methods for results evaluation. This problem makes the task of comparing results extremely difficult. Another identified issue is the lack of large-scale theoretical subjective studies on image appeal. Scientific community would then probably assist to a serious improvement on feature quality. Finally, the authors suggested future approaches should include content-describing features, since users tend to be influenced by the meaning of images. Other recommended direction was the creation of personalized systems, instead of general one, in order to overpass the subjectivity issue and develop custom solutions for each user.

The problem of aesthetical assessment on photographs reached a fairly complex level. Lots of different features were developed. However, there is always room for new ones: this is an exploration problem and new additions are always welcome. As mentioned in [34], there are no standardized methods for results evaluation which makes the algorithmic comparison to be extremely difficult. Even when the same dataset is used, there are many other variables which are not implemented in the same way. The parameter selection for the machine learning algorithms, for example, is rarely addressed on the analysed papers. This is a crucial step when it comes to over-fitting avoidance. In this work, I want to test some new features and, at the same time, introduce some practices on the testing phase which can make the whole process more clear and reliable.

3 ELEMENTS OF A GOOD PICTURE

In this chapter, I will explore some aesthetical concepts which can lead to good photography. This study assumes a big importance since the features to be developed should be highly correlated with the aesthetical principles which guide professional photographers on their demand for great pictures.

Among all the books and research papers, there is a generalized principle which seems to assume the highest importance: the existence of a strong point of emphasis. This prominent area is also known as subject of a picture. A good photography should be simple in the way that the observer knows exactly where to look at. Such simplicity can be achieved by creating strong distinction between the subject and the background. Several techniques can be used to produce that characteristic. In [35], the authors point out positioning and contrast as characteristics to take into account. In terms of positioning, an amateur photographer intuitively places the subject at the centre of an image. However, there is a viewer-researched classical guide, called *Rule of Thirds*, which is believed to take into account much more prominent positions. This rule has some interesting concepts behind, so it is worth explaining its origin. We need to go far back in time: not to the renaissance, not to any of the ancient civilizations, but to the origin of the Universe. There is a number which seems to rule the course of Nature. People name it Φ, or Phi, and its value is approximately 1.618. A large number of phenomena are related to this proportion: the organization of galaxies' spirals, the growth of sunflowers' seeds, and the hawks' diving path are just some examples of such singularity. Because of this property, this number is also known as Divine proportion.

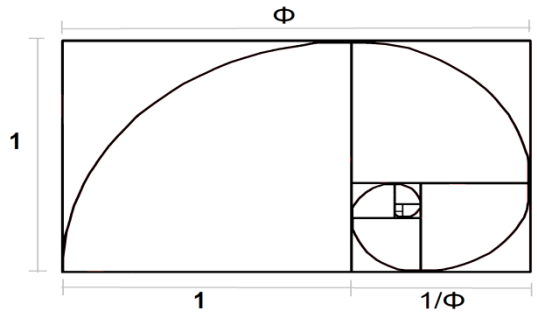


Figure 3-1. The Divine Proportion.

Figure 3-1 shows a spiral which follows the Divine proportion. In each division, a square is created and the remainder is used in the new iteration, following the mathematical formula

$$\Phi = 1 + \frac{1}{\Phi}, \tag{3.1}$$

which has a possible solution of $\frac{1+\sqrt{5}}{2}$, the golden number. The spiral converges to a point which is believed to be a crucial position to aesthetics. Photographers use a rule of thumb which tries to approximate this behavior: the rule of thirds. In Figure 3-2 we can observe that by dividing the image in 3 equal parts, both vertically and horizontally, we can approximate the spiral convergence points. The rule of thirds states that the subject of an image should be placed near one of those key points.

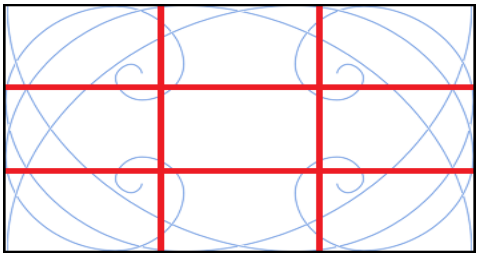


Figure 3-2. The Rule of Thirds.

The other mentioned characteristic is the contrast, so that the subject and the background are highly distinguishable. In [35], the referred contrast is essentially connected with brightness. Figure 3-3 shows an example of that concept.

Figure 3-3. Subject region enhanced by brightness contrast.

However, there are other contrast types. In [11], besides the already stated lighting dissimilarity, it is also mentioned colour and blur contrast. These two concepts are shown in Figure 3-4 and Figure 3-5, respectively.

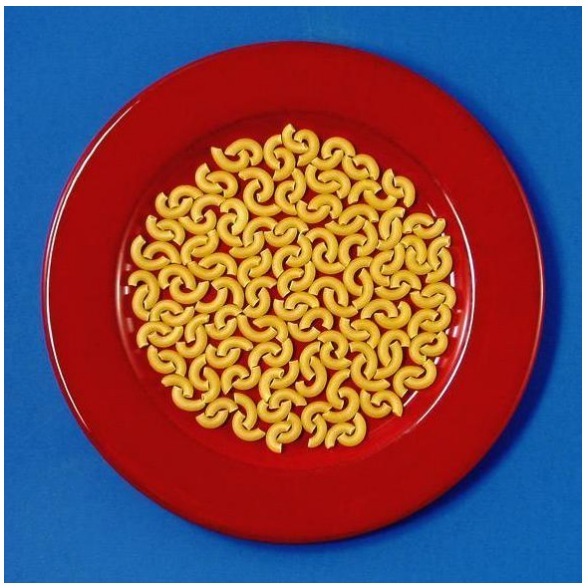


Figure 3-4. Colour contrast.

The so important simplicity concept states that someone who is observing a picture should know exactly where the subject is. Thus, another relevant characteristic is the simplicity of the background, so the viewer doesn't get distracted from the core of the image.

Figure 3-5. Blur contrast.

The aesthetical study performed in [11] was mainly based on [12] and [13]. In addition to the mentioned contrast principles, the authors highlighted the importance of realism. Contrarily to snapshots, which are realistic scenes,

professional photos often use a set of techniques to produce surreal representations of a view. Professionals have the main objective of impressing their audience, which can be achieved by using a large number of intense and saturated colours. Figure 3-6 is an example of the described characteristic. In (a), the photographer was careful in choosing a specific time of the day in order to achieve the presented colour palette. Possibly, this picture was also post-processed.

(a) (b)

Figure 3-6. Photographic realism. (a): "Golden Gate Bridge at Sunset" by Buzz Andersen, 2005, has a certain level of surrealism; (b): "Golden Gate 3" by Justin Burns, 2005, is a realistic view of the scene.

It may seem that just a few aesthetical characteristics were described in this section. However, from these simple rules, we can derive a fairly large set of features. By taking in account principles used by professional photographers, it is expected the developed features to have a high correlation with the concept of good quality picture.

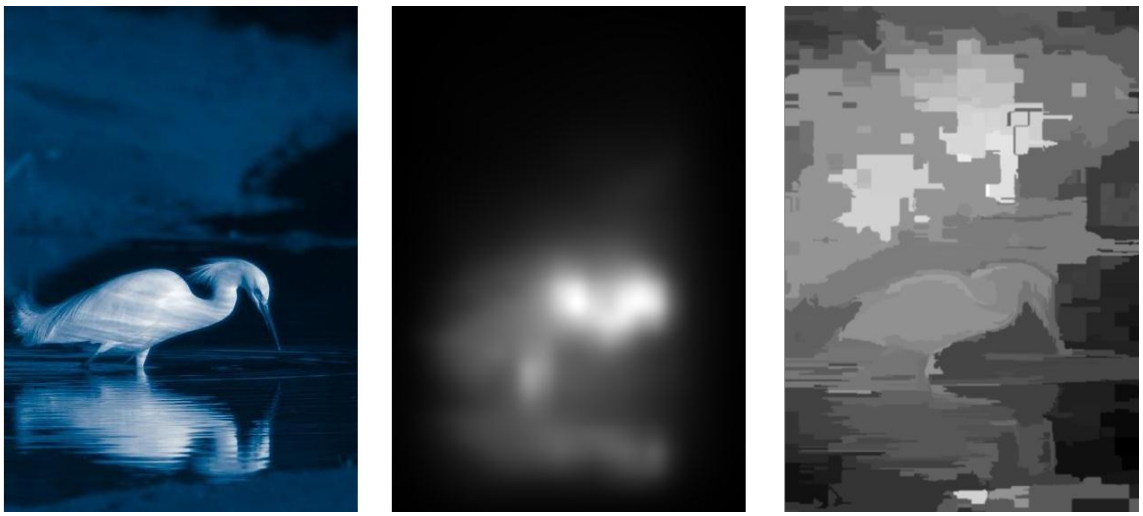
4 VISUAL FEATURES EXTRACTION

In this chapter I introduce the developed and extracted features for aesthetics assessment. A feature can be seen as a measurement of an image characteristic. The objective is evaluating the images and calculating numbers corresponding to the considered characteristics. Obviously, in this project, the introduced features try to report information relative to the studied aesthetics principles.

For a better organization of this document, I will divide this chapter in four sections, corresponding to different types of features: global, subject, background and subject/background relation.

4.1 Saliency Map

The saliency map tells us the saliency relation between pixels in an image. Figure 4-1 shows two different examples of saliency maps in (b) and (c). These kinds of maps describe pixels saliency by associating them with a value between 0 and 1, where 0 corresponds to the minimum possible saliency and 1 to the maximum.



(a) (b) (c)

Figure 4-1. Example of saliency maps. (a): original image; (b): graph-based visual saliency; (c): region based contrast saliency.

There are several different algorithms which compute saliency maps. This work uses two of them: the graph-based visual saliency from [18] and the region based contrast saliency from [21]. As depicted in Figure 4-1 these are quite different from each other. In the GBVS, pixels are treated independently, so the corresponding saliency map does not provide a strong region separation. On the other hand, the RBCS groups similar pixels so the saliency map is composed by saliency regions. I take advantage of the characteristics of both approaches and use them in different contexts, which will be described later on this document.

The first used saliency map technique was the GBVS. This approach involves three main steps: first extracting certain features, then forming activation maps by using the extracted feature channels, and then normalizing and combining the different activation maps. The used features are four orientation maps corresponding to orientations $\phi = \{0^o, 45^o, 90^o, 135^o\}$ using Gabor filters, a contrast map computed by means of the luminance variance in a local neighbourhood of size 80×80, and the luminance map itself. Each retrieved feature has a correspondent activation map. For that, the whole map is interpreted as a fully-connected graph, with each edge having a weighting measurement which depends on the dissimilarity between the two connected nodes, which can be seen as pixels, and on the distance between them. The dissimilarity is defined as

$$d\big((i,j)\|(p,q)\big) = \left|\log\frac{M(i,j)}{M(p,q)}\right|, \tag{4.1}$$

where M is the feature map. Thus, the final weight is calculated

$$w\big((i,j),(p,q)\big) = d\big((i,j)\|(p,q)\big).F(i-p, j-q), \tag{4.2}$$

$$F(a,b) = exp(-a^2 - b^2), \tag{4.3}$$

which means w will be proportional to the dissimilarity and to the closeness of two nodes in the graph. From the constructed graph, a Markov chain is created by normalizing the outbound edges' weights to 1. The equilibrium distribution of

this chain would naturally accumulate mass at nodes with high dissimilarity, since such transitions are associated with a higher likelihood than similar nodes. This way, we end up with an activation map which describes the saliency on a certain feature domain, which is derived from pairwise contrast. The objective is to combine all the calculated activation maps so the whole feature set is considered. For that, it is necessary to normalize each of them, since after combination the master map may be too nearly uniform, which is not very informative. Hence, it is important to concentrate activation in a few key locations. For that, it is performed a new weighting step

$$w2\big((i,j),(p,q)\big) = A(p,q).F(i-p,j-q), \tag{4.4}$$

where A is the previously calculated activation map. After that, the Markov chain process is repeated once more in order to calculate the mass concentration at each node. Different activation maps can then be combined in order to present the final image saliency.

The RBCS is a totally different method. The main difference of this algorithm is that it uses a region-based approach. Thus, the image is segmented in first place using a graph-based method [20]. For each resulting segment, it is calculated the saliency by comparing its colour contrast to all other regions in the image, such that

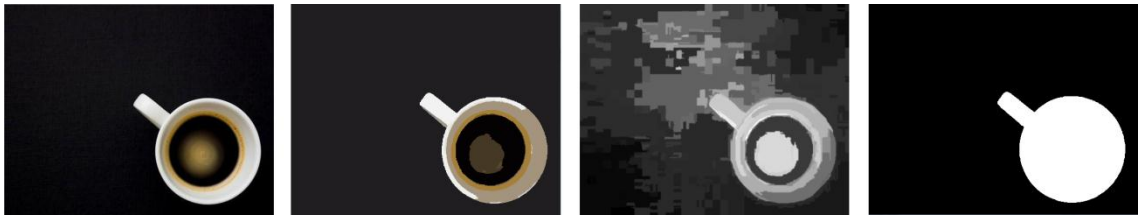
$$S(r_k) = \sum_{r_k \neq r_i} exp\big(-D_s(r_k,r_i)\big).w(r_i).D_r(r_k,r_i), \tag{4.5}$$

where D_s is the Euclidean spatial distance between the two regions' centroids, w is the relative region size, and D_r is the Euclidean colour distance in the L*a*b* colour space. This way, not only the colour disparity between regions is being considered, but also their closeness and size.

4.2 Subject Mask

The subject mask is a binary image which makes the distinction between subject and background. Developing such an algorithm is not a trivial task.

Many methods exist, but none of them achieves perfect results. However, some give a good insight about the subject/background distinction.



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

Figure 4-2. Elements of the subject mask calculation. (a): original image; (b): meanshift image; (c): RBCS map; (d): subject mask.

The subject region is known by having a higher saliency comparing to the background. It suggests that we can use the saliency map to predict which image pixels constitute the subject and which ones belong to the background.

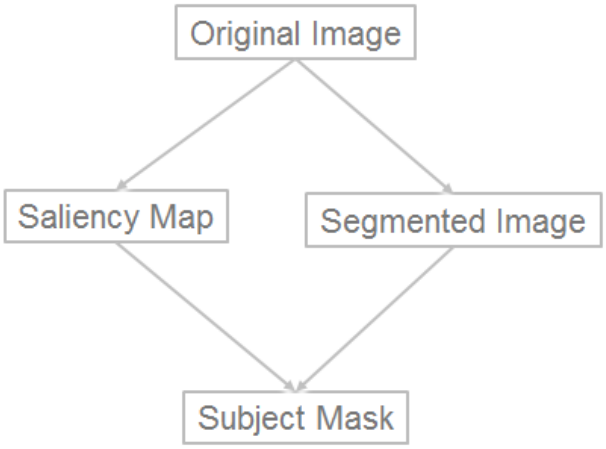


Figure 4-3. Subject mask calculation diagram.

I use the method developed in [36]. Firstly, the meanshift algorithm segments the image. This step is useful since the subject is interpreted as an area and not independent pixels. Thus, similar pixels can be grouped to form connected components within the image. For each segment, it is calculated the average saliency value using the RBCS map. The subject is constituted by those segments which have average saliency twice the average image saliency. This way, we end up with a binary image which assigns each image pixel to one of the subject or the background. Figure 4-3 shows a simple flow of the described

process. I use the RBCS map because it provides us a little insight about region saliency. However, the saliency map segmentation is not refined enough, so the referred extra segmentation step is required. Figure 4-2 shows the different elements of the subject mask calculation.

4.3 Global Features

Global features are those which measure characteristics from the whole image. Every single pixel is treated in the same way without any kind of distinctions.

4.3.1 Simplicity

Simplicity is an important characteristic to evaluate, since it tells us if the image has a large number of components, spreading the observer attention.

0	1	0
1	-4	1
0	1	0

Figure 4-4. Laplacian filter.

The Laplacian filter, shown in Figure 4-4, is widely used on edge calculation. Convolving this filter with an image allows identifying the pixels which are significantly different from its neighbourhood. Complex areas are known for having lots of distinctive pixels. Figure 4-5 shows two examples of image filtering using the Laplacian.

Figure 4-5. Laplacian image example. (a),(c): original image; (b),(d): result after applying Laplacian filter.

The first developed feature, *f1*, calculates the normalized area of the bounding box which encloses 75% of the energy present in the Laplacian image. The smaller the box, the simpler the image is. This feature is based on the ones used in [11] [24]. Such a calculation is performed by computing the vertical and horizontal histograms of the Laplacian and discarding $\left(1-\sqrt{0.75}\right)/2$ of the energy in each end. Figure 4-5 (a) returns a value of 0.283, while Figure 4-5 (c) result is 0.736.

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

Figure 4-6. Saliency map characteristics. (a): distinctive subject image; (b): saliency map of (a), with mean value of 0.341 and standard deviation of 0.239; (c): complex subject image; (d): saliency map of (c), with mean value of 0.444 and standard deviation of 0.178.

The saliency map provides information about the pixel prominence. Thus, we can use it to determine the scene complexity. Two features are extracted from the saliency map, which correspond to its first two statistical moments, the mean value, in (4.6), and the standard deviation, in (4.7).

$$f2 = \frac{\sum_i \sum_j I(i,j)}{N}, \tag{4.6}$$

$$f3 = \sqrt{\frac{\sum_i \sum_j |f2 - I(i,j)|}{N}}, \tag{4.7}$$

where *I* represents the image, *i* and *j* the pixel coordinates and *N* the total size of the image.

The mean value can tell the contribution that both high and low saliency areas have in the image. On the other hand, the standard deviation reports in which way the saliency values are close to each other. It is expectable that professional images have a higher standard deviation than snapshots, since

subject regions should have much higher saliency than the background. Figure 4-6 shows the example of two different images. (a) has a much more distinctive subject than (c), which translates in having lower saliency mean and higher standard deviation. *f2* and *f3* are original features, which means they were not tested before. It will be interesting to observe their performance during the classification phase.

(a) (b)

Figure 4-7. Hue count feature. (a): with few hue values, scores 18; (b): with the presence of several hue values, scores 7.

In the same way, the count of present hues can be a good measure of complexity. If an image has a large number of different hues, we expect it to contain a much more distracting scene. A professional photography usually achieves its dynamics mostly by mixing different tones of a few hue values, while a snapshot often contains many objects, each with its own distinctive colour. In this work I include the hue count feature from [11]. A 20-bin histogram of the hue component is created. Just pixels with brightness values in the range [0.15; 0.95] and saturation greater than 0.2 are considered, since hue calculation would become inaccurate otherwise.

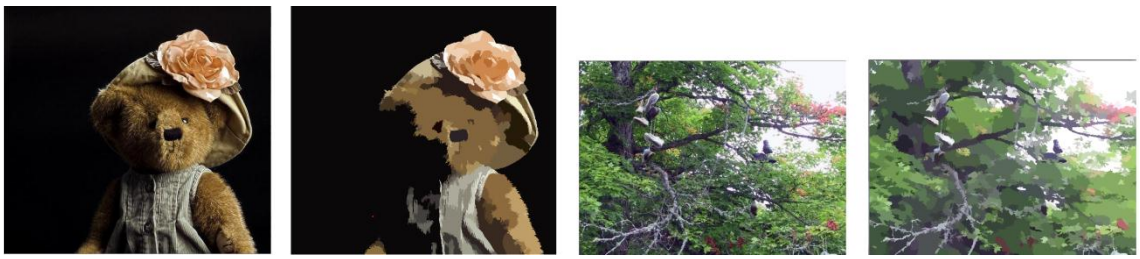
$$N = \{i \mid H(i) > \alpha m\}, \quad \textbf{(4.8)}$$

where *H* is the calculated histogram, *m* is the maximum occurrence in *H* and *α* is the sensitivity parameter. In this work *α* = 0.05 is used.

The hue count feature, *f4*, will be

$$f4 = 20 - \|N\|. \quad \textbf{(4.9)}$$

Figure 4-7 shows two examples of the application of this feature. (a) is much simpler than (b), achieving a much higher score.



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

Figure 4-8. Number mean shift segments feature. (a),(c): original images; (b): segmented image with 226 produced segments; (d): segmented image with 1108 produced segments.

As previously mentioned, mean shift algorithm is a clustering technique widely used in computer vision and image processing. We can use it to perform colour segmentation in an image by grouping similar connected pixels into a single colour area. Each region has the average value of its pixels. A monotonic image is naturally segmented since it just includes one colour. On the other hand, applying mean shift to an image with a large complexity will not make a big difference, since it is difficult to find similar connected pixels. This characteristic suggests a new way of measuring the simplicity of a picture. I include another feature from [24], *f5*, which count the number of segments after performing the mean shift algorithm in an image. Figure 4-8 shows two examples of segmented images. (a) is much simpler than (c). Applying the mean shift algorithm will result in the production of 226 segments in (b) and 1108 in (d), suggesting this measure can be used to quantify simplicity.

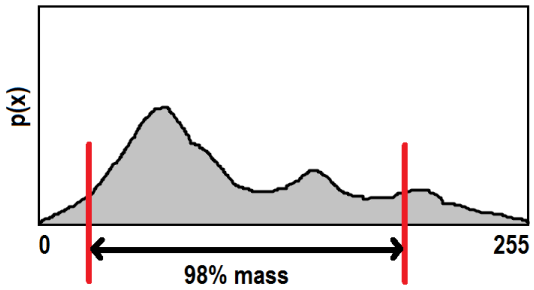


Figure 4-9. Example of the luminance histogram width feature, where the middle 98% of the histogram mass is computed.

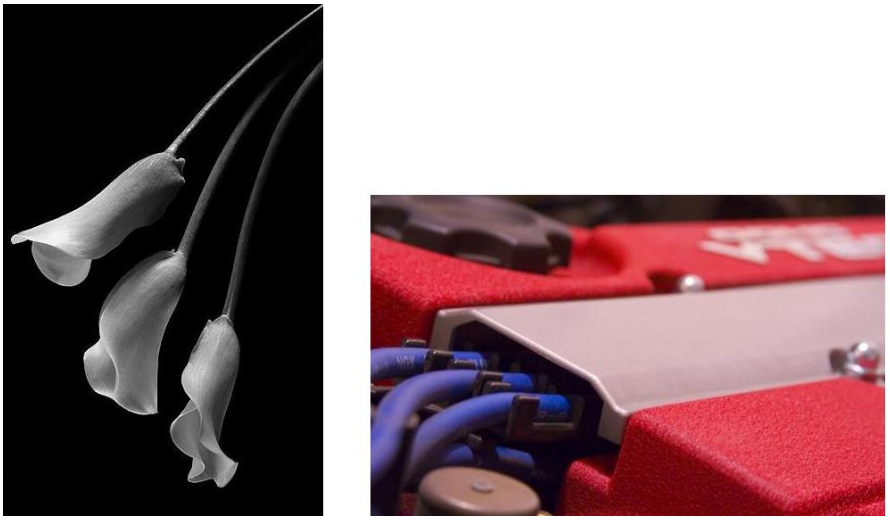
Another possible way of measuring the complexity of a photograph is by analyzing the distribution of its intensity histogram. If the histogram is nearly uniform, it will have a high number of different intensities, thus a high complexity. I use the luminance histogram width feature, *f6*, from [21], which is a modified version of a similar feature included in [11]. In this case, as shown in Figure 4-9, the width of the middle 98% luminance mass is reported. Figure 4-10 shows two practical examples of the use of *f6*.

(a) (b)

Figure 4-10. Luminance histogram width feature. (a): Low luminance variance, scores 173; (b): High luminance variance, scores 239.

4.3.2 Contrast

Contrast is an extremely important characteristic on determining aesthetically appealing imagery. Usually, professional photographers use this tool to enhance certain areas, making them more noticeable to the observers.



(a) (b)

Figure 4-11. Global contrast features. (a): high contrast image with Weber contrast = 256.13 and Michelson contrast = 1; (b) high contrast image with Weber contrast = 148.95 and Michelson contrast = 0.98.

Weber contrast is a commonly used measure to determine the contrast level of an image. Yeh et al. use it in [26], by calculating the average normalized difference of each pixel to the mean pixel value. It is described by

$$f7 = \frac{1}{N}\sum_{i}\sum_{j}\frac{L(i,j) - L_{avg}}{L_{avg}}, \quad \textbf{(4.10)}$$

where L represents the luminance image, L_{avg} the average luminance value, i and j the pixel coordinates and N the total size of the image.

Another well-known contrast measure is the Michelson Contrast, used in [24]. This calculation just considers the higher and the lower luminance values, evaluating the used luminance range.

$$f8 = \frac{L_{max} - L_{min}}{L_{max} + L_{min}}, \quad \textbf{(4.11)}$$

where L_{max} and L_{min} are the maximum and the minimum luminance values, respectively.

Figure 4-11 shows the difference between high and low contrast images. (a) has unarguably higher contrast than (b). Both Weber and Michelson measurements are consistence with this empirical observation. (a) has Weber and Michelson contrast values of 256.13 and 1, respectively. In the same way, (b) scores 148.95 and 0.98.

(a) (b)

Figure 4-12. Brightness average comparison. (a): brightness dynamics is used to enhance the subject, having an average brightness value of 22.35%; (b): snapshot with automatically adapted brightness, having an average brightness value of 53.14%.

Common cameras are programmed to automatically adjust some image characteristics in order to make them look reasonably good with small effort by the photographer. One of those automatic enhancements is the brightness correction. Most cameras adapt the average brightness to be 50% gray. However, professional photographers manually control the exposure so the subject and the background are significantly distinguishable, causing the average brightness to deviate from 50% gray. The larger the deviation, the more likely the photo was taken by a professional. This feature was previously used in [11] [21].

$$f9 = \frac{\sum_i \sum_j B(i,j)}{N}, \tag{4.12}$$

where *B* represents the brightness image component, *i* and *j* the pixel coordinates and *N* the total size of the image. Figure 4-12 shows the average brightness of two images. (a) looks like a professional photograph and scores 22.35%, while (b), a snapshot, achieves 53.14%.

4.3.3 Colour

(a) (b)

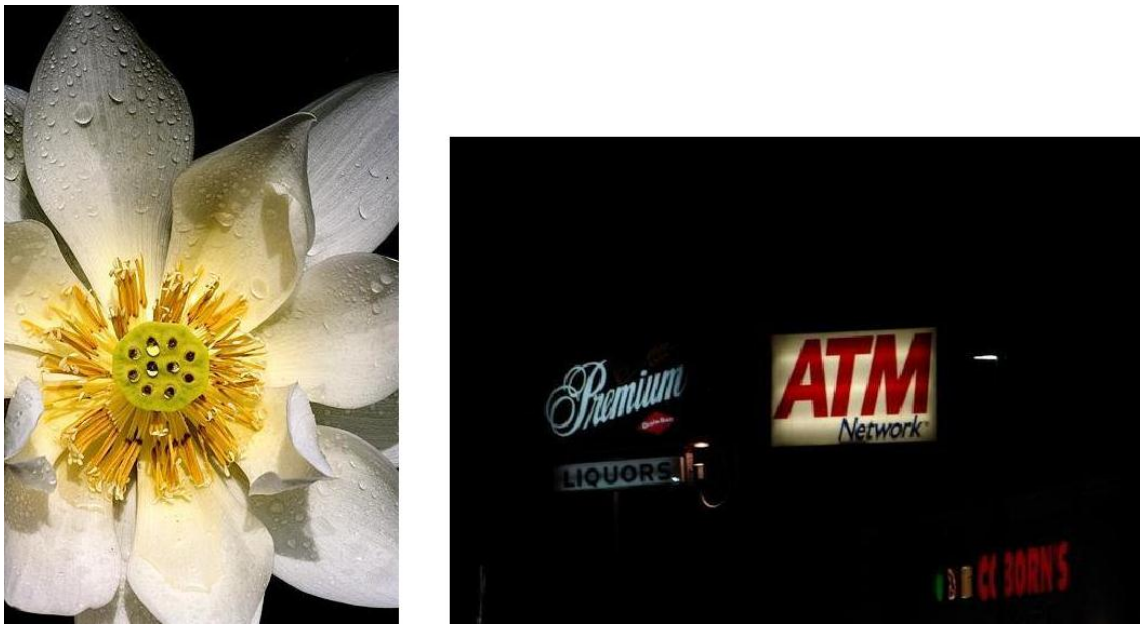
Figure 4-13. Average saturation comparison. (a): example of picture with pure colour usage, presenting a 57.59% value on the average saturation; (b): a much duller image, having an average saturation value of 35.63%.

Pure colours tend to be more appealing than dull ones. Chromatic purity can be measured by the saturation component of the HSV colour space. Some professional photographers even use specialized film to obtain deeper colours. Thus, the average saturation of images can be useful to distinguish between professional photographs from snapshots. I introduce the same feature as in [14],

$$f10 = \frac{\sum_i \sum_j S(i,j)}{N}, \tag{4.13}$$

where *S* represents the saturation image component, *i* and *j* the pixel coordinates and *N* the total size of the image.

Figure 4-13 compares two images concerning to their average saturation values. (a) uses much purer colours comparing to (b), achieving a value of 57.59% against 35.63%.



(a) (b)

Figure 4-14. Colour harmony comparison. (a): example of a picture which colours are closer to the high quality model, presenting a feature value of 1.52; (b): an image which is closer to the low quality model, having a feature value of 0.95.

I also use a modified version of the colour harmony feature introduced in [15]. A model representing each high and low quality images is created, so it is possible to determine how close each new image is to both components of the model. For the high and low quality photos in the training dataset, we can obtain the histograms of the average levels for each component of the HSV colour space such that

$$M_{high,hue}(i) = Average(H_{high,hue}(i)), \quad \textbf{(4.14)}$$

$$M_{low,hue}(i) = Average(H_{low,hue}(i)), \quad \textbf{(4.15)}$$

where $M_{high,hue}$ *and* $M_{low,hue}$ respectively represent the model for high and low quality images for the hue component, *i* is the histogram bin, and $H_{high,hue}$ *and* $H_{low,hue}$ are the histograms of each high and low quality images, respectively, for

the hue component. The other model components, $M_{high,saturation}$, $M_{low,saturation,}$ $M_{high,brightness}$ *and* $M_{low,brightness}$, can be similarly calculated. In order to reduce the computational complexity, each histogram is limited to its 50-bin representation.

The original implementation model was unnecessarily complex: the authors computed the combination of each histogram bin pair by adding both components. However, this step does not provide any advantage since it is just replying existing information and increasing the problem complexity. The applied simplification allows us to deal with much smaller histograms (50-dimensional) than the ones created in [15] (1274-dimensional).

Now that we have a model for the high and low quality images, the images can start being compared with it. For that I compute the cross-product distance between both the model and the image histogram for each component such that

$$Distance_{high,hue} = \sqrt{\left|M_{high,hue}\right|^2 |H_{hue}|^2 - \left(M_{high,hue}.H_{hue}\right)^2}, \quad \textbf{(4.16)}$$

$$Distance_{low,hue} = \sqrt{\left|M_{low,hue}\right|^2 |H_{hue}|^2 - \left(M_{low,hue}.H_{hue}\right)^2}, \quad \textbf{(4.17)}$$

where H_{hue} is the 50-bin hue histogram of the image to be evaluated. Similarly the other distances, $Distance_{high,saturation}$, $Distance_{low,saturation}$, $Distance_{high,brightness}$, and $Distance_{low,brightness}$, can also be computed.

The final feature is calculated by

$$f11 = \frac{Distance_{high,hue}}{Distance_{low,hue}} \times \frac{Distance_{high,saturation}}{Distance_{low,saturation}} \times \frac{Distance_{high,brightness}}{Distance_{low,brightness}}. \quad \textbf{(4.18)}$$

Figure 4-14 shows two example images: (a) is much closer to the high quality model, while (b) is more similar with the low quality one. (a) and (b) presents feature values of 1.52 and 0.95, respectively.

4.3.4 Blur

Blur is a key concept on photography. To evaluate the general level of blurriness I use the same feature as in [11]. A blurred image is described as

$$I_b = G_\sigma * I_o, \tag{4.19}$$

where I_b is the blurred image, I_o is the original image and G_σ is the blurring filter. The objective is recovering the σ parameter, given only I_b. The blurred image can be converted to the frequency domain by means of the Fourier transform such as

$$F = FFT(Ib), \tag{4.20}$$

where FFT denotes the Fast Fourier Transform. If we assume that the frequency distribution of the original image, I_o, is approximately uniform, we can count the number of frequencies with power above a certain threshold like

$$C = \|\{(u, v) \mid F(u, v) > \alpha m\}\|, \tag{4.21}$$

where *u* and *v* are the frequency coordinates, *m* is the maximum power in *F* and α is the sensitivity parameter. α = 0.2 was used. Since a smoothing filter just removes high frequencies, *C* can be an approximation for the maximum frequency present in the blurred image. Thus, the final feature is

$$f12 = \frac{C}{\|I_b\|} \sim \frac{1}{\sigma}. \tag{4.22}$$

4.3.5 Others

There are other image characteristics which potentially correlate with photographic aesthetics. As suggested in [14], the aspect ratio can be a relevant measurement. Ratios which approximate the already mentioned golden ratio are considered more pleasant. For some reason 4:3 and 16:9 formats are standards for television or movie screens. This feature is introduced by calculating

$$f13 = \frac{w}{h}, \tag{4.23}$$

where *w* and *h* are the image width and height, respectively.

(a) (b)

Figure 4-15. Intensity balance example. (a): high balance image, with a feature value of 0.007; (b): low balance image, with a feature value of 0.050.

Balance provides a sense of equilibrium and is a fundamental principle of visual perception. The eyes of the observers usually seek visual balance within a photograph. A good photographer uses composition to balance the image weights with respect to some key aspects. One possible measure is the left-right balancing used in [26]. I start dividing the image in two halves using a vertical line. For each part, an intensity histogram is computed, which allows us to compare both considered areas. In [26], the authors directly perform the comparing operation on both histograms. However, the human eye has some difficulties distinguishing close intensities. In order to achieve a more accurate result, I convolve the histograms with a Gaussian filter. This makes each histogram position considering its neighbourhood, which reflects the relation between similar pixels. Both histograms can now be compared, by calculating the summed squared difference each histogram position

$$f14 = \frac{\sqrt{\sum_i (H_l(i) - H_r(i))^2}}{N}, \quad \textbf{(4.24)}$$

where H_l and H_r respectively represent the left and right histograms, i the histogram bin, and N the total size of the image. The normalization process is required so we can compare images with different sizes. Figure 4-15 shows a practical example of the intensity balance feature. (a), with a value of 0.007, as a higher balance than (b), with a value of 0.050.

4.4 Subject Features

The subject features only analyse the image pixels which are set as foreground in the subject mask. Although being based in previously developed features, I originally introduce all the measurements in this section to the context of subject evaluation.

4.4.1 Contrast

I again use the average brightness measurement, *f15*. The process is similar to the one descried in (4.12), this time performed over the subject area.

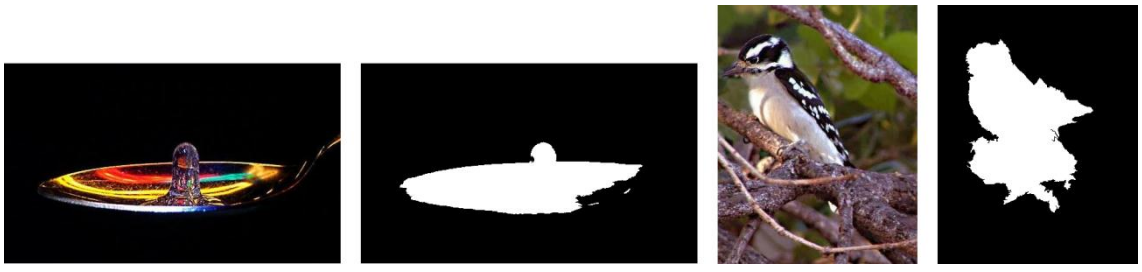
Similarly to (4.11), Michelson contrast is also used, creating a new contrast feature, *f16*.

The root mean square, RMS, measures the standard deviation of a distribution. I introduce three of these measurements, one for each of the luminance, *f17*, hue, *f18*, and saturation, *f19*. They follow the expression

$$f17{,}18{,}19 = \sqrt{\frac{1}{S_{area}}\sum_{s}\left(X(i,j) - X_{avg}\right)^2}, \tag{4.25}$$

where X represents one of the luminance, hue or saturation images, X_{avg} the same component average value, s the subject pixel coordinates and S_{area} the size of the subject area.

4.4.2 Colour



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

Figure 4-16. Subject average saturation. (a): high saturation subject, with feature value of 165.16; (b): subject mask of (a); (c): low saturation subject, with feature value of 80.11; (d): subject mask of (c).

Similarly to the global feature, shown in (4.13), I calculate the average saturation value, but this time in the subject area. The reasons of this inclusion are exactly the same as before. However, this feature, *f20*, is even more important since the subject is the most prominent area in a picture. In Figure 4-16, examples of this feature are presented: (a) has a much stronger subject in terms of saturation comparing to (c).

4.5 Background Features

The set of background features is calculated on the pixels which are considered part of the background. This is the part of the image which is complementary to the subject. The set of features described in this section assumes a great importance because the background of a good image has very distinctive characteristics: it usually is unobtrusive, so the subject area can have a much higher impact on the observer.

4.5.1 Simplicity

Colour simplicity is a good way of evaluating the background area. If an image background has high colour complexity, it will retain a lot of attention, distracting the observer from the real subject.

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

Figure 4-17. Background colour simplicity. (a): high simplicity, with feature value of 0.0034; (b): subject mask of (a); (c): low simplicity, with feature value of 0.0491; (d): subject mask of (c).

Similarly to [15], I use a feature, *f21*, which calculates the number of background colours which have a significant presence in the image. In order to reduce the computational complexity, each component of the RGB colour space

is firstly reduced to a 16-values scale. A 4096-histogram (16^3) can be then computed, with each bin corresponding to a different colour. The colour simplicity feature can be defined as

$$S = \{i \mid H(i) \geq \alpha m\}, \tag{4.26}$$

$$f21 = \|S\|/4096, \tag{4.27}$$

where *H* is the calculated histogram, *i* is the histogram bin, *m* is the maximum occurrence in *H* and *α* is the sensitivity parameter. *α* = 0.01 was used. This means the feature reports the number of histogram bins which are above a certain relative value. Figure 4-17 shows two images with different background complexity. The one in (a) has an almost uniform background, unlike (b) which is rather complex.

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

Figure 4-18. Background hues count. (a): low number of hues, with feature value of 18; (b): subject mask of (a); (c): high number of hues, with feature value of 6; (d): subject mask of (c).

The global hue count feature has potential to be applied to the background domain. Actually, it is probably even more relevant in such a context since the background simplicity has a higher influence on the picture quality than the subject. Thus, *f22* was created following the same process as (4.9), but computed on the background area.

4.5.2 Contrast

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

Figure 4-19. Background contrast features. (a): low contrast background, with Michelson contrast value of 0.736, luminance RMS of 0.068, hue RMS of 0.115, and saturation RMS of 0.287; (b): subject mask of (a); (c): high contrast background, with Michelson contrast value of 1, luminance RMS of 0.209, hue RMS of 0.211, and saturation RMS of 0.262; (d): subject mask of (c).

Still in the simplicity domain, we can expect the background of aesthetically pleasant images to have low contrast pixels. That way, I use some contrast measures already described in this document: Michelson contrast (*f23*), luminance RMS (*f24*), hue RMS (*f25*), and saturation RMS (*f26*). In Figure 4-19, it is possible to compare two images with completely different background types: (a) is much more uniform than (c), which is confirmed by the four considered background contrast values.

4.6 Subject/Background Relation Features

In this section, I will show the features which describe the relation between the subject and the background. This is probably the most important set of features since the most distinctive characteristic of a good photograph is a clear separation between these two areas.

4.6.1 Contrast

Since aesthetically pleasant images have high subject/background distinction it is worth measuring the contrast between these two areas.

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

Figure 4-20. Lighting ratio feature. (a): high subject/background disparity in terms of brightness, with feature value of 2.241; (b): subject mask of (a); (c): low brightness disparity, with feature value of 0.139; (d): subject mask of (c).

To measure the subject/background brightness disparity I used the lighting feature introduced in [15], which is described by

$$f27 = \left|\log\left(\frac{B_{subject}}{B_{background}}\right)\right|, \tag{4.28}$$

where $B_{subject}$ and $B_{background}$ are the mean brightness values of the subject and background areas, respectively. Figure 4-20 shows two images with a totally different use of lighting: in (a) the subject is much more distinguishable than in (c).

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

Figure 4-21. Subject/background HSV average difference. (a): high distinction, with hue average difference of 6844.88, saturation average difference of 6767.15, and brightness average difference of 21017.20; (b): subject mask of (a); (c): low distinction, with hue average difference of 869.91, saturation average difference of 1422.92, and brightness average difference of 2839.53; (d): subject mask of (c).

In order to extend the comparison between foreground and background, I included the features suggested in [21], where the authors compute the squared difference of the average values between the subject and the background for each of the brightness, *f28*, hue, *f29*, and saturation, *f30*, colour channels. The measurement is described by

$$f28,29,30 = \left(\frac{\sum_s X(s)}{S_{area}} - \frac{\sum_b X(b)}{B_{area}}\right)^2, \quad \textbf{(4.29)}$$

where *X* represents one of the brightness, hue or saturation channel images, *s* and b the subject and background pixel coordinates, respectively, and S_{area} and B_{area} the size of the subject and background areas, respectively. Figure 4-21 reports two examples of subject/background distinction on the HSV channels: in (a) all the difference of averages are higher than in (b), which means (a) has a more distinctive subject.

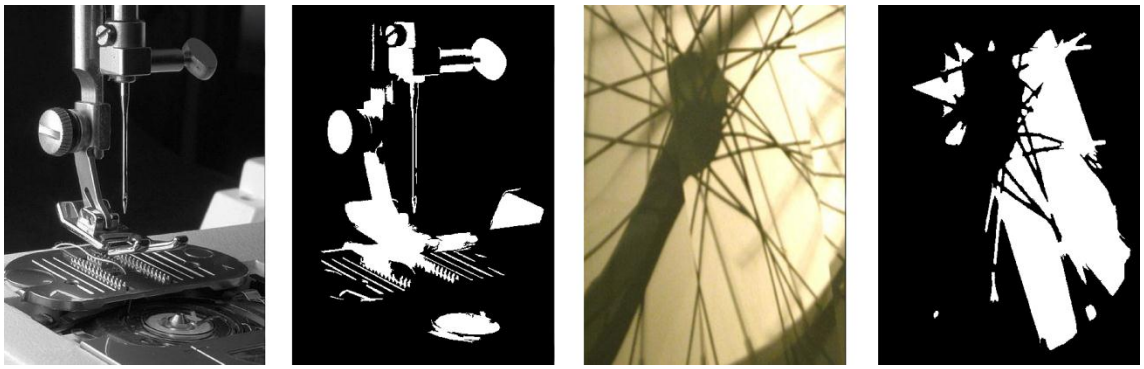
(a) (b) (c) (d)

Figure 4-22. Subject/background Weber contrast. (a): high contrast image, with a feature value of 1.963; (b): subject mask of (a); (c): low contrast image, with a feature value of 0.418; (d): subject mask of (c).

I also use an improved version of the Weber contrast. This is a measurement which can be easily applied to the subject/background context. It was not tried in any of the analyzed works but I believe it can be highly correlated to the aesthetical evaluation we are interested in performing. In this measure, it is calculated the normalized average distance of the subject to the average background value in the luminance domain. It allows us to measure the average luminance disparity between the two areas. The feature is described by

$$f31 = \frac{1}{S_{Area}} \sum_{s} \frac{L(s) - L_{background}}{L_{background}}, \quad \textbf{(4.30)}$$

where L represents the luminance image, $L_{background}$ the average luminance of the background area, s the subject pixel coordinates and S_{area} the total size of the subject. Figure 4-22 shows the application of *f31* on two images. (a) has a much higher subject/background luminance contrast than (b).



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

Figure 4-23. Subject/background Michelson contrast. (a): high contrast image, with a feature value of 1; (b): subject mask of (a); (c): low contrast image, with a feature value of 0.764; (d): subject mask of (c).

The subject/background Michelson contrast is another example of a contrast measurement which can efficiently integrate the set of features. I introduce this feature by using a similar approach to (4.11), but this time considering the maximum luminance value of the subject and the lowest luminance value of the background.

$$f32 = \frac{L_{MaxSubj} - L_{MinBack}}{L_{MaxSubj} + L_{MinBack}}, \quad \textbf{(4.31)}$$

where $L_{MaxSubj}$ represents the subject maximum luminance value and $L_{MinBack}$ the background minimum luminance value. Expectedly, most of the images are going to achieve a value of 1, which means this feature can be efficient on identifying snapshots rather than detecting good quality photographs. Figure 4-23 exemplifies the described feature. (a), with a feature value of 1, has higher contrast than (b), which has a feature value of 0.764.

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

Figure 4-24. Subject/background RMS contrast features. (a): high distinction, with luminance RMS of 0.500, hue RMS of 0.354, and saturation RMS of 0.536; (b): subject mask of (a); (c): low distinction, with luminance RMS of 0.250, hue RMS of 0.087, and saturation RMS of 0.206; (d): subject mask of (c).

Similarly to the RMS features previously introduced, this measurement is computed for the luminance, *f33*, hue, *f34*, and saturation, *f35*, domains, concerning to the subject/background distinction, so that

$$f33,34,35 = \sqrt{\frac{1}{S_{area}}\sum_{s}\left(X(s) - X_{BackAvg}\right)^2}, \quad \textbf{(4.32)}$$

where *X* represents one of the luminance, hue or saturation images, $X_{BackAvg}$ the same component background average value, *s* the subject pixel coordinates and S_{area} the size of the subject area. In Figure 4-24, (a) has higher values than (b) for the described set of features. It means (a) has a higher subject/background variance in each of the analysed components.

4.6.2 Composition

The composition of a picture has high influence on the way the observer percepts aesthetics. However, this is a high-level concept which is difficult to evaluate.

The rule of thirds is one of the most used features among photo aesthetical assessment works [2] [14] [15] [21] [25] [26] [28]. In this project I include a novel way of determining the subject position by means of the Graph-based visual saliency map. With this map, it is possible to use the pixels' weight depending on their saliency and calculate the correspondent centre of mass. In order to increase the importance of high-saliency pixels, an exponential value of the saliency map is used as follows

$$total_{weighted_{saliency}} = \sum_i \sum_j S(i,j)^4, \quad \textbf{(4.33)}$$

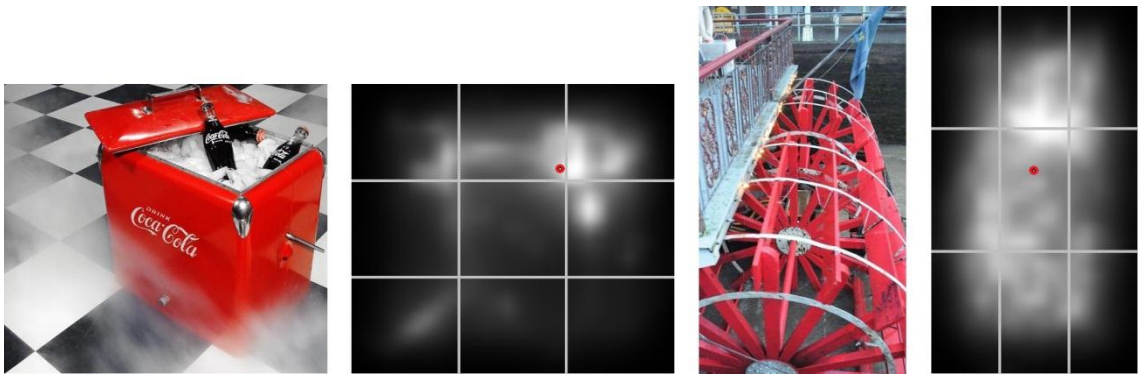
$$Subject_x = \frac{\sum_i \sum_j (iS(i,j)^4)}{total\_weighted\_saliency}, \quad \textbf{(4.34)}$$

$$Subject_y = \frac{\sum_i \sum_j (jS(i,j)^4)}{total\_weighted\_saliency}, \quad \textbf{(4.35)}$$

where *S* corresponds to the saliency map and *i* and *j* are the picture vertical and horizontal coordinates, respectively. The final feature reports the distance of the subject to the closest key-point defined by the rule of thirds.

$$f36 = min_{i=1,2,3,4}\left(\sqrt{\frac{(Subject_x - P_{ix})^2}{Image_{height}{}^2} + \frac{(Subject_y - P_{iy})^2}{Image_{width}{}^2}}\right), \quad \textbf{(4.36)}$$

where P_i represents each of the rule of thirds' key points, $Image_{height}$ the height of the image and $Image_{width}$ the image width. Figure 4-25 shows this feature applied to two different images. (a) has a better score than (c), since its subject is closer to one of the intersection points of the rule of thirds.



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

Figure 4-25. Example of the rule of thirds. (a): original high quality image; (b): feature applied on the saliency map of (a), with a relative distance to the closest intersection point of 0.044; (c): original low quality image; (d): feature applied on the saliency map of (c), with a relative distance to the closest intersection point of 0.200.

Another compositional characteristic which may be correlated with aesthetics is the subject size. In this case, I calculate the percentage of the image which is considered part of the subject, such that

$$f37 = \frac{S_{area}}{T_{area}}, \tag{4.37}$$

where S_{area} is the size of the subject area and T_{area} is the total image size. Figure 4-26 shows two examples of this feature usage.

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

Figure 4-26. Subject ratio feature. (a): image with subject ratio value of 0.146; (b): subject mask of (a); (c): image with subject ratio value of 0.037; (d): subject mask of (c).

4.6.3 Blur

It is possible to enhance the subject saliency by blurring the background areas. Actually, this is one of the most used techniques by photographers.

(a) (b) (c)

Figure 4-27. Low Laplacian contrast image. (a): original image with a subject/background Laplacian ratio of 0.753; (b): subject mask; (c): Laplacian image.

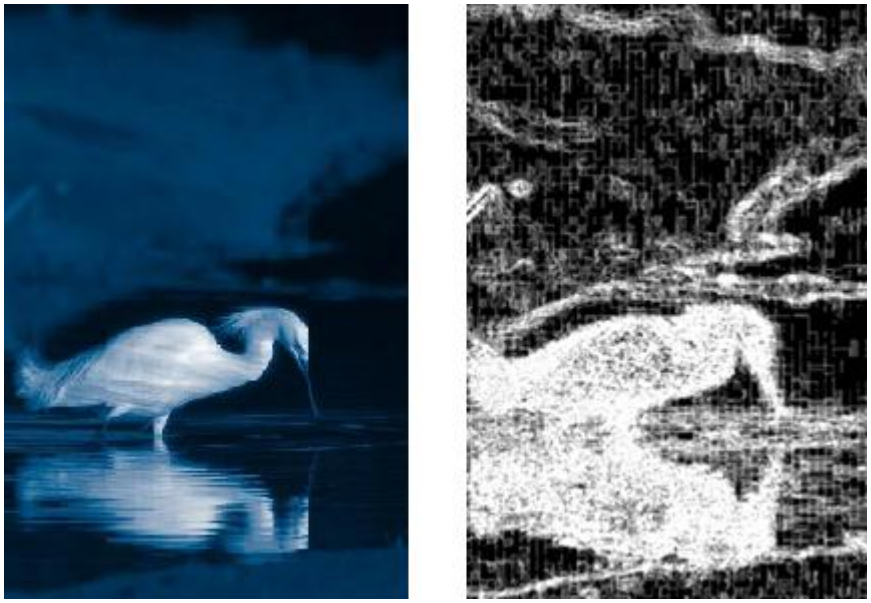
(a) (b) (c)

Figure 4-28. High Laplacian contrast image. (a): original image with a subject/background Laplacian ratio of 38.429; (b): subject mask; (c): Laplacian image.

A blurred area is known for having its details smoothed. This means it has low sharpness values. I decided to measure the difference in sharpness by means of the previously described Laplacian filter. By convolving an image with the Laplacian filter we obtain a sharpness map: the higher the pixel values, the more distinctive they are. This map can be used to calculate the average sharpness values for both the subject and the background, so we can relate these values as follows

$$f38 = \frac{\frac{1}{S_{area}}\sum_s L(s)}{\frac{1}{B_{area}}\sum_b L(b)}, \quad \textbf{(4.38)}$$

where *L* is the Laplacian image, *s* and *b* are respectively the subject and background pixels, and S_{area} and B_{area} are the subject and background areas, respectively. The highest the value of *f38*, the more enhanced is the image subject comparing to the background. Figure 4-28, represents a high Laplacian contrast image while Figure 4-27 exemplifies a low Laplacian contrast image. These are two practical examples of the use of *f38*.



(a) (b)

Figure 4-29. Probabilistic blur detection. (a): original image; (b): blur detection.

In [30], Meur et al. propose a depth of field measurement. The depth of field is the distance between the nearest and the farthest objects in a scene that appear sharp in an image. The feature introduced by [30] is obviously just an approximation to that distance. It is based on blurriness detection. I decided to use this feature on the subject/background comparison context. For that, three different blurring kernels of size $k \times k$, with $k = \{3,5,7\}$, are applied to the luminance of the original image. The vertical and horizontal derivatives histograms are then computed, such that

$$p_{xk} \propto hist(I * f_k * d_x), \tag{4.39}$$

$$p_{yk} \propto hist\big(I * f_k * d_y\big), \tag{4.40}$$

where *I* is the original image, f_k is the blurring kernel, $d_x = [1, -1]$ and $d_y = [1, -1]^T$. This two distributions are used to calculate the divergence with the original distribution, p_{x1}, for each image coordinate (i, j), so that

$$D_k(i,j) = \sum_{(m,n)\in W_{ij}} \Big(KL(p_{xk}|p_{x1})(m,n) + KL\big(p_{yk}|p_{y1}\big)(m,n)\Big), \tag{4.41}$$

where W_{ij} is a 3×3 window centered on the considered position and *KL* is the KL-divergence calculated by

$$KL(p|q)(i,j) = p_{ij} \log\left(\frac{p_{ij}}{q_{ij}}\right), \tag{4.42}$$

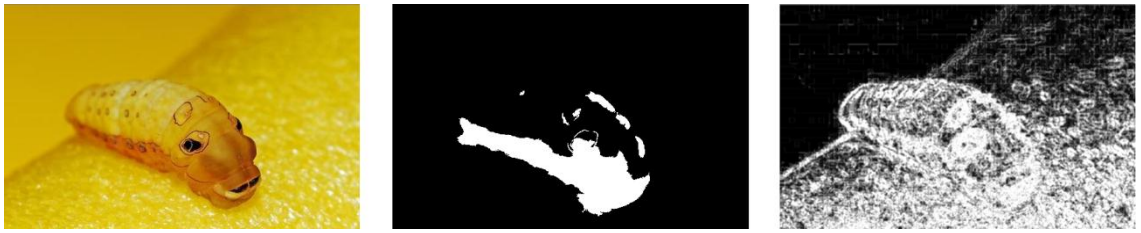
where *KL* measures how close *p* is to *q*. If p_{ij} and q_{ij} are exactly the same, their *KL* value will be 0 (mathematical indeterminations like $\log 0$ are treated as 0). This way, D_k reports the similarity of a region, before and after applying the $k \times k$ blurring kernel. If an area is originally blurred, its variation will not be very prominent. The final blurring value of a pixel is

$$B(i,j) = \sum_k D_k(i,j), \tag{4.43}$$

which after normalization will create an image similar to the one presented in Figure 4-29 (b). The final feature is the subject/ background relation concerning to the average *B* values, such that

$$f39 = \frac{\frac{1}{Subject_{area}} \sum_s B(s)}{\frac{1}{Background_{area}} \sum_b B(b)}, \tag{4.44}$$

where $Subject_{area}$ and $Background_{area}$ are respectively the subject and background areas, and similarly *s* and *b* are the subject and background pixels. Figure 4-30 and Figure 4-31 show the use of *f39*.



(a) (b) (c)

Figure 4-30. High Blur contrast image. (a): original image with a subject/background Blur ratio of 1.944; (b): subject mask; (c): Laplacian image.

(a) (b) (c)

Figure 4-31. Low Blur contrast image. (a): original image with a subject/background Blur ratio of 0.766; (b): subject mask; (c): Laplacian image.

5 LEARNING AND CLASSIFICATION TECHNIQUES

This chapter introduces the techniques used in the learning and classification phases. I start by describing the classifications methods. Then, some other used techniques are presented, in particular the ones related to feature selection.

5.1 Classification Methods

In this section, I will introduce the classification methods which will be used. The objective is to present a light overview of the selected techniques, so it is easier to understand my choices. The model will be trained on a labelled dataset, which means just supervised learning algorithms are selected.

The objective of machine learning methods is to create a model based on observed data patterns, so it is possible to take informed decisions on further examples.

5.1.1 Support Vector Machine

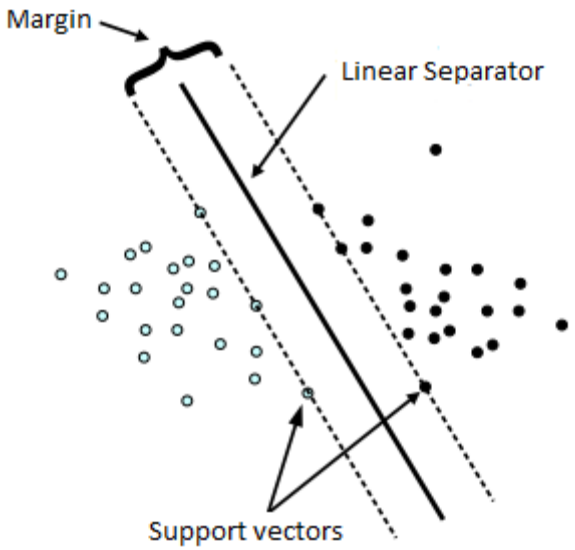
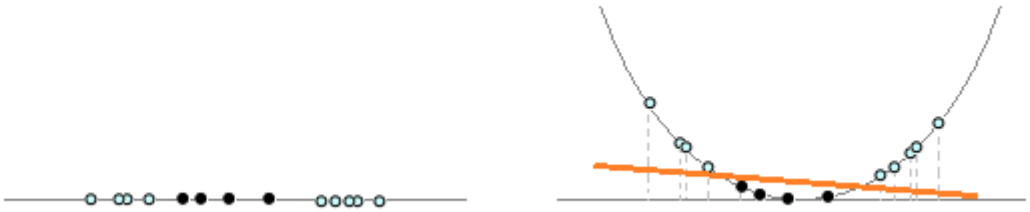


Figure 5-1. Linear separable data on SVM.

SVM is one of the most used machine learning techniques. It was introduced in the context of binary classification by Cortes et al. in [37]. The objective of this technique is to find a linear data separator. Figure 5-1 shows an example of

linear separable data in the 2D space. The support vectors, which are the points lying on the boundaries, are used to calculate the optimal linear separator, such that the margin is minimized.

However, in the real world it is normal to find data which is not linearly separable, which makes impossible the application of the previous approach. Although, there is a way of overcoming the problem: by projecting the data in a higher-dimensional space, the points can actually become separable. Figure 5-2 shows an example of a 1D to 2D projection, by means of a quadratic function. There are several auxiliary functions, or kernels, which can be used. Some examples are the linear, radial basis function, sigmoid, and polynomial kernels.



(a) (b)

Figure 5-2. Projection in higher dimension. (a): non-linearly separable space; (b): linearly separable space.

SVMs are known for being very robust to noise and less over-fitting prone comparing to other methods. However, it is computationally expensive, which means it will run slower than many other classifiers.

5.1.2 Gentle AdaBoost

Boosting is a technique which combines a large number of weak classifiers. A weak classifier is defined to be only slightly correlated with the true classification, having just above chance accuracy. In boosting, we keep training weak classifiers and adding them to the global model. In each iteration, it is attributed a weight to each classifier depending on its accuracy. The training data has also associated weights, so the misclassified examples keep gaining importance. Thus, future weak learners have a stronger focus on the

misclassified examples. During the classification phase, the model can use a number of techniques, like weighted voting. Figure 5-3 shows an example of the data flux in a boosting classifier: the input example is introduced to each weak classifier which output a classification value. The final system output is calculated by considering each of those values depending on their weights.

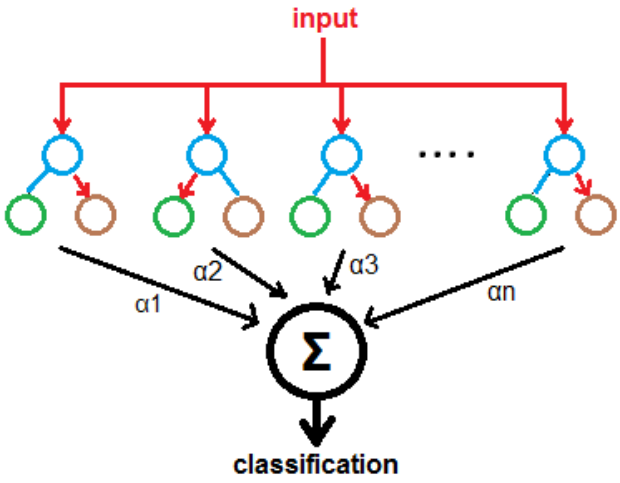


Figure 5-3. Boosting example.

Gentle AdaBoost is a boosting algorithm introduced in [38]. It has a particular way of weighting data points, since it puts less weight on outliers. This feature makes Gentle AdaBoost one of the most successful boosting algorithms. In this case, the weak classifiers are trees.

Gentle AdaBoost has a huge advantage over SVM: it can select informative features from a potential very large feature pool. This can be useful to this work, since we are in the presence of a fairly large feature set. However, AdaBoost is more sensible to noise than SVM.

5.1.3 Random Forest

Random forests [39] are somehow similar to boosting. They also combine several classifiers, which in this case are decision trees. The main difference is in the training phase, where, contrarily to boosting, the classifiers are grown independently from each other. This characteristic allows random forests to be less overfit prone. For the construction of the decision trees which constitute a random forest, it is just used a random subset of the total feature set. This

makes each tree to be a relatively weak classifier, and significantly reduces the computational complexity. This work addresses a binary classification problem. Thus, the final classification is determined by majority vote of the decision trees which are part of the forest.

5.2 Feature Normalization

It has been shown [40] that normalization is a pre-processing stage which can play an important role on machine learning algorithms performance. This characteristic is also mentioned in [41]. In order to take advantage of this property, I pre-processed the data by normalizing the feature set values, every time it is possible, so they can lie in the $[0,1]$ interval.

5.3 Feature Selection

Sometimes, it is not useful to have a large feature set. The existence of a large number of attributes can lead to high system complexity. By removing features, the dimensionality of the problem is reduced. It helps improving the performance of the learning models by speeding up the process, defying the curse of dimensionality[1] [42] or enhancing the generalization capability.

Several techniques can be used to reduce the feature space. The most popular approaches are probably the ones which use greedy strategies. This kind of algorithms is known for making the locally optimal choice at each stage. Although greedy strategies do not always find optimal solutions, they can approximate such a value in a reasonable amount of time.

In this project, I use a mix of forward and backward selection. Forward selection starts with an empty feature set and sequentially adds the one which maximizes the accuracy of the system. It can be described by the following steps:

1. Start with an empty set $Y_0 = \{\emptyset\}$

[1] The curse of dimensionality refers to the problem caused by high dimensional spaces, where data become easily sparse and dissimilar. This characteristic creates some difficulties to machine learning algorithms, since data patterns can be destroyed.

2. Select the next best non selected feature x^+ which maximizes the current feature set accuracy $x^+ = \arg max_{x \notin Y_k} [accuracy(Y_k + x)]$
3. Update the feature set $Y_{k+1} = Y_k + x^+; k = k + 1$
4. Go to 2

On the other hand, backward selection starts with the whole features included on the set. It sequentially removes the attribute which allows the new feature set to have the higher accuracy. It can be described as:

1. Start with the full set $Y_0 = X$
2. Select the existing feature $x^+ \in Y_k$ which maximizes the current feature set accuracy $x^- = \arg max_{x \in Y_k} [accuracy(Y_k - x)]$
3. Update the feature set $Y_{k+1} = Y_k - x^-; k = k + 1$
4. Go to 2

As previously mentioned, this work combines these two approaches in order to select a feature set which allows the system to have better performance. I introduce a new selection process. Firstly, forward selection is applied till the number of selected features is equal to the number we want to consider, *N*, plus one. The feature set size is excided so it is also possible to apply backward selection. Thus, by performing backward selection in two steps, the number of features is reduced to *N-1*. Again, forward selection is used, so we end up with the final feature set of size *N*. Forward selection is performed in first place because it is expected the feature set to be much smaller than the total number of features. If backward selection was used instead, the greedy solution could be much more distinct from the optimal solution, because of the required extra steps. On the other hand, performing backward selection in the end of the forward selection allows getting rid of some features which may not be relevant on the context anymore.

6 RESULTS

In this chapter I will report the results obtained by applying my algorithm to an existent dataset. This will allow measuring the performance of the overall system and compare it with other existing solutions.

6.1 Dataset

In order to have a base of comparison, I use the dataset which was originally created for evaluation of [11]. This dataset was chosen because it was used by a large number of other works, [4] [11] [15] [26] [31]. With this decision, I expect to contribute to a more uniform approach, allowing the scientific community to have a much stronger base of comparison.

The considered dataset was created by acquiring a set of images from DPChallenge.com, a photo contest website. In this website, users compete against each other by submitting their photographs, which are voted by the other website users in a scale from 1 to 10.

	Training	Testing
Professional photos	3000	3000
Snapshots	3000	3000

Table 6-1. Dataset composition in terms of number of photos.

From the DPChallenge website, 60000 photos, submitted by more than 40000 users, were retrieved. Each of the downloaded photos has been rated by at least 100 users, which minimizes the interference of possible outliers. The objective was to create two image subsets: one including good photographs examples and the other being composed by bad quality samples. Although, the subjectivity of the performing task makes the distinction quite difficult. In order to minimize dissensions, just the top and bottom 10% of the total photos were selected and assigned as high quality professional photographs and snapshots, respectively. Thus, each subset has 6000 images, building a total dataset of

12000 photos. From each set, half of the photos were assigned to the training set and the other set to the testing set. Table 6-1 shows such a distribution of photographs.

6.2 Classification Results

I start the classification process by using the whole feature set on the referred data set. For that, the classifiers described in section 5.1 are trained, which means it is necessary to select the corresponding parameters. In order to try to maximize the system accuracy, I perform grid search on a predefined parameter domain. The used grid search domains and steps are presented in Table 6-2, Table 6-3 and Table 6-4.

	weak count	weight trim rate	max depth
Domain	$[100,300[$	$[0.95,1[$	$[1,5[$
Step	10	0.1	1

Table 6-2. AdaBoost parameter grid search.

	SVM type	kernel	degree / γ / coef0 / C / ν / ε
Domain	$\{CSVC, \nu SVC\}$	{Linear, Polynomial, Radial Basis Function, $Sigmoid$}	[2]

Table 6-3. SVM parameter grid search.

	max depth
Domain	$[10,50[$
Step	1

Table 6-4. Random forest parameter grid search.

[2] Automatically generated by the OpenCV function *CvSVM::get_default_grid*.

In grid search, the objective is to find the parameters which achieve the highest accuracy scores on the considered data. Obviously, the results will possibly be adjusted to the noise contained in the used dataset, which can lead to some data overfitting. In order to efficiently evaluate the performance of the model, it is necessary to test the results in a completely different dataset comparing with the one used in the learning phase. For grid search, I apply a 10-fold cross validation technique to the training set, so the parameters can be chosen independently from the testing set. The data split of the total dataset is represented in Figure 6-1.

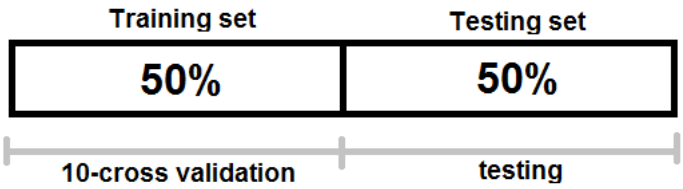


Figure 6-1. Data set usage on the parameter grid search.

In 10-fold cross validation, the considered set is randomly partitioned into 10 subsamples. Of the 10 subsamples, a single one is retained as the validation data while the remaining 9 are used to train the model. The cross-validation process is repeated 10 times, with each subsample being used exactly once as the validation data. In the end, I average the 10 results obtained from the folds, so a single estimation is produced. The 7th fold of the 10-fold cross validation process is represented in Figure 6-2.

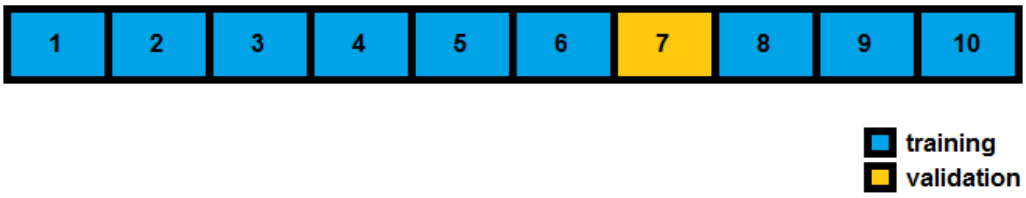


Figure 6-2. 7th fold of the 10-fold cross validation process

As previously mentioned, the objective of grid search is to maximize the calculated accuracy on the considered set. In this case, it will try to maximize the average accuracy returned from the 10-fold cross validation which

corresponds to a configuration of the parameters. For the considered classifiers, the following values are determined:

- AdaBoost
    - Parameter values
        - weak count: 110
        - weight trim rate: 0.98
        - max depth: 3
    - 10-fold cross validation accuracy: 91.5%
- SVM[3]
    - Parameter values
        - SVM type: C_SVC
        - kernel: Sigmoid
        - degree: 0
        - γ: 0.00015
        - coef0: 0.1
        - C: 62.5
        - ν: 0
        - ε: 0
    - 10-fold cross validation accuracy: 73.5%
- Random Forrest
    - Parameter values
        - max depth: 32
    - 10-fold cross validation accuracy: 88.9%

[3] By means of the OpenCV function *CvSVM::train_auto*

		Predicted class	
		Good	Bad
Actual class	Good	TP	FN
	Bad	FP	TN

Table 6-5. Confusion matrix fields.

In order to have an unbiased accuracy measure, it is necessary to test the selected parameters on the testing set, which was not used till now. The resulting confusion matrices are shown in Table 6-6, Table 6-7 and Table 6-8. The confusion matrices report the number of elements which are correct and the misclassified ones, for each considered class. For purposes of simplicity, I will refer to each of the field values of the confusion matrices by True Positives (TP), False Positives (FP), False Negatives (FN) and True Negatives (TN), as presented in Table 6-5.

		Predicted class	
		Good	Bad
Actual class	Good	2224	776
	Bad	567	2433

Table 6-6. Confusion matrix of the AdaBoost classifier applied to the whole feature set.

		Predicted class	
		Good	Bad
Actual class	Good	2068	932
	Bad	391	5609

Table 6-7. Confusion matrix of the SVM classifier applied to the whole feature set.

		Predicted class	
		Good	Bad
Actual class	Good	2466	534
	Bad	286	2714

Table 6-8. Confusion matrix of the Random Forest classifier applied to the whole feature set.

I decided to try another classifier which combines all of the previously described ones. For that, a voting strategy is implemented, where each of the classifiers report its classification. The final classification is the one which gathers more votes. Table 6-9 presents the confusion matrix of the combined classifier.

		Predicted class	
		Good	Bad
Actual class	Good	2242	758
	Bad	212	2788

Table 6-9. Confusion matrix of the combined classifiers applied to the whole feature set.

From the confusion matrices, it is possible to extract several relevant statistics. One of the most important is the accuracy, which measures the ratio of correct cases, and is described in (6.1). The accuracy measures of each considered classifier are presented in Table 6-10.

$$accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad \textbf{(6.1)}$$

	AdaBoost	SVM	Random Forest	Combination
Accuracy	77.6%	78.0%	86.3%	83.8%

Table 6-10. Accuracy of the different classifiers applied on the whole feature set.

As mentioned, there are other relevant statistics which can be extracted from the presented confusion matrices. The ones described in (6.2), (6.3), (6.4) and (6.5) are shown in A.2 (Appendix 2).

$$precision = \frac{TP}{TP + FP}, \quad \textbf{(6.2)}$$

$$recall = \frac{TP}{TP + FN}, \quad \textbf{(6.3)}$$

$$true\ negative\ rate = \frac{TN}{FP + TN}, \quad \textbf{(6.4)}$$

$$kappa = \frac{accuracy - 0.5}{0.5}. \quad \textbf{(6.5)}$$

In order to shoe the effectiveness of the features, a table with the accuracy of each of them is included in A.3. The correspondent chart is presented in A.4. The used parameters are those defined by the described grid search.

As mentioned in section 5.3, feature selection may reveal to be useful when trying to improve the overall performance of a system. In order to test this premise, the process described on the referred section is followed. I decided to find a feature set of size 10, since at that point almost all the features have a stable accuracy over 85%, as shown in A.5. As mentioned, the features are

chosen depending on the observed accuracy, after adding, on forward selection, or removing a feature, on backward selection. The accuracy cannot be measured on the testing set. Otherwise the trained model could overfit to the testing data, which would lead to doubtful results. Thus, it is again used 10-fold cross validation on the training set. It works exactly on the same way as previously described. Figure 6-1 and Figure 6-2 show an overview of the process. I decided to use the C-Support Vector classification with a linear kernel. This decision was based on the fact that most meaningful features should allow a linear separation between professional pictures and snapshots.

As described in section 5.3, the first step to take is performing forward selection. Such a technique is used till the feature set has size 10. In A.5, it is represented the table used on the forward selection for the first 10 steps. By the end of this run, the feature set $F = \{f1, f9, f11, f17, f22, f28, f30, f31, f32, f37\}$ is selected.

In order to determine the performance of the system at this stage, I perform the exact same steps as the ones used to test the whole feature set. By means of grid search, the following parameters are determined:

- AdaBoost
    - Parameter values
        - weak count: 160
        - weight trim rate: 0.98
        - max depth: 4
    - 10-fold cross validation accuracy: 90.3%
- SVM[4]
    - Parameter values
        - SVM type: NU_SVC
        - kernel: Sigmoid
        - degree: 0
        - γ: 0.03375

[4] By means of the OpenCV function *CvSVM::train_auto*

        - coef0: 0.1
        - C: 0
        - ν: 0.09
        - ε: 0
    - 10-fold cross validation accuracy: 36.9%
- Random Forrest
    - Parameter values
        - max depth: 22
    - 10-fold cross validation accuracy: 79.5%

The correspondent confusion matrices are presented in Table 6-11, Table 6-12 and Table 6-13. Table 6-14 shows the combined classifier previously described. The achieved accuracies are presented in Table 6-15. At this point, we can see there is a better overall performance. In addition, the problem complexity was severely reduced. A.6 contains further statistical information extracted from the confusion matrices, which can help us to understand the efficiency of the current system. A.7 and A.8 report the accuracy of each individual selected feature.

		Predicted class	
		Good	Bad
Actual class	Good	2017	983
	Bad	735	2265

Table 6-11. Confusion matrix of the AdaBoost classifier applied to the 10 feature set defined after forward selection.

		Predicted class	
		Good	Bad
Actual class	Good	2428	572
	Bad	29	2265

Table 6-12. Confusion matrix of the SVM classifier applied to the 10 feature set defined after forward selection.

		Predicted class	
		Good	Bad
Actual class	Good	2377	623
	Bad	254	2746

Table 6-13. Confusion matrix of the Random Forrest classifier applied to the 10 feature set defined after forward selection.

		Predicted class	
		Good	Bad
Actual class	Good	2433	567
	Bad	169	2831

Table 6-14. Confusion matrix of the combined classifier applied to the 10 feature set defined after forward selection.

	AdaBoost	SVM	Random Forest	Combination
Accuracy	71.4%	90.0%	85.4%	87.7%

Table 6-15. Accuracy of the different classifiers applied to the 10 feature set defined after forward selection.

As mentioned in 5.3, forward selection is a greedy strategy which locally searches for the best possible solution at each step. If we keep applying forward selection we may end up with a very inefficient solution. This kind of behaviour comes from the non-linear nature of the feature combination, which allows two good features to perform badly when combined. To minimize this characteristic, I apply a combination of forward and backward selection around the size 10. This action will remove features which lost their relevance during the forward selection process. Thus, forward selection is applied to the previously obtained feature set so an extra feature is chosen. 2 backward selection steps are then performed to the 11-sized feature set, reducing its dimensionality to 9. However, the objective is ending up with a 10-dimensonal feature set, which is achieved by using forward selection again. The final feature set is represented by $F = \{f1, f4, f9, f10, f11, f17, f22, f30, f31, f32\}$. The data process is presented in A.9.

The previously described grid search is again performed, so it is possible to select the learning parameters. The following values are found:

- AdaBoost
    - Parameter values
        - weak count: 250
        - weight trim rate: 0.96
        - max depth: 3
    - 10-fold cross validation accuracy: 89.9%
- SVM[5]
    - Parameter values

[5] By means of the OpenCV function *CvSVM::train_auto*

        - SVM type: NU_SVC
        - kernel: Sigmoid
        - degree: 0
        - γ: 0.00015
        - coef0: 0.1
        - C: 0
        - ν: 0.09
        - ε: 0
    - 10-fold cross validation accuracy: 39.9%
- Random Forrest
    - Parameter values
        - max depth: 24
    - 10-fold cross validation accuracy: 78.5%

The confusion matrices which show the performance of these classifiers are shown in Table 6-16, Table 6-17, Table 6-18 and Table 6-19. Each classifier's accuracy is presented in Table 6-20. Further statistical measures are shown in A.10. A.11 and A.12 present the accuracy of each individual feature.

		Predicted class	
		Good	Bad
Actual class	Good	2136	864
	Bad	562	2438

Table 6-16. Confusion matrix of the AdaBoost classifier applied to the 10 feature set defined after forward and backward selection.

		Predicted class	
		Good	Bad
Actual class	Good	2636	364
	Bad	313	2687

Table 6-17. Confusion matrix of the SVM classifier applied to the 10 feature set defined after forward and backward selection.

		Predicted class	
		Good	Bad
Actual class	Good	2498	502
	Bad	271	2729

Table 6-18. Confusion matrix of the Random Forrest classifier applied to the 10 feature set defined after forward and backward selection.

		Predicted class	
		Good	Bad
Actual class	Good	2611	389
	Bad	256	2744

Table 6-19. Confusion matrix of the combined classifier applied to the 10 feature set defined after forward and backward selection.

	AdaBoost	SVM	Random Forest	Combination
Accuracy	76.2%	88.7%	87.1%	89.3%

Table 6-20. Accuracy of the different classifiers applied to the 10 feature set defined after forward and backward selection.

7 DISCUSSIONS

In this chapter, I will discuss the results from the previous section and compare them with other authors' works.

The first obtained result was the performance of the classifiers using the full feature set. Analysing the data from A.2, it is easily seen that the Random Forest outperforms the other methods in terms of accuracy, with a value of 86.33% comparing with the 83.83% value performed by the combined classifier, which is the second best in this category. Also in terms of recall, Random Forest performs better than the combined classifier, with a value of 82.2% against 74.7%. It indicates this is a strong classifier when it comes to identifying the accurately reported professional photos among all the ones which belong to such a class. However, under certain constraints, the combined classifier may be considered the best choice. It reports a precision value of 91.4% comparing with the 89.6% value returned by the Random Forest, which can be valuable if our main concern is having the best possible ratio of photos correctly classified as professional among all the ones receiving that classification. The True Negative Rate indicator is also favourable to the combined classifier, which achieves 92.9% against the Random Forest value of 90.4%. This means the combined classifier is a better option when it comes to identifying the rate of snapshots which were correctly classified as so. In the considered feature set, both AdaBoost and SVM present a much poorer performance than the other classifiers.

This first algorithm run allows us to get an idea of how each feature performs. Although this analysis cannot be interpreted in a linear way, since the classifier parameters are optimized to the whole feature set, it is possible to get a taste of the developed features' efficiency. Data presented in A.3 and A.4 shows each feature's accuracy when independently run by the classifiers, which are optimized for the whole feature set. At first sight, it is possible to see that SVM reports a much higher standard deviation of the feature values. It produces features with both the highest and lowest accuracy. The other classifiers have much more balanced accuracy values, with Random Forest and combined

classifier slightly outperforming Gentle AdaBoost. Arguably, the best presented features are *f1* (75% energy Laplacian box), *f4* (hue count), *f11* (colour harmony), *f21* (subject colour simplicity), *f22* (subject hue count) and *f38* (subject/background Laplacian relation). Most of these features (*f1*, *f4*, *f22* and *f38*) are related to simplicity, which shows the importance of this aspect on the classification process. *f22* and *f38* were originally introduced in this paper; being part of the mentioned restricted feature set implies that they have a relevant role on image aesthetics assessment.

After performing forward selection, as shown in A.5, the feature set got reduced to 10 attributes: *f1* (75% energy Laplacian box), *f9* (average brightness), *f11* (colour harmony), *f17* (subject luminance RMS), *f22* (background hue count), *f28* (subject/background brightness squared difference), *f30* (subject/ background saturation squared difference), *f31* (subject/background Weber contrast), *f32* (subject/background Michelson contrast) and *f37* (subject relative size). The individual feature performance, presented in A.7, shows that *f1*, *f11* and *f22* have a much higher accuracy than the other attributes. All these three features were part of the high accuracy set when using the whole features.

In terms of performance of the overall system with the reduced feature set, from A.6, we can observe a notorious improvement. The SVM classifier achieves an accuracy of 90.0%, outperforming the other classifiers in every measurement but recall, where the combined classifier performs slightly better, with 81.1% against 80.9% from the SVM. However, all the other statistics give advantage to the SVM classifier, mainly when it comes to precision and TNR where it achieves a round number of 99%.

The combined forward and backward selection led to a new set of features, as shown in A.9. This process substituted *f28* and *f37* by *f4* and *f10*. A.11 and A.12 show us that *f1*, *f4*, *f11* and *f22* are the most distinctive features. Once again, these features are all part of the best set determined while using the whole attributes. In general terms, the accuracies seem to be higher comparing with the ones calculated after the forward selection phase. Although there are

shared features between these two steps, the accuracy values are different because of the different parameters of the classifiers.

In terms of accuracy, we assist to an overall improvement of the results, as presented in A.10. Excluding SVM, all the classifiers improve their general performance. In this case, the combined classifier achieved a value of 89.3%, achieving the best accuracy among the considered methods. However, it is not enough to perform better than the SVM used after forward selection. Figure 7-1 shows an overview of the accuracies achieved by the different methods under the three tested conditions. There is a notorious overall improvement through the process, despite the system having its best performance in the SVM classifier after forward selection.

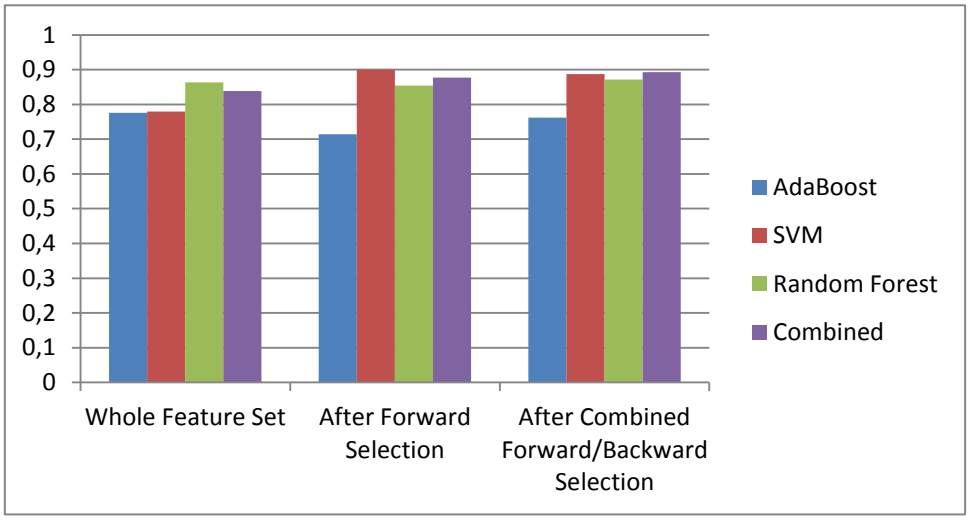


Figure 7-1. System accuracy.

Concerning to the other measurements, we can observe different behaviours. As shown in Figure 7-3, the recall values follow an increasing pattern. Both SVM and combined classifier, applied after forward and background selection, perform over 87%. However, both precision and TNR values decrease after applying the combined forward and background selection. Precision presents a higher value of 91.1% using the combined classifier, which is way lower than the 98.8% achieved by the SVM classifier after forward selection. The TNR has its higher value of 91.5%, using the combined classifier, which is much lower than the SVM value after forward selection, which presented a value of 99.0%.

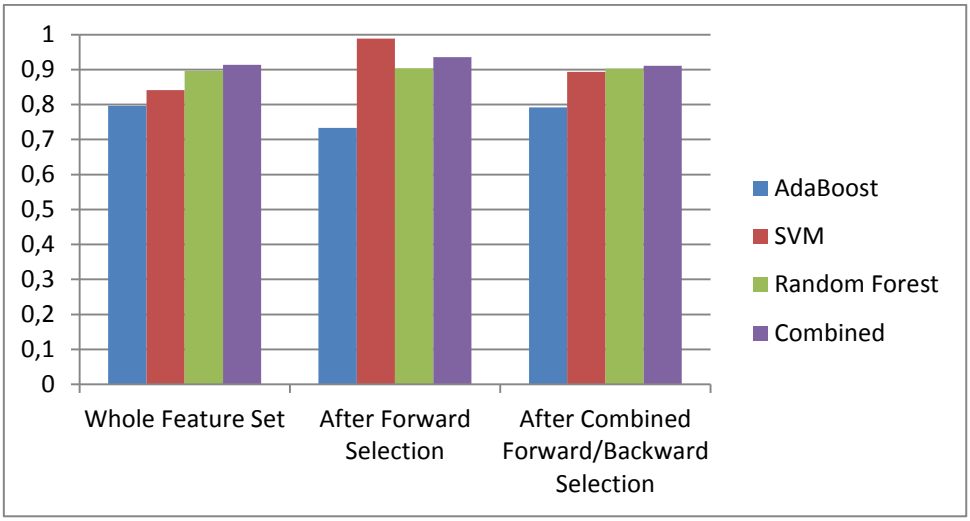


Figure 7-2. System Precision.

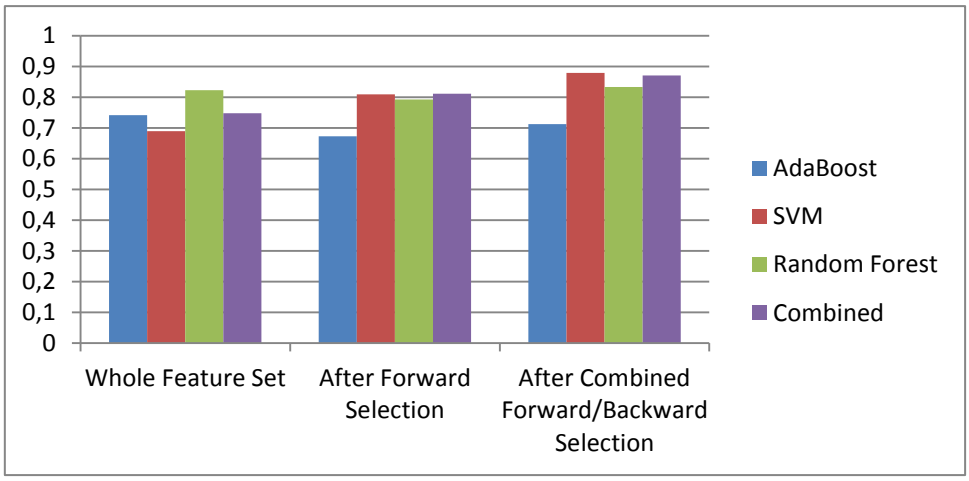


Figure 7-3. System Recall.

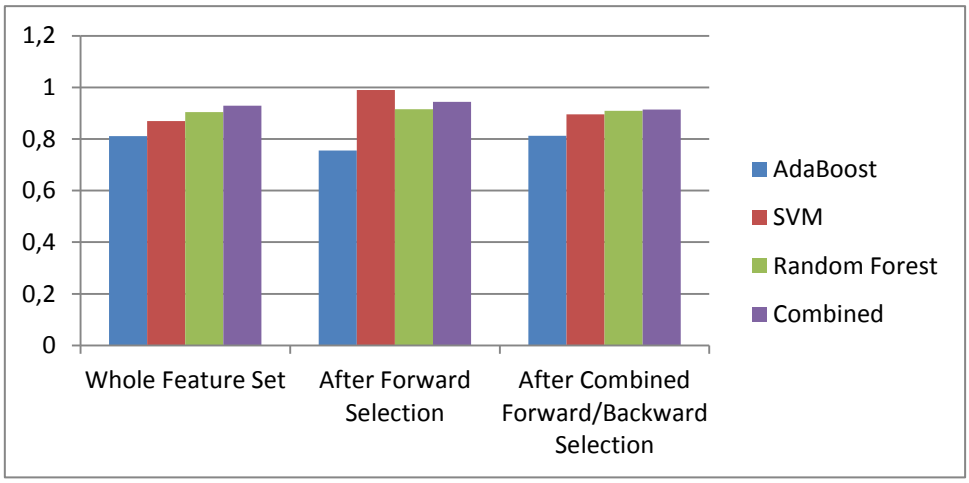


Figure 7-4. System TNR.

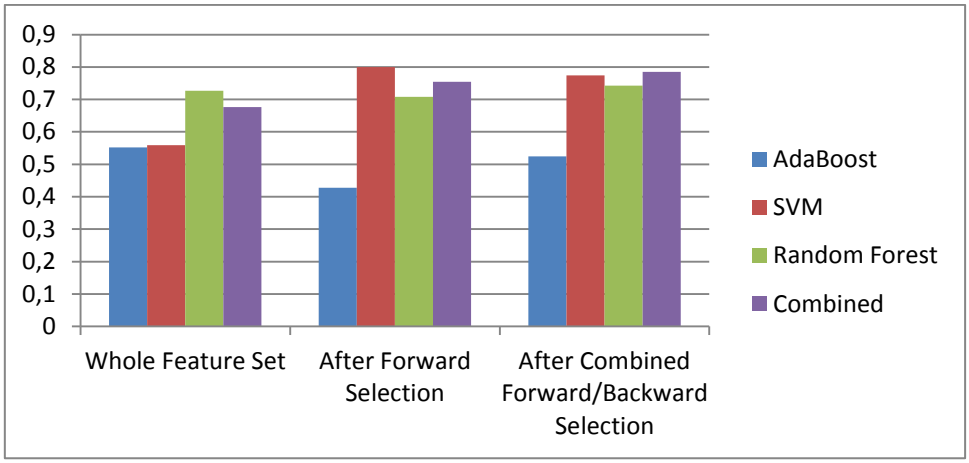


Figure 7-5. System Kappa Statistic.

This analysis shows that different classifiers should be used for different tasks. If the objective is accurately classifying the highest number of images independently of the considered class, then we may prefer using SVM after forward selection. However, if it is preferable to report the highest number of professional photos, it is better to use a classifier with higher precision, like SVM after combined forward and backward selection. If we want a more balanced system, we may want to use the combined classifier after combined forward and backward selection.

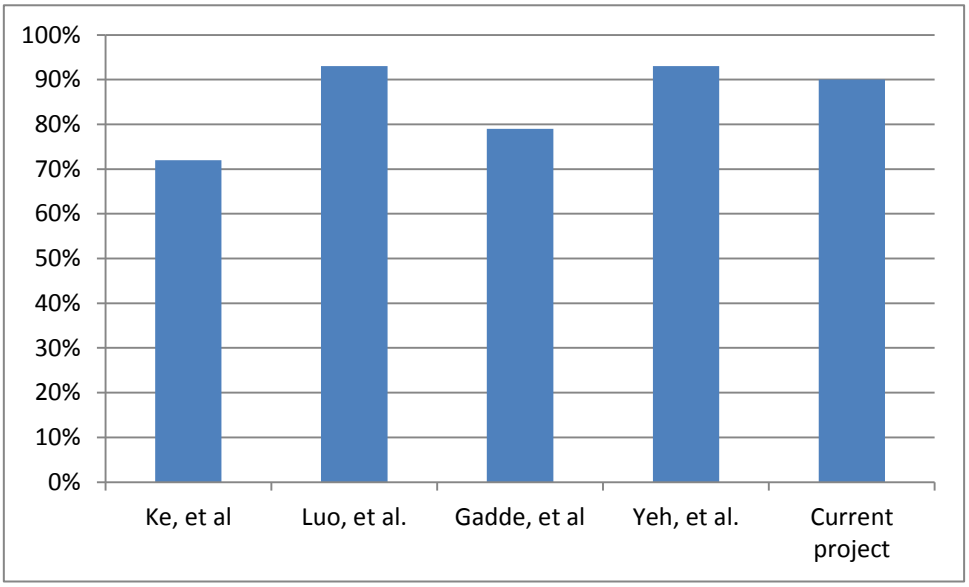


Figure 7-6. Comparison of the different algorithms which use the same dataset.

It is important to compare the developed classifiers with the existing related works. This step is extremely difficult since all the environments seem to be different. There are lots of variables involved which mean we cannot directly compare all the algorithms. In order to reduce a variable, I just considered works which share the dataset used in this project: Ke et al. [11], Luo et al. [15], Gadde et al. [4] and Yeh et al. [26]. Figure 7-6 shows a comparison of the accuracy achieved by those algorithms and the SVM classifier which presented the highest accuracy in this project. In this set, my classifier report the 3rd best results, right behind the work developed by Luo et al. and Yeh et al., which achieve an overall accuracy of 93%. However, there are other variables which interfere with the final system score. As shown during this process, parameter selection may have a huge role on the final accuracy. On the other works, there are no evidences of the way the authors performed such a step. If the parameters are selected using the testing set accuracy, the final results will probably be overfitted to the data. This property will lead to a much higher reported accuracy than the one which will be observed in unseen data.

There is also another relevant comparison: the precision-recall relation. In this project, the SVM classifier achieved precision and recall values of 99% and 81%, respectively. Excepting Gadde et al., all other authors present information of their precision-recall values. Those numbers are shown in Table 7-1 and Table 7-2. In this particular domain, the algorithm developed in this project clearly outperforms the other implementations.

	Recall
Ke et al.	<1%
Luo et al.	16%
Yeh et al.	16%
Current project	81%

Table 7-1. Recall values when precision = 99%.

	Precision
Ke et al.	65%
Luo et al.	86%
Yeh et al.	79%
Current project	99%

Table 7-2. Precision values when recall = 81%.

8 CONCLUSIONS

In this thesis, I developed an aesthetics-driven system for classifying images as professional photos or snapshots. The project is based on several works which address the same problem. However, as an open research field, there is always room for improvements. I expanded the used features by introducing a few new ones. The originally introduced *f22* (background hue count), *f31* (subject/background Weber contrast relation), *f32* (subject/background Michelson contrast relation) and *f38* (subject/background Laplacian relation) are among the set of features with best accuracy. Nevertheless, the main subject of study was the learning phase. In this project, I introduced a classifier which was never tried in any of the related works: Random Forests. Although not achieving the best possible classification accuracy, this classifier has proved to be robust enough to rival with the most widely used techniques. It constantly beat AdaBoost which is a common choice among the scientific community. In addition, a combined classifier was created, by using a voting strategy taking into account the three considered classifiers. In some cases, this technique outperforms all the other classification methods. Lastly, I put a lot of effort in building a clear process which follows the rules of a good machine learning method. The first important step was selecting a widely used dataset, since it is crucial to have a similar base of comparison. Then, I developed attribute and parameter selection techniques using grid search and forward/backward selection. In these methods, the testing dataset is never used, so it is possible to avoid overfitting.

The results of the project show that this system has slightly lower accuracy than the state of the art algorithms. However, it outperforms the other solutions in terms of precision and recall values. In addition, all the details were described in this document, so one can easily reproduce the work. I experienced some difficulties trying to replicate some other authors' implementations. With this project, I tried to make the whole process clear enough so anyone can achieve the same results. Thus, I can say the project goals were achieved.

Despite the promising result of this work, I believe there are some ways of further improving the solution. Subject extraction relies on saliency maps and image segmentation. Both of them are open problems with results which are not completely related with humans' perception. It is possible that the extracted subject region does not match the actual subject, decreasing the overall classification performance. Refining the subject extraction algorithm would certainly improve the system accuracy. It would also be useful trying other types of classifiers and selection techniques. Most authors are mainly concerned with the feature set. However, in order to extract the best performance from the features, it is necessary to have a strong classifier. Data patterns are often complex and the application of relevant machine learning techniques may heavily influence the system's performance. Finally, applying the system on a real application would allow understanding its real impact.

REFERENCES

[1] M. W. Sandler, Photography: An Illustrated Historical Overview, New York, USA: Oxford University Press, 2002.

[2] S. Bhattacharya, R. Sukthankar and M. Shah, “A Framework for Photo-Quality Assessment and Enhancement based on Visual Aesthetics,” in *Proceedings of ACM Multimedia Conference*, Firenze, Italy, 2010.

[3] S. Bhattacharya, R. Sukthankar and M. Shah, “A Holistic Approach to Aesthetic Enhancement,” *ACM Transactions on Multimedia Computing, Communications, and Applications,* vol. 7S, no. 1, 2011.

[4] R. Gadde and K. Karlapalem, “Aesthetic Guideline Driven Photography by Robots,” in *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence*, Barcelona, Spain, 2011.

[5] H. Tong, M. Li, H. Zhang, J. He and C. Zhang, “Classification of Digital Photos Taken by Photographers or Home Users,” in *Proceedings of the 5th Pacific Rim conference on Advances in Multimedia Information Processing*, Tokyo, Japan, 2004.

[6] N. Serrano, A. Svakis and J. Luo, “A Computationally Efficient Approach to Indoor/Outdoor Scene Classification,” in *Proceedings of the 16th International Conference on Pattern Recognition*, Quebec City, Canada, 2002.

[7] C. Oliveira, A. Araujo, C. J. Severiano and D. Gomes, “Classifying Images Collected on the World Wide Web,” in *Proceedings of the 15th Brazilian Symposium on Computer Graphics and Image Processing*, Fortaleza, Brazil, 2002.

[8] A. Vailaya, A. Jain and H. J. Zhang, “On Image Classification: City vs. Landscape,” in *Proceedings of the IEEE Workshop on Content-Based*

Access of Image and Video Libraries, Bombay, India, 1998.

[9] H. Tong, M. Li, H. Zhang and C. Zhang, “Blur Detection for Digital Images Using Wavelet Transform,” in *Proceedings of the IEEE International Conference on Multimedia and Expo*, Taipei, Taiwan, 2004.

[10] Y. Ma, L. Lu, H. Zhang and M. Li, “A User Attention Model for Video Summarization,” in *Proceedings of the 10th ACM international conference on Multimedia* , San Francisco, USA, 2002.

[11] Y. Ke, X. Tang and F. Jing, “The Design of High-Level Features for Photo Quality Assessment,” in *Proceedings of Computer Vision and Pattern Recognition*, New York, USA, 2006.

[12] L. Frost, The A-Z of Creative Photography, New York, USA: Amphoto Books, 1998.

[13] B. Peterson, Learning to See Creatively, New York, USA: Amphoto Books, 2003.

[14] R. Datta, D. Joshi, J. Li and J. Wang, “Studying Aesthetics in Photographic Images Using a Computational Approach,” in *Proceedings of the European Conference on Computer Vision*, Graz, Austria, 2006.

[15] Y. Luo and X. Tang, “Photo and Video Quality Evaluation: Focusing on the Subject,” in *Proceedings of the 10th European Conference on Computer Vision*, Marseille, France, 2008.

[16] A. Levin, “Blind motion deblurring using image statistics,” in *Proceedings of the 20th Annual Conference on Neural Information Processing Systems*, Vancouver, Canada, 2006.

[17] L. Itti, C. Koch and E. Niebur, “A Model of Saliency-Based Visual Attention,” in *Proceedings of the IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1998.

[18] J. Harel, C. Koch and P. Perona, “Graph-Based Visual Saliency,” in *Proceedings of the 20th Annual Conference on Neural Information Processing Systems*, Vancouver, Canada, 2006.

[19] M. Cheng, G. Zhang, N. J. Mitra, X. Huang and S. Hu, “Global Contrast based Salient Region Detection,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Colorado Springs, USA, 2011.

[20] P. F. Felzenszwalb and D. P. Huttenlocher, “Efficient Graph-Based Image Segmentation,” *International Journal of Computer Vision,* vol. 59, no. 2, pp. 167-181, 2004.

[21] L. Wong and K. Low, “Saliency-enhanced Image Aesthetics Class Prediction,” in *Proceedings of the IEEE International Conference on Image Processing*, Cairo, Egypt, 2009.

[22] X. Sun, H. Yao, R. Ji and S. Liu, “Photo Assessment based on Computational Visual Attention Model,” in *Proceedings of the 17th ACM international conference on Multimedia*, Beijing, China, 2009.

[23] B. Cheng, B. Ni, S. Yan and Q. Tian, “Learning to Photograph,” in *Proceedings of ACM Multimedia Conference*, Firenze, Italy, 2010.

[24] M. Desnoyer and D. Wettergreen, “Aesthetic Image Classification for Autonomous Agents,” in *Proceedings of the 20th International Conference on Pattern Recognition*, Istanbul, Turkey, 2010.

[25] C. Li, A. Gallagher, A. C. Loui and T. Chen, “Aesthetic Quality Assessment of Consumer Photos with Faces,” in *Proceedings of IEEE 17th International Conference on Image Processing*, Hong Kong, China, 2010.

[26] C. Yeh, Y. Ho, B. A. Barski and M. Ouhyoung, “Personalized Photograph Ranking and Selection System,” in *Proceedings of ACM Multimedia Conference*, Firenze, Italy, 2010.

[27] L. Huang, T. Xia, J. Wan, Y. Zhang and S. Lin, “Personalized Portraits Ranking,” in *Proceedings of the 19th ACM international conference on Multimedia* , Scottsdale, USA, 2011.

[28] S. Dhar, V. Ordonez and T. Berg, “High Level Describable Attributes for Predicting Aesthetics and Interestingness,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Colorado Springs, USA, 2011.

[29] B. Geng, L. Yang, C. Xu, X. Hua and S. Li, “The Role of Attractiveness in Web Image Search,” in *Proceedings of the 19th ACM international conference on Multimedia*, Scottsdale, USA, 2011.

[30] O. Meur, T. Baccino and A. Roumy, “Prediction of the Inter-Observer Visual Congruency (IOVC) and Application to Image Ranking,” in *Proceedings of the 19th ACM international conference on Multimedia*, Scottsdale, USA, 2011.

[31] H. Su, T. Chen, C. Kao, W. Hsu and S. Chien, “Scenic Photo Quality Assessment with Bag of Aesthetics-Preserving Features,” in *Proceedings of the 19th ACM international conference on Multimedia*, Scottsdale, USA, 2011.

[32] P. Isola, D. Parikh, A. Torralba and A. Oliva, “What makes an image memorable?,” in *Proceedings of 24th IEEE Conference on Computer Vision and Pattern Recognition*, Colorado Springs, USA, 2011.

[33] P. Isola, D. Parikh, A. Torralba and A. Oliva, “Understanding the Intrinsic Memorability of Images,” in *Proceedings of the 25th Annual Conference on Neural Information Processing Systems*, Granada, Spain, 2011.

[34] A. K. Moorthy and A. C. Bovik, “Visual Quality Assessment Algorithms : What Does the Future Hold?,” *Multimedia Tools and Applications,* vol. 51, no. 2, pp. 675-696 , 2011.

[35] L. Langford and P. Andrews, Langford's Starting Photography, 6th edition, Oxford, UK: Focal Press, 2009.

[36] R. Achanta, S. Hemami, F. Estrada and S. Susstrunk, “Frequency-tuned Salient Region Detection,” in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, Miami, USA, 2009.

[37] C. Cortes and V. Vapnik, “Support-Vector Networks,” *Machine Learning,* vol. 20, pp. 273-297, 1995.

[38] A. Torralba, K. P. Murphy and W. T. Freeman, “Sharing features: efficient boosting procedures for multiclass object detection,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Washington, DC, USA, 2004.

[39] L. Breiman, “Random Forests,” *Machine Learning,* vol. 45, pp. 5-32, 2001.

[40] R. Herbrich and T. Graepel, “A PAC-Bayesian Margin Bound for Linear Classifiers: Why SVMs work,” in *Proceedings of the 15th Annual Conference on Neural Information Processing Systems*, Vancouver, Canada, 2001.

[41] A. B. A. Graf and S. Borer, “Normalization in Support Vector Machines,” in *Proceedings of the DAGM Symposium for Pattern Recognition*, Munich, Germany, 2001.

[42] I. Guyon and A. Elisseeff, “An Introduction to Variable and Feature Selection,” *The Journal of Machine Learning Research,* vol. 3, pp. 1157-1182, 2003.

[43] R. Datta, J. Li and J. Wang, “Algorithmic Inferencing of Aesthetics and Emotion in Natural Images: An Exposition,” in *Proceedings of the IEEE International Conference on Image Processing*, San Diego, USA, 2008.

APPENDICES

A.1 Feature Names

Global Features	
f1	75% energy Laplacian box
f2	Average saliency
f3	Saliency standard deviation
f4	Hue count
f5	Number of meanshift segments
f6	98% luminance histogram mass width
f7	Weber contrast
f8	Michelson contrast
f9	Average brightness
f10	Average saturation
f11	Colour harmony
f12	FFT blur estimation
f13	Aspect ratio
f14	Intensity balance
Subject Features	
f15	Average brightness
f16	Michelson contrast
f17	Luminance RMS
f18	Hue RMS
f19	Saturation RMS

f20	Average saturation
Background Features	
f21	Colour simplicity
f22	Hue count
f23	Michelson contrast
f24	Luminance RMS
f25	Hue RMS
f26	Saturation RMS
Subject/background Relation Features	
f27	Lightning ratio
f28	Brightness squared difference
f29	Hue squared difference
f30	Saturation squared difference
f31	Weber contrast
f32	Michelson contrast
f33	Luminance RMS
f34	Hue RMS
f35	Saturation RMS
f36	Rule of thirds
f37	Subject relative size
f38	Laplacian relation
f39	Blur relation

A.2 Statistical Information of the Whole Feature Set

	Accuracy	Precision	Recall	TNR	Kappa
AdaBoost	0,776167	0,796847	0,741333	0,811	0,552333
SVM	0,7795	0,840992	0,689333	0,869667	0,559
Random Forest	0,863333	0,896076	0,822	0,904667	0,726667
Combined	0,838333	0,91361	0,747333	0,929333	0,676667

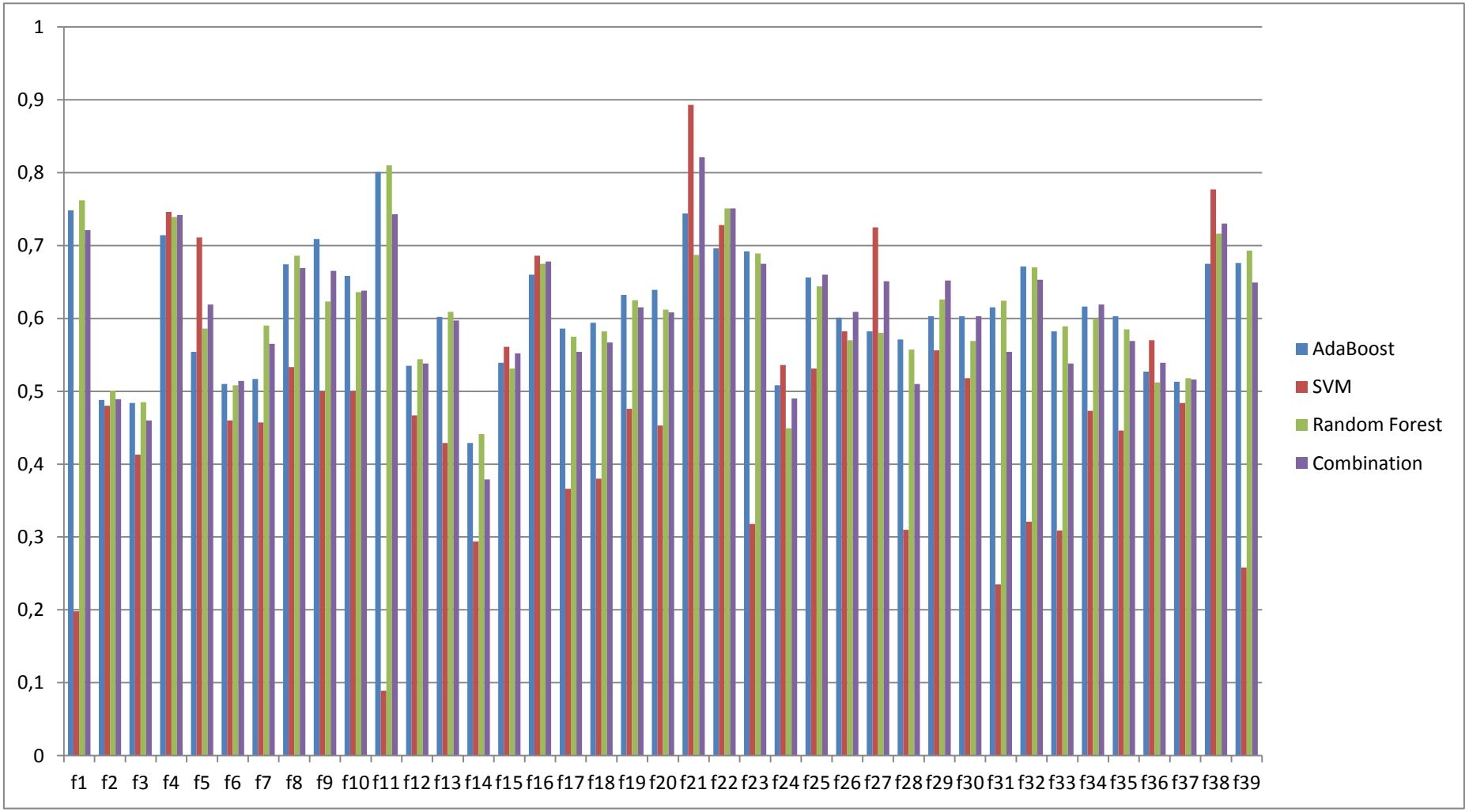
A.3 Feature Accuracy Using the Whole Feature Set Parameters: Table

Feature	AdaBoost	SVM	Random Forest	Combined
f1	0,748	0,198	0,762	0,721
f2	0,488	0,48	0,501	0,489
f3	0,484	0,413	0,485	0,46
f4	0,714	0,746	0,739	0,742
f5	0,554	0,711	0,586	0,619
f6	0,51	0,46	0,508	0,514
f7	0,517	0,457	0,59	0,565
f8	0,674	0,533	0,686	0,669
f9	0,709	0,5	0,623	0,665
f10	0,658	0,5	0,636	0,638
f11	0,801	0,089	0,81	0,743
f12	0,535	0,467	0,544	0,538
f13	0,602	0,429	0,609	0,597
f14	0,429	0,294	0,441	0,379
f15	0,539	0,561	0,531	0,552
f16	0,66	0,686	0,675	0,678
f17	0,586	0,366	0,575	0,554
f18	0,594	0,38	0,582	0,567
f19	0,632	0,476	0,625	0,615

f20	0,639	0,453	0,612	0,608
f21	0,744	0,893	0,687	0,821
f22	0,696	0,728	0,751	0,751
f23	0,692	0,318	0,689	0,675
f24	0,508	0,536	0,449	0,49
f25	0,656	0,531	0,644	0,66
f26	0,601	0,582	0,57	0,609
f27	0,582	0,725	0,58	0,651
f28	0,571	0,31	0,557	0,51
f29	0,603	0,556	0,626	0,652
f30	0,603	0,518	0,569	0,603
f31	0,615	0,235	0,624	0,554
f32	0,671	0,321	0,67	0,653
f33	0,582	0,309	0,589	0,538
f34	0,616	0,473	0,6	0,619
f35	0,603	0,446	0,585	0,569
f36	0,527	0,57	0,512	0,539
f37	0,513	0,484	0,518	0,516
f38	0,675	0,777	0,716	0,73
f39	0,676	0,258	0,693	0,649

Green cells report accuracy over 70%.

A.4 Feature Accuracy Using the Whole Feature Set Parameters: Chart



A.5 Feature Selection: Forward Selection Phase

Feature	Step1	Step2	Step3	Step4	Step5	Step6	Step7	Step8	Step9	Step10
f1	0,859									
f2	0,51	0,856	0,612	0,854	0,851	0,853	0,852	0,638	0,851	0,857
f3	0,444	0,85	0,852	0,854	0,855	0,855	0,854	0,853	0,858	0,857
f4	0,54	0,649	0,854	0,852	0,856	0,672	0,528	0,853	0,857	0,858
f5	0,853	0,495	0,853	0,853	0,855	0,855	0,854	0,854	0,854	0,851
f6	0,491	0,843	0,855	0,551	0,854	0,851	0,853	0,643	0,853	0,85
f7	0,505	0,853	0,854	0,853	0,857	0,854	0,852	0,856	0,855	0,855
f8	0,852	0,855	0,851	0,851	0,854	0,854	0,857	0,853	0,855	0,856
f9	0,851	0,619	0,849	0,83	0,854	0,852	0,861			
f10	0,851	0,853	0,855	0,852	0,699	0,855	0,854	0,858	0,858	0,852
f11	0,55	0,855	0,855	0,852	0,653	0,853	0,852	0,851	0,863	
f12	0,852	0,855	0,539	0,853	0,859	0,853	0,853	0,575	0,852	0,853
f13	0,44	0,779	0,854	0,856	0,852	0,664	0,765	0,854	0,855	0,855
f14	0,615	0,854	0,85	0,598	0,854	0,854	0,851	0,854	0,854	0,854
f15	0,522	0,856	0,854	0,856	0,853	0,855	0,85	0,856	0,598	0,853
f16	0,851	0,855	0,854	0,849	0,854	0,854	0,853	0,856	0,853	0,853
f17	0,854	0,715	0,853	0,859						
f18	0,581	0,467	0,854	0,853	0,855	0,746	0,58	0,855	0,606	0,855
f19	0,857	0,64	0,5	0,854	0,728	0,855	0,853	0,85	0,856	0,853
f20	0,512	0,835	0,858	0,828	0,831	0,853	0,851	0,855	0,854	0,855
f21	0,783	0,855	0,62	0,855	0,852	0,758	0,856	0,85	0,855	0,846
f22	0,852	0,857								
f23	0,853	0,854	0,853	0,48	0,857	0,543	0,853	0,857	0,854	0,856
f24	0,855	0,523	0,852	0,551	0,854	0,542	0,854	0,854	0,852	0,854
f25	0,555	0,853	0,856	0,853	0,852	0,679	0,852	0,854	0,857	0,854
f26	0,51	0,383	0,852	0,853	0,853	0,854	0,854	0,755	0,856	0,852
f27	0,854	0,853	0,855	0,852	0,853	0,856	0,851	0,777	0,853	0,853
f28	0,851	0,855	0,852	0,854	0,542	0,853	0,851	0,859		
f29	0,495	0,854	0,852	0,848	0,852	0,538	0,855	0,856	0,851	0,854
f30	0,855	0,843	0,849	0,73	0,863					
f31	0,852	0,438	0,483	0,855	0,855	0,86				
f32	0,854	0,853	0,858							
f33	0,622	0,855	0,854	0,549	0,853	0,553	0,853	0,853	0,855	0,842
f34	0,516	0,854	0,853	0,852	0,853	0,854	0,711	0,854	0,854	0,855
f35	0,501	0,854	0,857	0,854	0,854	0,858	0,854	0,854	0,852	0,766
f36	0,856	0,854	0,854	0,853	0,854	0,85	0,856	0,855	0,855	0,694
f37	0,563	0,85	0,854	0,853	0,748	0,854	0,852	0,766	0,608	0,863
f38	0,832	0,59	0,529	0,44	0,78	0,479	0,655	0,682	0,551	0,658
f39	0,687	0,613	0,506	0,812	0,556	0,679	0,688	0,5	0,525	0,621

Green cells correspond to the selected features. The feature picked at a certain step is maintained through the whole run.

A.6 Statistical Information of the Feature Set after Forward Selection

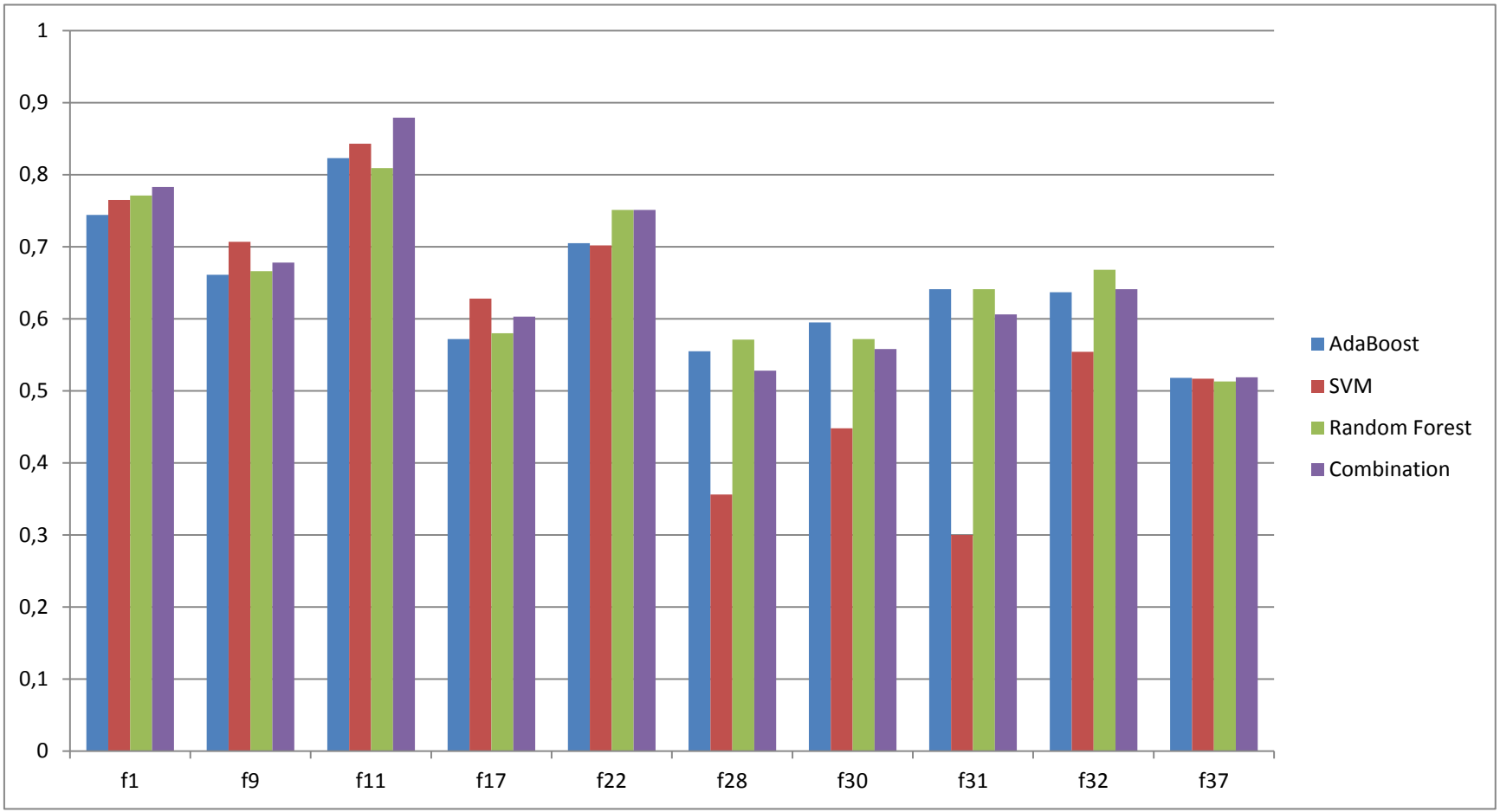
	Accuracy	Precision	Recall	TNR	Kappa
AdaBoost	0,713667	0,732922	0,672333	0,755	0,427333
SVM	0,899833	0,988197	0,809333	0,990333	0,799667
Random Forest	0,853833	0,903459	0,792333	0,915333	0,707667
Combined	0,877333	0,93505	0,811	0,943667	0,754667

A.7 Feature Accuracy after Forward Selection: Table

Feature	AdaBoost	SVM	Forest	All
f1	0,744	0,765	0,771	0,783
f9	0,661	0,707	0,666	0,678
f11	0,823	0,843	0,809	0,879
f17	0,572	0,628	0,58	0,603
f22	0,705	0,702	0,751	0,751
f28	0,555	0,356	0,571	0,528
f30	0,595	0,448	0,572	0,558
f31	0,641	0,3	0,641	0,606
f32	0,637	0,554	0,668	0,641
f37	0,518	0,517	0,513	0,519

Green cells report accuracy over 70%.

A.8 Feature Accuracy after Forward Selection: Chart



A.9 Feature Selection: Combined Forward and Backward Selection Phase

Feature	Forward1	Backward2	Backward3	Forward4
f1		0,857	0,855	
f2	0,849			0,857
f3	0,72			0,716
f4	0,858	0,854	0,854	
f5	0,854			0,851
f6	0,852			0,853
f7	0,853			0,852
f8	0,853			0,857
f9		0,854	0,853	
f10	0,852			0,859
f11		0,853	0,857	
f12	0,854			0,851
f13	0,767			0,854
f14	0,851			0,854
f15	0,855			0,854
f16	0,854			0,854
f17		0,854	0,852	
f18	0,853			0,854
f19	0,854			0,679
f20	0,644			0,854
f21	0,853			0,855
f22		0,854	0,853	

f23	0,854			0,853
f24	0,563			0,695
f25	0,855			0,689
f26	0,856			0,855
f27	0,855			0,855
f28		0,862		0,854
f29	0,857			0,853
f30		0,853	0,855	
f31		0,853	0,855	
f32		0,853	0,853	
f33	0,855			0,855
f34	0,659			0,847
f35	0,853			0,855
f36	0,853			0,856
f37		0,855	0,856	0,853
f38	0,461			0,681
f39	0,527			0,702

Dark green cells correspond to features included during forward selection, while dark red ones represent excluded features during backward selection. Light green cells show attributes which constitute the feature set at each stage.

A.10 Statistical Information of the Feature Set after Combined Forward and Backward Selection

	Accuracy	Precision	Recall	TNR	Kappa
AdaBoost	0,791698	0,712	0,812667	0,524667	0,791698
SVM	0,893862	0,878667	0,895667	0,774333	0,893862
Random Forest	0,902131	0,832667	0,909667	0,742333	0,902131
Combined	0,910708	0,870333	0,914667	0,785	0,910708

A.11 Feature Accuracy after Combined Forward and Backward Selection: Table

Feature	AdaBoost	SVM	Forest	All
f1	0,722	0,765	0,767	0,777
f4	0,716	0,738	0,742	0,742
f9	0,697	0,5	0,654	0,685
f10	0,659	0,5	0,65	0,65
f11	0,825	0,844	0,806	0,855
f17	0,568	0,628	0,578	0,604
f22	0,666	0,738	0,751	0,751
f30	0,584	0,448	0,573	0,55
f31	0,651	0,289	0,632	0,604
f32	0,649	0,554	0,668	0,649

Green cells report accuracy over 70%.

A.12 Feature Accuracy after Combined Forward and Backward Selection: Chart