FAKULTÄT FÜR !INFORMATIK

Affective Image Classification

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Master

im Rahmen des Studiums

Computergraphik/Digitale Bildverarbeitung

eingereicht von

Jana Machajdik

Matrikelnummer 0326272

an der
Fakultät für Informatik der Technischen Universität Wien

Betreuung:
Betreuer: Priv.-doz. Dr. Allan Hanbury

Wien, 1.10.2009 _____ _____

(Unterschrift Verfasser) (Unterschrift Betreuer)

Affective Image Classification

Jana Machajdik

1.10.2009

Abstract

Images speak more than thousand words. One of their aspects is that they can affect people on an emotional level. Since the emotions that arise in the viewer of an image are highly subjective, they are rarely indexed. However there are situations when it would be helpful if images could be retrieved based on their emotional content. Our goal is to use image-processing methods to investigate or develop methods to extract and combine low-level features that represent the emotional content of an image and build a framework that classifies images automatically. Specifically, we exploit theoretical and empirical concepts from psychology and art theory to extract image features that are specific to the domain of artworks with emotional expression.

For our work we choose a dimensional approach to emotions that is known from the field of psychophysiology [52]. According to this approach an emotion can be classified by coordinates in a two-dimensional emotion space where one axis represents valence (the type of emotion), ranging from pleasant to unpleasant, and the second axis is defined as arousal (the intensity of the emotion), ranging from calm to exciting/thrilling. Emotions mapped onto this space can be translated into words like angry, sad, exciting etc. and these can be used for automatic indexing of images.

Machine learning methods are used to learn classification based on these features.

For testing and training, we use three types of data sets, the International Affective Picture System (IAPS) [43] (which is also used by Yanulevskaya et al. [78]), a set of artistic photography downloaded from a photo sharing site (to investigate whether the conscious use of colors and textures displayed by the artists improves the classification) and a set of peer rated abstract paintings to investigate the influence of the features' performance and ratings on pictures without contextual content. Improved classification results are obtained on the IAPS set, compared to Yanulevskaya et al. [78]), who use general purpose image features.

Abstrakt - Deutsch

Bilder sprechen mehr als tausend Worte. Einer ihrer Aspekte ist, dass sie auf Menschen eine emotionale Wirkung haben. Da diese meist subjektiv wird sie selten indexiert und kann somit nicht von Suchmaschinen wiedergefunden werden. In einigen Bereichen ist aber die Wahl des gesuchten Bildes vom emotionalen Ausdrucks abhängig. Das Ziel dieser Arbeit ist es, Methoden zu untersuchen um Bilder anhand ihres emotionalen Ausdrucks zu klassifizieren. Dabei werden theoretische und em-

pirische Konzepte aus der Psychologie, Kunst und Digitaler Bildverarbeitung verwendet. Wir setzen theoretische Konzepte in mathematische Formeln um und berechnen spezifische Merkmale der Bilder. Diese werden dann von Machine Learning Algorithmen verarbeitet um eine Klassifizierung anhand von Emotionen zu erzielen. Mit unseren Merkmalen erzielen wir bessere Resultate als vergleichbare Publikationen in diesem Bereich.

Contents

CHAPTER 1

Introduction

1.1 Motivation

It is said that images speak more than a thousand words. One of their aspects is that they can affect people on an emotional level. In recent years, with the increasing use of digital photography technology by the general public, the amount of images has exploded into yet unseen numbers. Huge image collections are available through the internet. Professional and press image databases grow by thousands of images per day. These rapidly growing digital repositories create a need for effective ways of retrieving information. Currently, most systems use textual indexing to find the relevant images. In other words, they try to describe the image with less than the thousand words it speaks, but still enough words to find it, if needed. However, the indexing process is not optimal in most cases, since the people who index images have different views of what is important to write about the image than those who wish to find it. Hence it happens that often the search words are not among the image labels, although the image contains that information. Since the emotion that arises in the viewer of an image is highly subjective, they are rarely indexed. However there are situations when it would be helpful if images could be retrieved based on their emotional content.

Content-based image retrieval (CBIR) systems are built to support image search based on low level visual features, such as colors, textures or shapes. However, human perception and understanding of images is subjective and rather on the semantic level [74]. Therefore, there is a current trend towards dealing with a higher-level of multimedia semantics [60]. In this context, two levels are recognized [30]:

- Cognitive level
- Affective level

While in cognitive domain “car” is always a car and there is usually not much discussion about the correctness of retrieving an image showing a tree in an African

savannah under the label “landscape”, there might be some discussion about whether the retrieved car is “cool” or just “nice” or whether the found landscape is “peaceful” or “dull”. Furthermore, a television broadcast could make the winning team’s fans happy, the losing team’s fan sad, and elicit no emotions at all from an audience that is not interested in soccer. Moreover, there are even people who laugh at horror movies [30].

In [74], systems which analyze and retrieve images at the *affective level* are called Emotional semantic image retrieval (ESIR). Most conventional applications lack the capability to utilize human intuition and emotion appropriately in creative applications such as architecture, art, music and design, as there is no clear measure to give evaluation of fitness other than the one in the human mind [13]. These ESIR systems are built to mimic these human decisions and give the user tools to incorporate the emotional component into his or her creative work. As such, ESIR lies at the crossroads of artificial intelligence, cognitive science, psychology and aesthetics and is at its very beginning stage [74].

As an analogy to bridging the “semantic gap” in cognitive content analysis, extracting the affective content information from audiovisual signals requires bridging the “affective gap”, which can be defined as “the lack of coincidence between the measurable signal properties, commonly referred to as features, and the expected affective state in which the user is brought by perceiving the signal” [30].

Automated extraction of the affective content information from visual signals can be beneficial to various content indexing and retrieval applications [30]. In [74] an overview of the general scheme of the state of the art in ESIR systems is given. The graphical representation can be seen in Figure 1.1. Generally ESIR consists of three

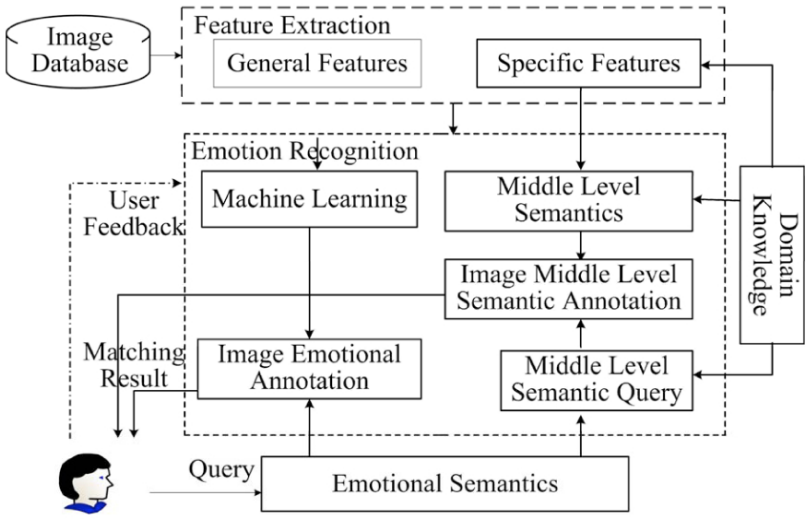

Figure 1.1: General scheme of ESIR. Image from [74].

parts: algorithms to extract perceptual features that can simulate human emotions, models to represent emotional semantics, and mechanisms to perform emotion recognition which usually involve machine learning or hierarchical models.

1.2 Issues

The typical three parts of ESIR indicate three types of key issues that have to be dealt with in this field of work, with one additional issue which is due to the character and young age of this topic.

Representation of emotion semantics

"Affective computing [55] [9] [31] seeks to provide better interaction with the user by understanding the user's emotional state and responding in a way which influences or takes into account the user's emotions" [44]. A straightforward approach to affective content analysis would be to apply the content classification methods that already proved to be promising in the cognitive domain. However, an application of these techniques to affective content classification clearly requires the prior specification and modeling of the affective content categories (e.g. happy, sad, etc.) that are to be searched for in data, which then needs to be followed by training these categories using a suitable data set.

So how do we model and represent the emotions in images?

When dealing with affective image classification it should be understood to consider the psychological aspects and categorization of emotions. Psychology provides useful emotion classification systems as well as approaches to carry out reasonable experiments with humans. We will look briefly at some of the psychological concepts regarding the field of affective classification in Chapter 3.

Features of affect

In the cognitive case the features describe the aspects of a real entity (e.g. the color red characterizing a red car). However, little is known regarding the relations between the features and something as abstract as affect. Which color combination or texture is related to happiness, disgust, or fear? The search for relations between features and the affective state they are likely to communicate may be approached by consulting a number of recent studies from the fields of art history, advertising, psychophysiology, and human-computer interaction [30]. We deal with this question in Chapter 4.

Emotion recognition

The function of emotion recognition is to bridge the gap between low-level features and high-level semantics. Basically there are two possibilities to solve this issue: the utilization of machine learning algorithms or building semantic decision rules in a hierarchical model to define the relationships. In this work, we decided to use machine learning. The precise methods and tools are described later in this chapter.

Collecting test data

While finding a representative training data set is already a considerable challenge in the cognitive domain, even for reasonably well-defined problems like face detection, this appears to be far more difficult in the affective domain. The main problem lies in the fact that the variety of content that can appear in happy, sad or exciting images is practically unlimited [30]. Moreover, the emotion inspired by a certain visual signal is extremely subjective and varies between observers. Manually labeling data with emotion is even more difficult than providing ground truth for problems such as object recognition in images. So instead of measuring the actual feeling or emotion of an individual that arises from looking at a picture, we search for methods to extract the affect that is *expected* to be evoked in a viewer and which is likely to be elicited from the majority of the audience given the stimulus. So while the individual feeling may vary from one person to another, the average or expected feeling can be considered objective, as it reflects the more or less unanimous response from a general audience [30] [78]. Our data sets and methods of labeling are described in Chapter 5.

1.3 System Flow of our Framework

In the course of this work, we implemented our own image processing framework with the goal to use image processing methods to investigate or develop methods to extract and combine low-level features that represent the emotional content of an image and build a framework that classifies images automatically.

The system flow of our image processing framework is the following (and is also illustrated in Figure 1.2):

First, the image database is set up. In our case this means placing the images in the appropriate folders along with files containing their manual emotion labels, i.e. the ground truth for classification. The presentation of our test data sets is given in detail in Chapter 5. Then, some preprocessing is done and involves resizing of the images, cropping away borders, converting the images from RGB to Improved HSL (IHSL) cylindrical color coordinate space, and segmentation of each image into continuous regions as described in Chapter 4. The segmentation result is saved along with the original image. Both are input to the feature extraction process, which presents the core of the framework (see Chapter 4 for details). During feature extraction all features are computed for each image and saved in a feature vector (or feature list). After all features of all images have been computed, the images (or to be precise, the feature vectors of the images) are split into a training and test set. The training set along with the appropriate manual emotional labels as ground truth is used to train the classifier in Weka. The resulting trained classifier is then used to automatically classify the images from the test set, i.e. each image receives a class label. During evaluation we compare the assigned automatic class labels to the ground truth (i.e. the manually assigned emotion words) and count the false or correct classifications. We conducted several different tests to explore our data. They are described in detail in Chapter 6, along with the results.

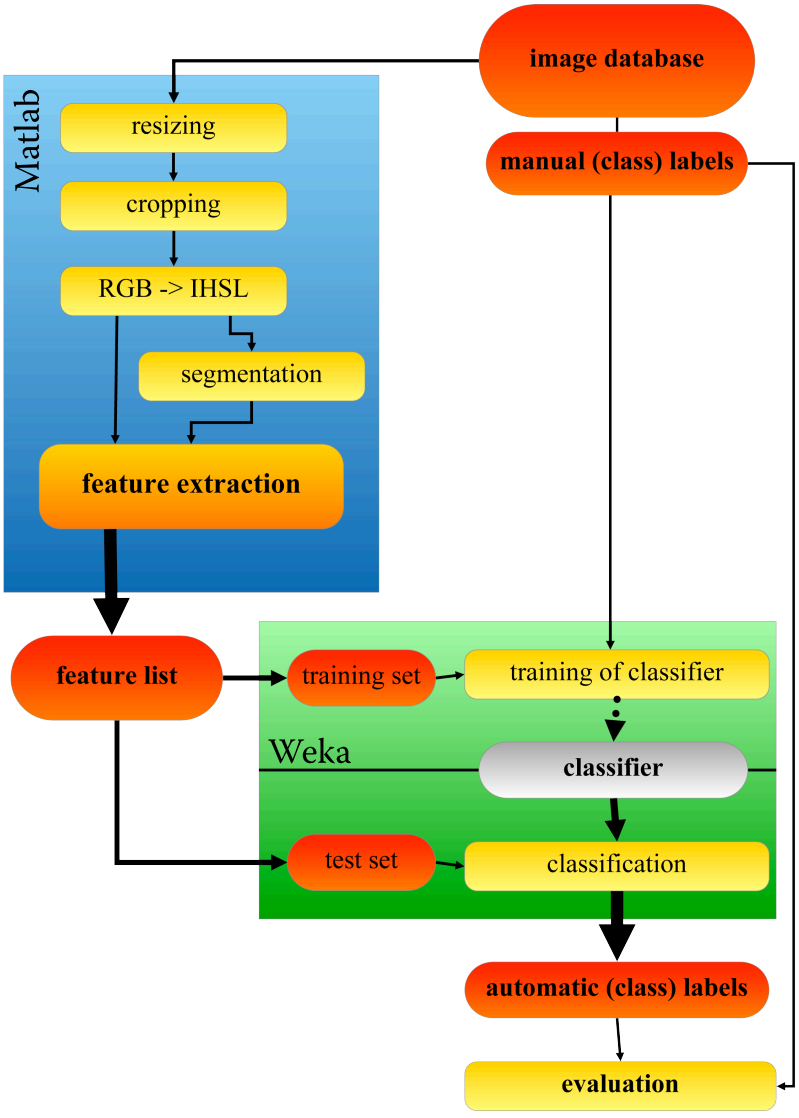

Figure 1.2: System flow.

Classifier

As classifier, we chose the Naive Bayes classifier [77] due to its good performance and speed. We also tried to use other popular classifiers, such as support vector machines (SVM), Random Forest [77] or the C.45 tree classifier [77], but in our case, the Naive Bayes proved to be the best, both in performance and speed. This may be because in our case we have non-uniform distributions of instances per class, which especially caused trouble when using the SVM. Since our data was not linearly separable, and the balance of the number of instances per class was unequal (especially during the one-against-all tests) the SVM always chose to ignore the smaller class altogether and we did not succeed in correcting this behavior by weighting the importance of each class. Moreover, the SVM was very slow compared to the other algorithms. The tree-based classifiers were fast, but delivered worse performance than Naive Bayes.

1.4 Tools

To build our image processing framework, we used the high-level technical computing language and interactive environment Matlab. **Matlab** [66] is a commercial, platform independent software built for solving mathematical problems and the graphical presentation of results. It is optimized for work with matrices, which is of great advantage when handling images. The Matlab interface provides a fast scripting language and an extended library with a wide range of image processing algorithms.

For data mining and machine learning we utilized **Weka** [77], a comprehensive open source Machine Learning toolkit, written in Java at the University of Waikato, New Zealand. It provides a library of many popular machine learning algorithms as well as tools for data mining, feature selection and visualization.

The final processing of the results and generation of the diagrams during evaluation (e.g. in Chapter 6) was done in Microsoft **Excel**.

1.5 Overview of this work

The rest of this work is organized as follows: In Chapter 2 we give a summary of the State of the Art in Emotional semantics for image retrieval. Then we examine the psychological background and possible emotional classification systems in psychology in Chapter 3. The feature extraction process is described in detail in Chapter 4 followed by a presentation of our test data sets in Chapter 5. The tests and results of the evaluation of our framework are presented in Chapter 6. Finally, we give our conclusions and discuss future work in Chapter 7.

CHAPTER 2

State of the Art

Many works dealing with object detection, scene categorization or content analysis on the cognitive level have been published, trying to bridge the semantic gap [44], but where affective retrieval and classification of visual or acoustic signals (digital media) is concerned, the publications are rare and few. Perception of emotions as evoked by visual scenes is an almost untouched area of research [15]. The expression "affective computing" was defined in 1997 by Picard [55] and the first noteworthy publication implementing a framework realizing some kind of image classification using subjective impression categories was, to our knowledge, done in 1998 ([34]).

2.1 Affective content analysis of static images

The K-DIME system presented in [10] lets the user fetch multimedia material from the Web using textual keywords (such as "airplane", "house" etc.) as input and filters the result by Kansei (Japanese for sensitivity) words, such as "romantic", "quiet" etc. The system is based on building a Kansei user model - a set of neural networks - for each user by evaluating the user's textual feedback on the images and learning a mapping between low-level features of the multimedia data and the impression words. The system requires multiple feedback-rounds from the user to build a model. With each round the model adapts itself based on the new feedback or expands if new words have been used by the user. The low level features extracted from the images are only roughly mentioned as common "color, texture and shape features". A prototype of K-DIME was implemented in the form of a personalized holiday planner where the users can search for "housing with a casual atmosphere", "romantic landscapes" etc. One of the problems stated in the article is that with an unlimited dictionary, the users had great difficulty to use the search system and the relevance feedbacks were highly inconsistent. However when a list of words was provided, they were able to perform the test tasks. The choice of the vocabulary further had great impact on the consistency of the results. However, no measurements on the

performance or quality have been published by the authors.

An early, but remarkable work was done by Colombo et al. [15]. The authors identified two distinct semantic levels: expressive and emotional level. The emotional level is formally constructed at the top of the significance hierarchy, from the levels below, namely the expressive and perceptual levels. Although there are examples that don't fit in with the hypothesis, rules are defined which map features of the lower levels to the characteristics of the top levels (i.e. low-level perceptual features are combined and mapped to expressive words and further to emotions). Such construction rules, however, depend on the specific data domain to which they refer as well as the culture in which they are to be used. Colombo et al. derived rules specific to the domain of art paintings and video commercials of the western culture. To define the rules for visual representation on an expressive level, Itten's theory of colors [37] and basic semiotic principles have been exploited. For the rules mapping to the emotional content, principles of psychology of visual communication have been translated into computational forms.
To extract the content representation of art paintings at the expressive level, the image is segmented (by a clustering in the CIE L*u*v space) into regions with uniform colors and a set of *intra-region* and *inter-region* properties are computed. The *intra-region* properties are computed for each segment individually and include color, warmth, hue, luminance, saturation, position and size of the segment, whereas *inter-region* features are based on Itten's theory and define hue, saturation, warmth and luminance contrast, and color harmony between distinct regions. A fuzzy representation model is used to describe and store each property's value as well as to express the vagueness of the problem at hand. Furthermore a grammar is defined for the representation of the features and finally each image's content representation is verified by a *model-checking* engine which computes the degree of truth of the representation over the image. An illustration of this concept is shown in Figure 2.1. To proceed with the representation at the emotional level, a set of four primary emotions is defined, consisting of the categories *action, relaxation, joy and uneasiness*. A set of plausible inputs is then defined for each of the primary emotions, and linear regression is done to achieve adaptation of proper weights for each input feature. The dependencies between the perceptual/expressive features and emotions are summarized in the following Table:

Emotion	**Dependency rule**
action	weighted measure of presence of *warmth*, presence of *hue contrasts* and presence of *slanted lines*
relaxation	weighted measure of presence of *luminance contrasts* presence of brown regions and presence of green regions
joy	weighted measure of presence of regions in *harmonic accordance*
uneasiness	weighted measure of absence of *contrast of hue*, presence of yellow and presence of purple regions

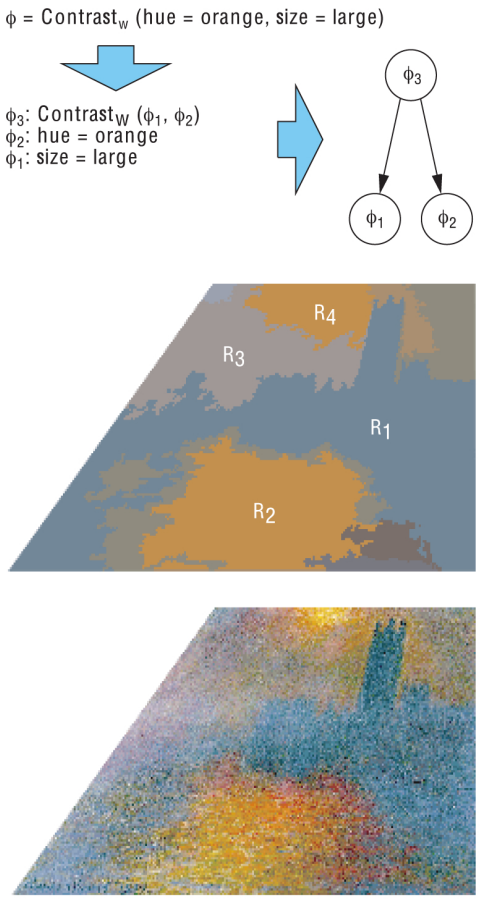

Figure 2.1: The idea of the Model-checking engine from [15]. Top: Formula decomposition. Middle: Segmentation. Bottom: Original image.

The content representation of commercial videos is done differently. At the expressive level, four semiotic categories of commercial videos are defined identified: *practical, playful, utopic and critical*. Similar to the image approach, the commercial is segmented into shots, all perceptual features are extracted and a fuzzy representation model as well as linear regression to adapt weights of dependencies are used. The *intra-frame* features measure the amount of horizontal and vertical or slanted lines and the saturation of colors, whereas the *inter-frame* features address the presence of cuts and dissolves, the presence or absence of colors recurring in many frames, the presence or absence of editing effects. The following table shows the mapping of the perceptual features to the categories at the expressional level (0 and 1 mean that the feature value should be near the minimal value 0 or maximal value 1, $\times$ indicates irrelevance):

Perceptual features	Semiotic categories			
	practical	playful	utopic	critical
saturation	~ 0	~ 1	×	×
recurrent colors	×	×	~ 1	~ 0
horizontal/vertical lines	~ 1	~ 0	×	~ 1
cuts	×	~ 1	~ 1	×
dissolves	~ 1	×	~ 1	×
editing effects	×	×	×	~ 0

The classification into emotional level categories is then done in two steps. The first classification separates commercials with action from those that induce quietness. Each class is then further separated to specify the character of the video. Action separates into suspense or excitement, quietness into relaxation and happiness. The dependencies between the perceptual/expressive features and emotional classes are summarized in the following Table:

Emotion	Dependency rule
action	weighted measure of presence of cuts, presence of high degree of motion, presence of slanted lines, presence of red and purple
excitement	categorized as action, plus weighted measure of presence of cuts and presence of short sequences
suspense	categorized as action, plus weighted measure of presence of frequent cuts, presence of both, long and short sequences
quietness	weighted measure of presence of long sequences, presence of dissolves between frames, presence of blue, orange, green and white colors and absence of black and purple colors
relaxation	categorized as quietness, plus weighted measure of absence of motion
happiness	categorized as quietness, plus weighted measure of presence of motion

To test the effectiveness of the expressional features, 40 classic paintings (from Renaissance to contemporary art) where ranked by 35 experts and the implemented retrieval system according to 4 queries addressing the contrast of luminance, warmth, saturation and harmony. The similarity of the ranking was measured with the resulting average effectiveness between 60% and 70% on all 4 queries. Further, 20 commercial videos were classified into the four semiotic categories by 5 experts and the classification was compared to the values of the emotional features automatically extracted by the retrieval system. Here, the agreement ranged from 60% for practical and critical, 80% for utopic to more than 90% for playful. The features on the emotional level were not tested.

Another work in the field of affective computing is done by Wu et al. [56]. Claiming a lack of unified emotion categorization systems, they created their own affective space by choosing 9 affective word pairs: beautiful - ugly, dynamic - static, cheerful - gloomy, active - passive, natural - artificial, tense - relaxed, simple - complex, bright - dark and hot - cold. An eigenvalue decomposition calculated by principal component analysis (PCA) was used to simplify the emotion space and transform it into an orthogonal three dimensional space. As in most works in this area, 3 groups of low-level visual features were extracted from the images: color, texture and shape features. To measure color, color correlograms [36] were computed after the conversion of the image into HSV space. The first three Tamura features [65] were used to describe texture, and moment invariants [35] were chosen to represent shapes. These features were used to train a support vector machine. They conducted their experiments with 150 carefully selected landscape images as the training set and test set respectively. The images in both sets were selected from a database of 1420 scenery images by excluding all ambiguous images. Twelve users then examined the 150 images in the sets and put a score for each of the adjective word pairs for each image. These scores were used as ground truth. The experimental results show that color features had higher accuracy (84%) than texture (accuracy rate: 79,6 %) and shape features (accuracy rate: 81.3%), but the combination of all features showed better results than using subsets of features. The accuracy rate for the combined (color + texture + shape) features was 93,3%, with 75 correct samples and 5 errors in the test set. However, it is possible that these good results are due to the very specific testing data.

One of the early systems realizing image query by impression words was presented by Hayashi and Hagiwara in [34]. They adopted the RGB color projection distributions in the vertical and horizontal axis as color image features. To calculate the projection the image was divided into L horizontal and L vertical *belts* and the average value over each *belt* was computed for each R,G and B channel, so for each image there were $(Height + Width)/L \times 3$ color features. These features were chosen because of their simplicity, fast processing time and to avoid image segmentation and object detection. Further, *spectrum characteristics* were computed by using the FFT on the horizontal and vertical directions of the grayscale-transformed image. Both FFTs are divided into N ranges from the DC component to higher frequency components, where the amplitudes over each of the N ranges are averaged and normalized. As impression words Hayashi and Hagiwara chose 27 adjectives and 8 words describing weather, time and seasons. Together this were the following 35 words: spring, summer, fall, winter, sunny, cloudy, sunset, night, soft, hard, rich, solemn, hot, warm, cool, cold, dry, fresh, pretty, pleasant, strong, rustic, romantic, lonely, country, mysterious, urban, clear, peaceful, calm, young, elegant, severe, active, *märchen*(fairy tale). A multi-layer neural network was trained with the back propagation algorithm to learn the correlations between the features and the impression words. For training and evaluation of the proposed *IQI system* 120 scenic 200×140 pixel images were rated by 10 users using an evaluation sheet with the 35 impression words. The answers were averaged to produce ground truth. By setting $L = 20$ and $N = 4$, 55 features where produced per image. 40 images where taken

to train the neural network. The final estimator was a 3-layer neural network with 55 nodes in the input layer, 35 nodes in the output layer and 13 nodes in the hidden layer. For evaluation, *"the rate where the most impressive word in the evaluation data for each image is within the top 5 in the impression words vector estimated automatically"* was computed 10 times for the 80 remaining images. The best rate was 78.8%. For detailed evaluation of each word, see [34].

A comparable approach was taken in [75]. Following the semantic differential technique [52] [63], 12 emotional word pairs (exhilarated - depressive (E1), warm - cool (E2), happy - sad (E3), light - heavy (E4), hard - soft (E5), brilliant - gloomy (E6), lively - tedious (E7), magnificent - modest (E8), vibrant - desolate (E9), showy - elegant (E10), clear -fuzzy (E11), fanciful - realistic (E12)) were selected to build up an orthogonal three-dimensional emotional factor space. The three factors F1, F2 and F3 determined by PCA had strong correlation with the above features (F1 correlated with E1-E7, F2 correlated with E8-E10 and F3 correlated strongly with E11-E12). To avoid huge feature vectors which result from using "common" features with implicit relationship between the features and the emotional semantics only, Wang et al. developed three novel image features designed for their three emotional factors specifically. By analyzing their image database, a combination of lightness and warm-cool description was determined as feature set for the first emotional factor F1. Using the fuzzy logic algorithm proposed in [73], a 5 value fuzzy membership function ranging from "very dark" to "very light", expressing the perceived lightness level, was created for each pixel. This is then combined with a warm-cool membership function as defined in [53] to create a 10-dimensional histogram describing the light-warm-cool distribution of the image. For the second factor, a 7-dimensional saturation-warm-cool fuzzy histogram was derived in a similar fashion as before, but with only three values used for saturation and an additional contrast measure. The third factor features were found to be best represented by a sharpness measurement computed by taking the top 5% average of the gradient values, and a contrast of lightness measure. For each factor a Support Vector Machine of Regression (SVR) was constructed and trained to learn the exact mapping between the features and the factors. A factor matrix then defines the mapping between the 3 factors and the 12 emotion words. To get ground truth, a web application was built and posted online. The users were asked to rate 100 art images by giving a score for each of the 12 emotional word pairs. 42 users completed the rating session and it took each user 1-2 hours. For evaluation 75% of the images were used as training set and 25% were in the test set. The results for the three SVRs were all above 85% correct. Compared to the method by Hayashi et. al [34] with a 65-dimensional non-specific feature vector and neural network with direct evaluation to emotion words, the presented approach performs better.

A different approach was chosen by Yanulevskaya et al. in [78]. They adapted a scene categorization algorithm from [69] and combined it with machine learning to perform emotional valence categorization. The training set was a subset of the International Affective Picture System (IAPS) [43] which was categorized into distinct emotional categories by Mikels et al. [49]. The emotional categories are: anger, awe,

disgust, fear, sadness, excitement, contentment, and amusement. Instead of the typical "color, texture and shape features", they utilized the "visual words vocabulary" from [69], where the features are defined as similarities to the 15 proto-concepts (building, car, charts, crowd, desert, fire, US-flag, maps, mountain, road, sky, smoke, snow, vegetation, water) proposed in [69]. By using similarities to the whole vocabulary, instead of just choosing one "word", it is possible to model scenes that consist of elements not in the code-book vocabulary. To measure the similarities to the 15 proto-concepts, the images were divided into r overlapping rectangular regions (with r =4 for coarse and r =17 for fine sampling) and holistic image statistics features, represented by Wiccest features [22] and Gabor filters [11], were extracted. These features were then used to train a Support Vector Machine classifier to distinguish between the various emotional valences. Optimization on the SVM parameters and feature selection were performed to get the best results. As ground truth, the 396 categorized IAPS-pictures were divided into a training (70%) and testing (30%) set. Overall, the proposed system performs slightly better than chance (50% one-versus-all). The trained system was also tested with some masterpieces from an art museum and for these paintings, the system is claimed to perform better than in the case of the IAPS data set.

Sung-Bae Cho [13] developed a human-computer interface for the purpose of aiding humans in the creative process of fields such as architecture, art, music and design. The proposed human-computer interface was built to help utilize human intuition and emotion appropriately in creative applications. The presented technique is called interactive genetic algorithm (IGA) and is a genetic algorithm [23] that performs optimization with the help of human evaluation. A human can obtain what he has in mind through repeated interaction with the method. The author suggests that this allows the development of effective human-oriented evolutionary systems, since this way a fitness function for the genetic search algorithm, which is not definable otherwise, is obtained directly and entirely from the human user. Therefore this system provides a solution that reflects the user preference of the particular user. The technique is implemented for image search and music search. For the image search, indices of wavelet coefficients are used to represent *chromosomes* for the algorithm. Each image is decomposed using the Haar wavelet transform, but only the signs of the 50 largest coefficients in the red, green and blue channels are stored in the *chromosome* (i.e. 50×3 array) and used for image search. With each generation, a selection of 12 images is presented to the user (i.e. the size of the population is 12). Each time the user evaluates which images seem to be fit on a 7-score basis. The system takes two images with the highest fitness, performs a crossover exchanging the color and shape features respectively and using an empirical similarity function to retrieve the 12 images from the database that would be displayed as the search result in the next generation. This generation can be again evaluated by the user etc. This process can be repeated until the user is satisfied with the result. The system was tested by letting 10 users search for images with a gloomy or cheerful impression manually and then using the system. The images were compared and the users were asked to rate the difference in expression between them. On average, the users were somewhat satisfied with the IGA search results, giving them average scores between -1 and 2 (from $[-3, 3]$) for cheerful images and 0 to 2 for gloomy images. A major

drawback of this method is that it needs many human evaluations each time, until the system gives suitable results and therefore the human fatigue problem is one that would prevent many users from actually using this system on a regular basis.

2.2 Affective content analysis in films

Audio-visual signals, such as movies are somewhat related to this topic. As was shown in [15] the connection or adaptation of methods for image retrieval and video retrieval can be profitable. More so, since there are so few publications in this area. In following, some of the studied approaches are presented.

An interesting work on the affective content analysis in movies is presented in [30] and [31] by Hanjalic et al. They propose a computational framework for affective video content representation and modeling. The representation part of the framework consists of a set of curves that depict the expected transition from one feeling to another along a video, measuring the *expected* emotion or major feeling as elicited from a general user. The set of *affect curves* consists of curves representing the changing of the value of the *valence* and *arousal* dimension of the three-dimensional emotion space as defined in [52] throughout the movie as illustrated by the example in Figure 2.2. *Valence* characterizes the level of "pleasure", ranging from "extremely unpleasant" to "highly pleasant" and *arousal* indicates the intensity of the emotional experience, ranging from "energized" and "alert" to "calm" and "peaceful". The modeling part addresses the problem of computing

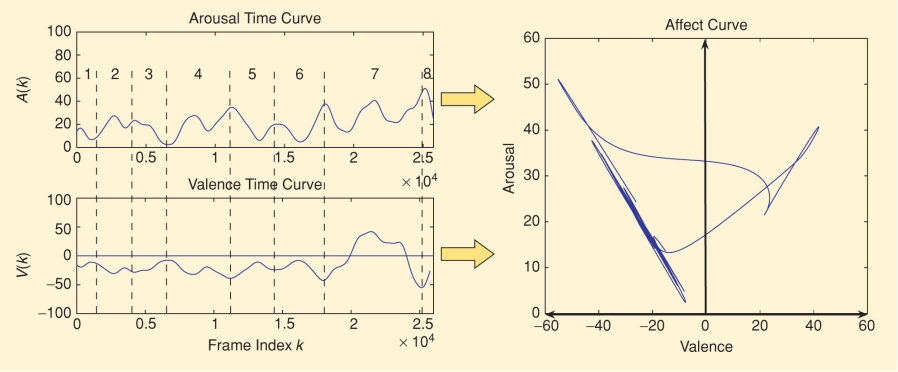

Figure 2.2: The arousal, valence and affect curve obtained for an excerpt from the movie *Saving Private Ryan* [30].

the values of the content representation curves on the basis of low-level features extracted from video as described in [31]. To model the *arousal time curve*, the weighted average of three feature time curves is computed. The *arousal time curve* is therefore represented by the following three features: the *motion component*, obtained on the basis of the overall motion activity measured between consecutive

video frames, the *rhythm component*, obtained by investigating the changes in shot lengths of the video, and the *sound energy component*, obtained in synchronization with video frame interval by computing the total energy in the sound track of a video. The *valence time curve* is modeled based solely on the smoothed time function of changing of signs of the the *pitch-average component* of the soundtrack of the video. During the development of the *affect curves*, a special emphasis was placed on the satisfaction of the following validity criteria: comparability, compatibility and smoothness. Further, in [30], techniques to use the proposed *affect time curves* to achieve automatic personalized video highlight extraction, and generating of program previews, ideas about personalized video delivery, and personalization based on the affective user-profile generation, and video retrieval through the 2-D affect space are presented. However, no large-scale evaluation of the presented techniques was done, only few case examples are shown and discussed. Based on the two examples (a soccer match and an excerpt from the movie *Saving Private Ryan*) it can be seen that the affect curves represent the affect of the video clips fairly well. However, the simplicity of the function representing the pitch-average component also leads to imperfections in the valence time curve causing inaccuracies in the evaluation of some scenes. Despite this, the results seem to be promising. Unfortunately this system doesn't use any visual features applicable to static images, neither does it output any emotional categories, only affect curves, which are open for interpretation, but none is given. Some of the ideas of application of affective analysis are, however, interesting and might find usage in the area of static images.

A work dealing with affective understanding in film was published in [72]. The first main contribution is a complementary approach to the identification of suitable emotional output categories for affective film categorization, which is grounded in the related fields of cinematography and psychology. In contrast to most other works in this field, the output emotional categories are complete (covering roughly the whole spectrum of emotions useful for film categorization) and chosen with clear reason which is theoretically founded and supported by psychological studies as well as on cinematographic basis and gives a meaningful classification for films. The authors identify two most dominant theoretical psychological perspectives on emotions: the Darwinian perspective and the cognitive perspective (both will be explained later in this work). Based on these theories a set of emotions and their mapping onto a two-dimensional emotion space (as shown in Figure 2.3) is defined and then modified to fit the cinematographic purpose and field of work to create the following emotional output categories: *Anger*, *Sad*, *Fear*, *Joyous*, *Surprise*, *Tender Affection* and *Neutral*. The second contribution is the development of a set of effective audio-visual features for the purpose of affective film classification. The system overview and listing of features is best illustrated by the flow chart in Figure 2.4. The input of the system are manually segmented scenes. For each scene the visual and audio signals are processed separately. The visual signal is further segmented into shots and key-frames. For each such shot or frame a list of features are computed. Visual features consist of shot duration, shot density, motion analysis corresponding to visual excitement, lighting key measurement by amount of shadow area and median level of brightness, color energy (a product of raw energy and color contrast) and a

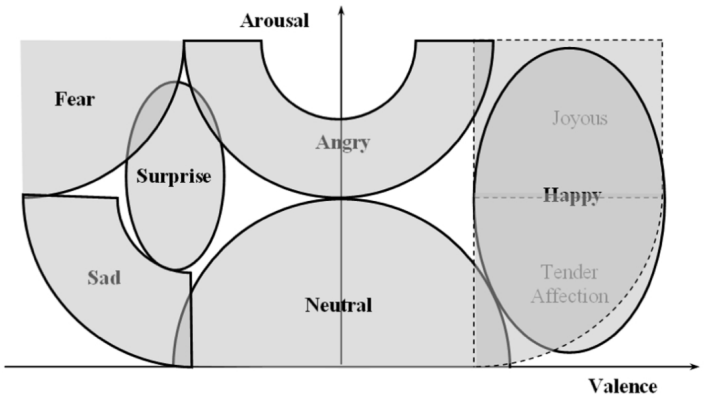

Figure 2.3: Emotional output categories defined by [72] and their mapping onto a 2-D emotional space.

measure for visual detail by average of a grey-level co-occurrence matrix. The audio signal is separated according to its audio-type. Here the types music, environmental noise, speech or silence are used to differentiate the signal. Each type is described with several specific features among which are sound energy statistics, low energy ratio, spectral roll-off and centroid, MFCC and its delta statistics, ZCR statistics, spectral flux, LFPC, LSTER, chroma and its statistics, octave bands and music scale. These separated sounds are then given to a support vector machine based probabilistic inference machine to recognize composition of moods in the audio signal. The output of the probabilistic SVM is integrated over all sound units of the scene and weighted by their duration to form the *scene audio affect vector*, a vector of high-level audio features at scene level. The visual and audio features are then put together to form a feature vector representing the whole scene, which are finally sent to the same inference machine as before to obtain probabilistic membership vectors. The training data for this work consisted of 36 full-length movies of various genres, all of which where relatively recent mainstream Hollywood movies chosen to represent popular films, and a diversity of director styles and emotions with the goal to get an unbiased data set. They were segmented into 2040 scenes which were labeled independently by three persons according to the perceived prevailing emotion. In most cases the labels agreed unanimously. Only in 14.08% of all scenes were assigned 2 different labels, with no cases of three different votes. For testing, a take-one-movie-out approach was chosen, where all scenes of one movie where removed from the data set and used for testing and all others where used for training. This was repeated for every movie. A thorough discussion of the results is given in the paper, but the most interesting part of the findings is an observed lack of strong correlation between the simple low-level visual features and the affective content of the scenes (only 42.86% correct rate when using only visual features). Except in cases of extremes, color doesn't seem to correspond well to the mood of the movie. Sound, on the other hand, seems to be

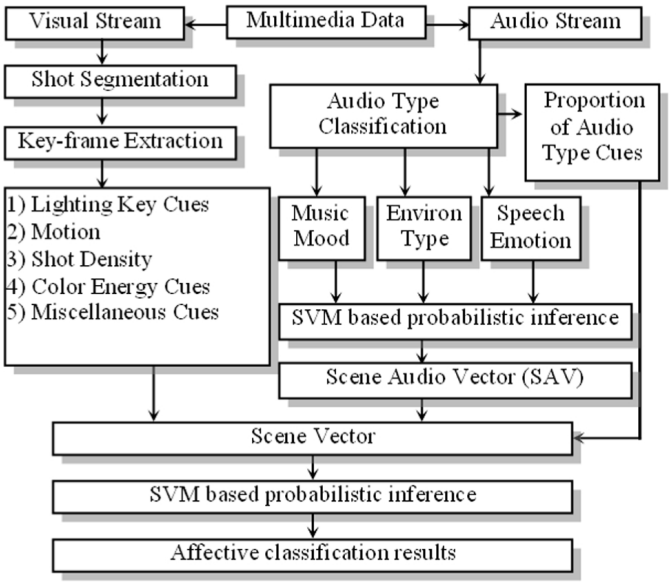

Figure 2.4: System overview for affective film classification by [72].

much more informative (61.39% correct rate with audio features only). The overall correct classification rate (using both visual and audio features) is 74.69% or 85.82% if scenes where the "alternate" label (in cases of dual labeling) is selected are added as correct.

Other works dealing with affect-related issues in film using visual features are [39], [79] and [57]. However, these approaches aren't as thorough and elaborate as the work by [72] mentioned above.

Kang et al. [39] utilize visual features such as the occurrence of 11 basic colors, as well as average light and saturation of each shot, further they categorize camera motion via frame differencing into 8 directions with the help of Hidden Markov Models (HMMs), and compute the shot durations. These features are computed on both shot and scene level and used in the attempt to automatically distinguish between scenes depicting fear, happiness or sadness.

Zhai et al. [79] propose the use of Finite State Machines (FSM) for detecting and classifying movie scenes into three types; conversation, suspense and action. The low level features used in this approach include motion and audio energy as well as a mid-level feature, face detection. The transitions of the FSMs are determined by the features computed from each shot in the scene. The framework has been tested on 60 video clips.

In [57] a mean-shift based clustering framework is proposed with the goal to classify films into genres such as Comedy, Action, Drama or Horror. The computed

features are selected according to cinematographic rules or guidelines and include four characteristics: average shot length, color variance, motion content and lighting key.

Yet another approach to extract information about emotions in films was explored by Salway and Graham in [60]. They decided to take advantage of audio descriptions for the visually impaired which are provided by an increasing number of television programs and films. Except for being an invaluable means for improving access to the visually impaired, these descriptions can be used as source for retrieving or generating descriptions of the semantic and emotional video content. Such audio descriptions are inserted between existing dialogues and describe the important information about on-screen scenes and events, about characters' actions and appearances. Salway and Graham scan for possible descriptions of visibly manifested emotions in the time-coded description texts and classify them. For classification they use 22 types of emotional classes proposed by [51] and produce a list of emotion keywords for each of class. In total 679 emotion keywords were selected for the 22 emotional classes by retrieving synonyms and hyponyms from WordNet [4]. Occurrences of these keywords in the audio descriptions where then plotted against the time-code of the description. From observation and interpretation of the plots found that clusters of emotion keywords appear to identify many of the dramatically important sequences of the movie and that there is a correspondence with the notions of story structure and conventional character behavior. Further they tested the performance of this system by comparing its results with human responses. Ten short (1-3 minutes length) film sequences were shown to 10 subjects, who had to report up to three emotions that they thought were depicted in the sequence. Those emotions which were reported by the majority of the subjects were considered *important*. The system retrieved 12 out of 19 (1-2 per sequence) such emotions correctly (63%). It retrieved another 11 out of which 7 were observed by only one or two subjects and 4 which were not reported at all by the subjects.

2.3 Critique

While the previously described works have certainly advanced research in the field of affective content analysis, many of them have at least one of the following issues/drawbacks:

general features - Although some of the works suggest that art and impression specific features are of advantage [75] [17], some of the works in this field use common, general or holistic features (e.g. in [10], [56], [78]).

incomplete emotional categories - Often, the emotional categories used as output of the given classification are ad hoc and incomplete. As was shown in [72] and many psychological studies [52] [49], choosing meaningful emotional categories is not an easy task and requires thorough consideration. The number of "emotional categories" occurring in the discussed works range from no cat-

egorization at all (e.g. [31]) to 35 “impression words” (in [34]). Furthermore, the output categories are on different *levels of significance* (according to [15]). Most of the “ *kansei* impression words” as used in [34], [56], [75] or [10] are at the *expression level*, whereas emotional adjectives as used e.g. in [78] or [72] are at the higher, *emotional level*.

unpublished or questionable data sets - The data sets are in most cases unknown (unpublished) and in some cases explicitly handpicked which brings up the question of how the images were selected or whether the prior manual filtering process wasn't too biased by the aims and methods of the selector. Especially in the cases where a few images/scenes/previews have been selected out of thousands without a good reason or description of the method, the suspicion lies near that this has been done with the goal (or at least a bias) to select pictures (or scenes) which would be discernible by the implemented methods.

missing or unclear evaluation - Another problem is presented by the unclear measures of success (e.g. in [78]), incomplete description (e.g. [13]) or even absolute lack of evaluation (e.g. in [31], [10]) of the presented methods.

Due to these factors, the works are incomparable.

In this work we overcome these problems by using a dimensional approach to emotions and choosing emotional categories defined in a proper psychological study [49], we use an image data set which is available to any scientist [43], we describe our evaluation measures extensively and show detailed results. The main goal of this work is to study features that are specific to the task at hand and we discuss them in detail in the following chapters.

CHAPTER 3

Psychological Background

Images invoke a wide range of emotions. However, to have a reasonable labeling for a classification process, we need a finite number of output labels or classes. Therefore, a fundamental challenge lies in selecting the appropriate output emotions that would be representative enough to cover a large part of the emotions that arise in humans from viewing images, as well as small enough for an automated classification system to work. Such emotional words should be universal (universally comprehended and experienced), distinctive (each emotion clearly distinguishable from another), usable (i.e. relevant in image classification context) and comprehensive [72].

The classification of emotions is a much investigated topic in psychology and various researchers provide useful emotion classification systems as well as approaches to carry out reasonable experiments with humans.

Wang and Cheong [72] made a survey of contemporary psychological theories and research in regard of affective film classification and showed that the emotion selection process is not a trivial task. They identify the two most dominant theoretical psychological perspectives on emotions, described in more detail in the next two subsections.

- the Darwinian perspective implying biological origins and universality of human emotions assuming their importance as survival functions, e.g. according to Ekman [21] there are six basic emotions: Happy, Surprise, Anger, Sad, Fear, Disgust

- the cognitive perspective postulating that *appraisal*, "a thought process that evaluates the desirability of circumstances", ultimately gives rise to emotion [72]. On account of this paradigm the dimensional approach to describe emotion is grounded

3.1 Darwinian perspective

The Darwinian perspective postulates that basic emotions are phenomena that have an important survival function to a species and therefore are of biological origin. This implies that certain human emotions (the basic emotions) are universal to the whole species. Ekman [21] provided an impressive body of evidence in his study of human facial expressions. He identified six basic emotions along with their accompanying facial expressions which were recognizable across many human cultures: Happy, Surprise, Anger, Sad, Fear, and Disgust. However, other researchers argue against this "basic emotion list", each having their reasons and evidence [50]. Hence, there exist several different lists of "basic emotions".

3.2 Cognitive perspective

The fundamental assumption in the cognitive perspective is that *appraisal*, a thought process that evaluates the desirability of circumstances, is the cause of emotions [72]. Based on this assumption a dimensional approach was developed. According to this approach, a human affective response or state can be represented using the following three basic dimensions [52]:

- Valence (V)
- Arousal (A)
- Control (Dominance) (C)

The *Valence* typically characterizes the level of "pleasure" that is related to a given affective state. It can take positive or negative values, ranging from "highly pleasant" to "extremely unpleasant". The value of *Arousal* stands for the intensity of the experience while in a given affective state. It ranges from "energized, excited, and alert" to "calm, drowsy, or peaceful". The third dimension - *Control (Dominance)* - is particularly useful in distinguishing among affective states having similar *Arousal* and *Valence* (e.g. differentiating between "grief" and "rage"), and typically ranges from "no control" to "full control". The main benefit of this three-dimensional space is that all emotions can be mapped into it [52]. While for every affective state there is the corresponding value in the three-dimensional coordinate space, not every point in the VAC space represents an affective state. Psychophysiological experiments, which included measurements of affective responses of a large group of subjects to calibrated audiovisual stimuli, the International Affective Picture System (IAPS) [43] and the International Affective Digitized Sounds system (IADS) [12], showed that only certain areas of this space are relevant [31], e.g. there is no highly aroused neutral feeling. An illustration of this three-dimensional space and the emotionally relevant regions can be seen in Figure 3.1. In the first experiments with IAPS and IADS, subjects affective responses to the stimuli were quantified either by evaluating their self-reports from a questionnaire or measuring physiological functions that are considered related to the particular affect dimensions. For example, heart rate reliably indicates *Valence*, while skin conductance is associated with *Arousal*. [31] [43]

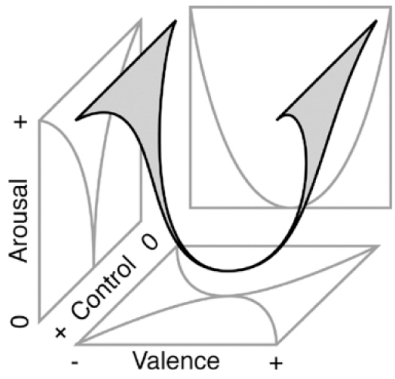

Figure 3.1: Valence-Arousal-Control space and the regions which are relevant for mapping emotions. Image from [31].

However, Greenwald et al. [27] showed that *Valence* and *Arousal* account for most of the independent variance in affective responses. Furthermore, numerous studies of human affective responses to media have shown that "emotion elicited by pictures, television, radio, computers, and sounds can be mapped onto an emotion space created by the arousal and valence axes" [20]. For these reasons, the *Control* dimension can be neglected when developing affective content analysis methods. [31]

By dropping the *Control* dimension we get a two-dimensional Valence-Arousal (VA) coordinate system. As in the 3D case, not all of the 2D space is taken up by emotions. The mappings of the obtained affective responses onto the 2D VA coordinate system is roughly parabolic [31], as illustrated in Figure 3.2.

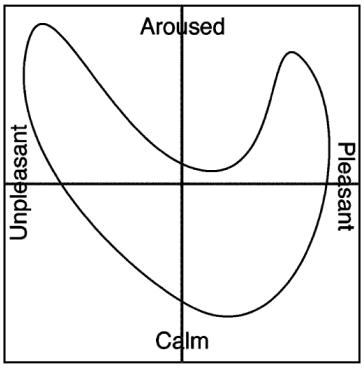

Figure 3.2: Valence-Arousal space. Image from [31].

3.3 Combined approach

Although there is no universally accepted list of "basic emotions", the Darwinian paradigm supports the usage of a very short list of emotional words to describe the complex range of human emotions. Researchers in the field of affective content analysis mainly use a method called Semantic Differential (SD) [52], a method from experimental psychology, to describe emotional semantics. Adjective pairs are used to represent emotions and an observer is asked to evaluate the emotion evoked by the shown image by putting a score to each emotional word pair. The average score from all observers is taken as the emotional rating of the image.

Another popular method is to do a two-step experiment, first asking the observers to freely describe the image they see by emotional words, then selecting the most frequently used words and repeating the experiment, but this time asking the users to rate how well each of the selected emotional words fits for the given image. This method was used by Mikels et al. [49] (and is described in more detail in Chapter 5) and we chose to use his results in terms of rated images and emotional categories, which included the following emotional words: *Amusement, Awe, Contentment, Excitement, Anger, Disgust, Fear, Sad*.

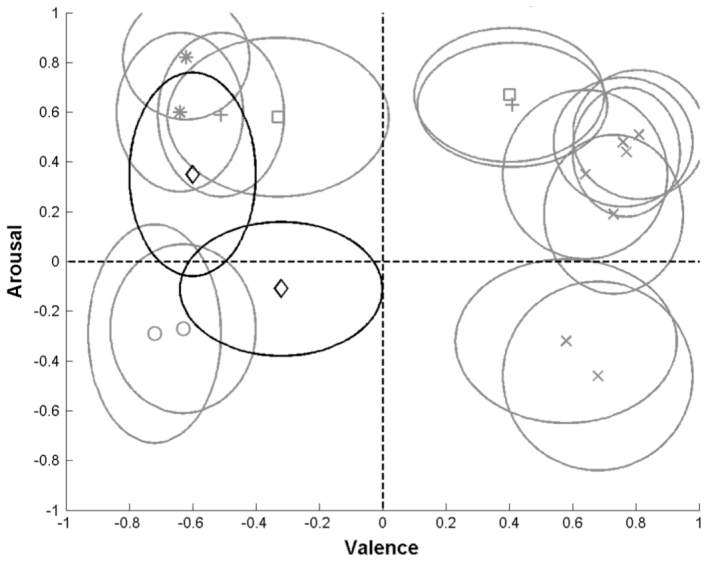

Figure 3.3: Influence of basic emotions in Valence-Arousal space. Anger + aggression (+), sad + depressed (*o*), fear + fearful + terrified (∗), happy + relaxed + joy + kind + affection + excitement (×), surprise + tense (square), disgust (diamond). Image from [72].

The concept of having just a few emotional words is better for the purpose of classification and labeling than the dimensional approach. However, the dimensional

approach to emotions is convenient for visual representation and clustering of emotionally categorized data. An example of mapping emotional words to the VA space is shown in Figure 3.3.

A particularly interesting application of this mapping was proposed in [30] and [31], and involved creating a user interface by mapping of the categorized movies onto the VA space, thereby providing a convenient overview of the possible genres and showing a list of movies in the region of interest that the user selected. This concept is illustrated in Figure 3.4 and Figure 3.5, and could be easily adapted to image retrieval.

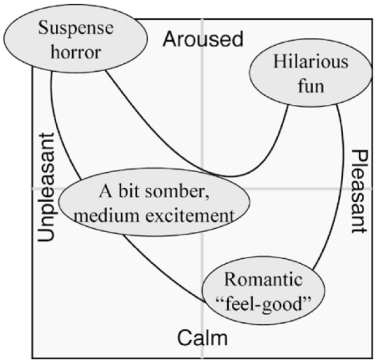

Figure 3.4: Illustration of the possibility for video content indexing and visualization. Image from [31].

We see the main advantage of a combined approach in the dual possibility to represent emotions to the end user. When searching for a specific emotion, the user could enter the search word specifying the emotion that is nearest to what he/she has in mind as we know it from the currently common search machines, but with the possibility to narrow his/her search by adjusting the Valence or Arousal coordinate and thereby specifying what he/she exactly means by the word. However, with the same labeling the user could take a browser-based approach, specifying more vaguely an emotional subspace and he/she could get clusters of thumbnails of images that map to that specific area, as well as their emotional labels. Such an implementation of the user-interface could be very handy when implementing such an enhancement into an image retrieval application. Implementing such a user interface goes beyond the scope of this work, but this model could be exploited in the future.

For now, we decided to use the results from the studies in [49]. We could have built our own system (as many other scientists did) with our own set of words and VA-mappings, but this would be contrary to our notion to make this work comparable to the works of others. Instead of conducting psychological experiments ourselves, we take the results of psychological studies which are conclusive and available. Moreover, these same emotional categories and image sets have been used by another recent publication in the field of affective content analysis, namely by Yanulevskaya et al. [78], which gives us a convenient work to which to compare our results.

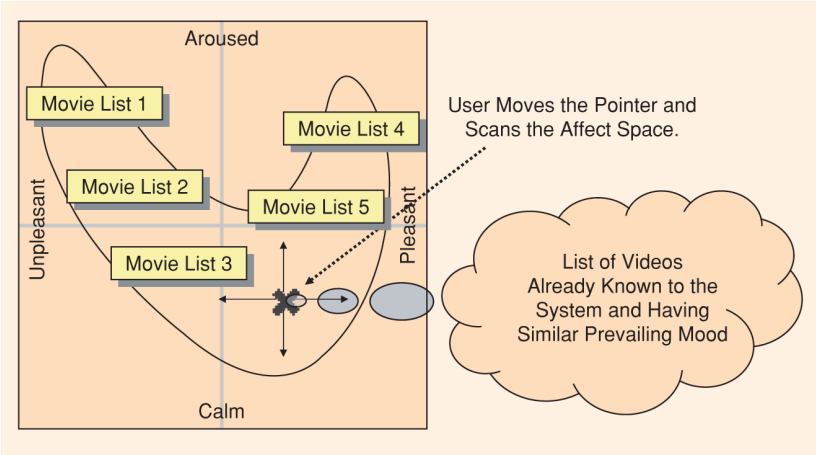

Figure 3.5: Personalized video delivery based on browsing the VA space. The user moves the pointer across the 2D affect space, first guided by the meaning of the description of the genres (as in Figure 3.4) and the affect bounds and then the video lists appearing at each location, each with the prevailing mood which the user picked. Image from [30] .

CHAPTER 4

Feature Extraction

In comparison with movies or videos we have a more difficult situation with static images. This is because much of the information available in videos is missing and as was shown in [72], the visual signal is much less informative than the audio. Furthermore we have even less information than the visual part of a video, since movement, cuts and shot duration are not available. However, since we study images which were never supposed to move, we assume that more care is put into the selection of colors and composition of the image than in a movie, and that there is more information contained in these characteristics, at least in the case of professional photographs or artwork.

This chapter deals with the feature extraction process and describes each feature in detail, giving examples of the feature values and representative images to illustrate their meaning. In the feature extraction part of the process, the procedure is as follows: First, the images are resized, cropped and segmented and converted into the IHSL cylindrical color space. Then features are computed on the resized and cropped images.

Instead of simply extracting common features like color histograms the goal was to find features that would model the relationships of the visual appearance of an image to its affective impact. There are several theories about how artists use colors and shapes to effectively express emotional content visually. Johannes Itten's color theories [37] are a famous example. We use a selection of features, many of which we have previously used in other image retrieval tasks [47], [64].

4.1 Preprocessing

Before the actual features are computed, several steps are taken to prepare the images for the feature extraction process.

Resizing

As the first preprocessing step, all images are resized to the size of 200 000 pixels, leaving their original aspect ratio unchanged, to normalize the data and to provide consistent input. This size was chosen because it is big enough to discern details of the image, but small enough to enable reasonably fast processing speeds. It may be argued that the size of the image affects the impact of the image on a human viewer. A large wall image probably draws more attention than a thumbnail on a corner of a magazine. However, we hypothesize that the emotional affect the image induces does not change with the size, only the impact or "amount" of the emotion may, depending on the context. A happy image will still look happy, and a sad image will still look sad... Moreover, the focus of this work is on digital images, which in most cases will be seen on a computer screen, where the differences of size are negligible and the images are easily resizable.

Cropping

As the next step, "digital frames" or single-color borders added to the image are cropped away. Artists and art photographers who publish their pictures online like to add digital frames around their photographs, but we hypothesize that these do not have any impact on the affect of the image itself. Although humans subconsciously prefer images with boundaries, consciously, or possibly semantically, we are able to eliminate them from our consideration of an image, or to view them as an extraneous feature [7]. Since the presence (or lack) of such frames strongly affects the feature values, but doesn't affect the perceived emotion, or at least not even in a similarly strong proportion (which can be clearly illustrated on the ACQUINE system [58]), we decided to remove the frames to eliminate their influence.

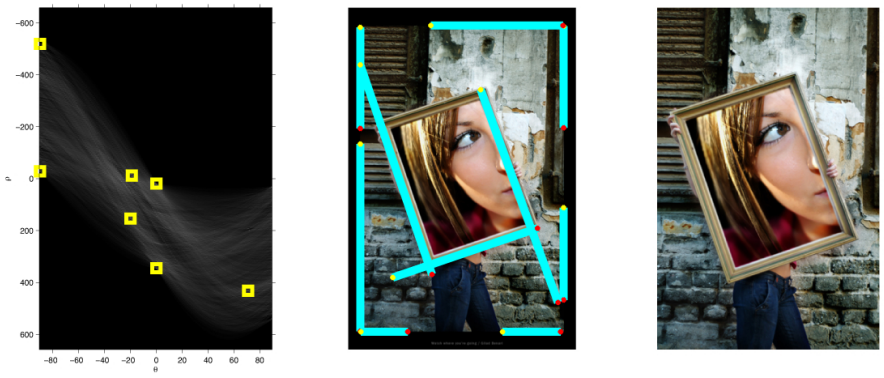

Figure 4.1: Cropping - normal case. Left: Hough space with peaks marked by squares. Middle: Original image with lines detected by Hough transform. Cropping is done along the lines. Right: Image after cropping.

Cropping is done by finding purely horizontal and vertical lines by the means of

Hough transform [26]. We observed that if there are artificial borders, the Hough transform has the strongest peaks at 0 and 90 degrees depicting the border lines. If such a peak histogram is observed, the image is cropped along the corresponding lines (additionally cropping away a few pixels more to provide for cases with a double stroked border line). However, this approach works only on image where the border has a strong contrast to the image on all four edges. In such cases, it would be sufficient to take the vertical and horizontal lines nearest to the image boundary and crop along these. To provide for cases where only parts of the borders are detected, because e.g. the contrast between the border and the image becomes to small (as in Figure 4.2), an additional routine is added. We employ Canny edge detection to find the start and end points of the lines found by the Hough transform. These lines are shown in the Figures 4.1 and 4.2. The start and end points nearest to the borders of the image are taken as coordinates for the cropping. In cases where more than a third of the picture width or height would be cropped away, it is assumed that this is because of missing lines. Further the assumption is made that the border is symmetrical and therefore the image is cropped symmetrically, taking the line with the minimum distance to the border as cropping distance in the appropriate direction. An example of this concept is shown and explained in Figure 4.2. The threshold for this corrective measure is set to a third of the width and length, because it does not matter if a small part of the image is cut away in a few cases. We are confident that a few pixels tighter cropping than chosen by the author of the image will not change an image's whole expression.

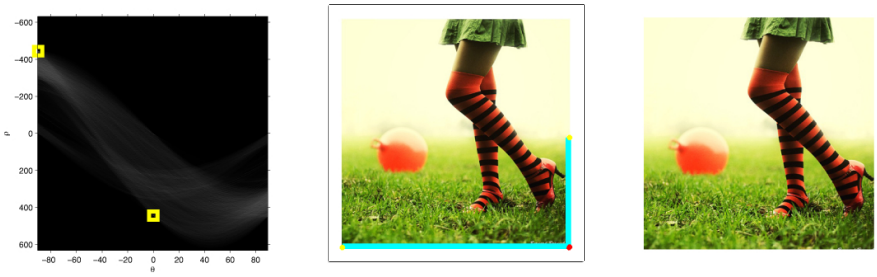

Figure 4.2: Cropping - unclear border. Left: Hough space with peaks at 0 and 90 degrees, marked by squares. Middle: Original image with lines detected by Hough transform. Due to the low contrast between the image background and the white border lines were only found on part of the image. Since more than a third of the image would be cut away when cropping along the lines, the minimum distance to the border is used for cropping. Right: Image after cropping.

Color Spaces

In this work, three color spaces are used to compute the different features: the RGB space, the L*a*b* space, and a cylindrical coordinate color representation in terms of Hue, Saturation and Lightness, which is the color representation most similar to the human perception and description of color. In color psychology, color tones and

saturation play important roles. Therefore, it makes sense to do this kind of image analysis in a color space that intuitively models these factors. Although there are many such cylindrical coordinate spaces available (HSV, HLS, HSI, etc.), it can be shown [29] that a "unified" set of cylindrical color coordinates suitable for image analysis can be derived [28]. The coordinates are calculated as follows:

$$lightness = \frac{1}{3}(R + G + B) \tag{4.1}$$

$$saturation = \max(R, G, B) - \min(R, G, B) \tag{4.2}$$

$$hue = \arctan\left(\frac{\sqrt{3}(G - B)}{2R - G - B}\right) \tag{4.3}$$

where $lightness \in [0, 1]$, $saturation \in [0, 1]$ and $hue \in [0°, 360°]$.

The main advantage of this particular color space is that it has removed the brightness dependence of the saturation. This results in a perceptively more correct Saturation channel than the HSV color space (in which the saturation is dependent on brightness). A visual comparison between the saturation channels of the two cylindrical color spaces is illustrated in Figure 4.3.

We compute this conversion for each image at the beginning of the feature extraction procedure as many features make use of this color space. The Hue, Saturation and Lightness channel of the image shown in Figure 4.3 can be seen in Figure 4.4.

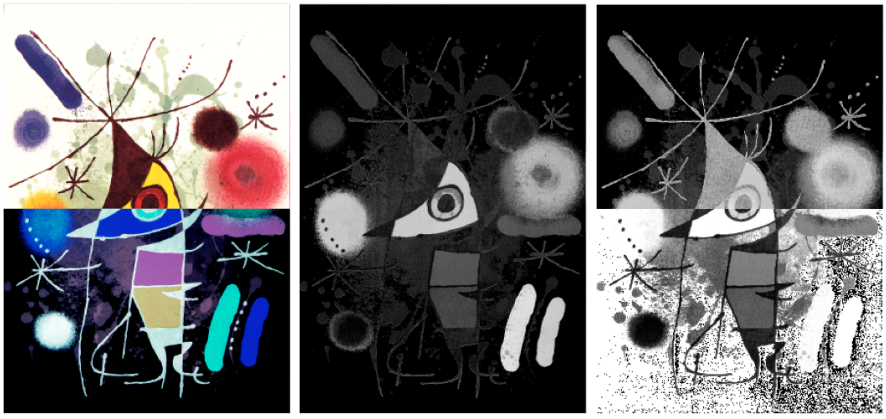

Figure 4.3: Improved HSL color space. Left: Le chanteur, Joan Mirò (bottom half inverted). Middle: Correct saturation channel computed by the Improved HSL color space. Right: For comparison - saturation channel created by the "classical" cylindrical HSV conversion.

Gray-scale images

Since we work with color in the 3D-cylindrical space, using the Hue, Saturation and Brightness channel to represent the image, we have to take special care when com-

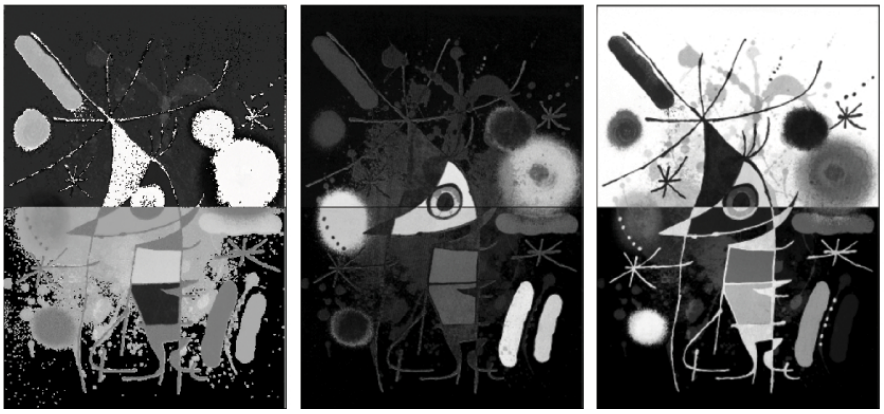

Figure 4.4: Improved HSL color space. The Hue (left), Saturation (middle) and Lightness (right) channel of the image shown in Figure 4.3.

puting features for gray-scale images. The problem with images without color is that the values in the Saturation channel are all zero and the values of the Hue channel are random. To prevent divisions by zero (when working with the Saturation channel) or random results (on the Hue channel), we set features which take these channels into account automatically to zero.
The detection of grayscale images is done by checking the original number of channels of the original image (if the original image has only 1 channel, it has to be gray-scale) and by setting a threshold on those images which have all 3 RGB channels, but contain no color. Our implementation is the following: If the maximum value of the Saturation channel is smaller than 0.1, we consider the image to be gray-scale and all smaller saturation values (if there are any) as noise.

Segmentation

Since *"harmonious composition is essential in a serious work of art"* [4], we need to consider it to analyze an image's character. As defined in [4], *composition* is *"the spatial property resulting from the arrangement of parts in relation to each other and to the whole."* To characterize the spatial organization of a picture, we use low-level image segmentation. There are many different types of image segmentation. However, instead of the popular segmentation by color clustering in CIELUV space used for example in [17], we made use of a watershed and waterfall segmentation. The advantage of the waterfall segmentation is that it takes spatial information as well as color information into account, resulting in regions that are more contiguous in comparison to color clustering.

Watershed and Waterfall segmentation

Watershed segmentation [8] uses a method from the field of mathematical morphology called the watershed transform. The basic idea behind this is to make use of the landscape analogy. Hereby, the concept of using a height-map to represent a landscape is adapted to the image analysis domain. A height map is a 2D array of values, which can be viewed as an evenly-spaced finite grid located in the (x,y)-plane. Each value represents the z height of the landscape at that point. The insight behind the Watershed transform [8] is that a greyscale image is nothing but such a 2D array of values, so it can be viewed as a landscape, where the heights are given by the grey levels in the image (see Figure 4.5).

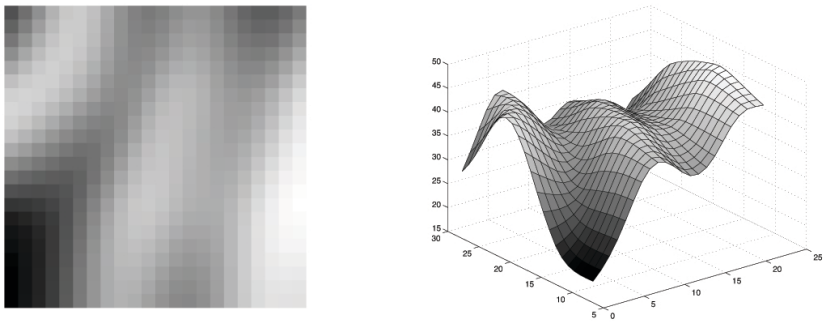

Figure 4.5: The landscape analogy. Left: 2D greyscale image. Right: 3D landscape constructed from the image on the left. The pixel values were interpreted as a height map.

To get the segmentation, a flooding process is simulated on the image landscape. Each "valley" is flooded from its local minimum. As the water rises, pools begin to form around each regional minimum. These pools are called *catchment basins*, each with it's associated minimum (see Figure 4.6). With the continuously rising water, some of the *catchment basins* will eventually meet. At these meeting points dams, or *watersheds*, are constructed to keep them apart. The flooding is continued until the whole landscape is fully "under water". The constructed dams (*watersheds*) separate the areas with the different local minima and represent the borders of the distinct regions, thus segmenting the image into a number of regions, each associated with a different regional minimum (see Figure 4.6).

However, without proper preprocessing of the input image, the Watershed approach often leads to over-segmentation of natural or noisy images because in such images most local minima are very often not associated with the objects of interest but instead only represent noise or a "irrelevant" details of a texture, etc. A general method for solving this problem involves trying to merge some of the regions together to reduce the overall number of regions in the image and to obtain a better segmentation. One algorithm which takes this approach is the Waterfall algorithm described in [45]. The Waterfall segmentation [45] is a Watershed performed considering only the

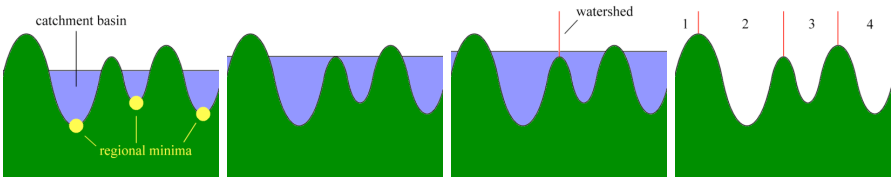

Figure 4.6: Watershed principle. (I) Regional minima and the associated catchment basins, (II) flooding of the landscape, catchment basins meet, (III) a dam (*watershed*) is constructed to keep the basins apart, (IV) final segmentation after the whole landscape has been flooded. (Images are taken from [25])

low pass points separating regions. In order get less pixel values, each region is filled with the value of the smallest pass point of its frontier (see Figure 4.7). This operation establishes a hierarchy among the frontiers produced by the Watershed. The process may be iterated until a single region covers the whole image.

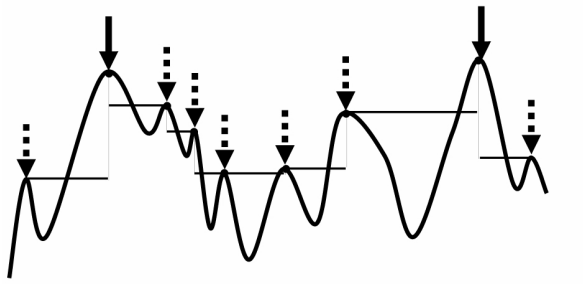

Figure 4.7: Waterfall principle. The Watershed lines are indicated by arrows and only solid line arrows will be preserved by the Waterfall algorithm. Image from [64]

In this work we use the efficient graph-based Waterfall algorithm implementation presented in [45]. It is the same implementation we used and described in [64]. The waterfall algorithm is carried out on the gradient of the color image. We use the gradient found to give the best results in a morphological waterfall segmentation in [5]. This is the saturation weighing-based color gradient applied in the cylindrical coordinate color space. This gradient gives a larger weight to the differences in hue when the saturation is high, and a larger weight to differences in luminance when the saturation is low. In order to simplify the image before segmenting it, thereby eliminating small regions, we make use of the morphological leveling [48]. The filter used to produce the marker for the leveling operator is the morphological alternating sequential filter [61], where the size of the filter refers to the number of subsequent opening and closing operations. We apply the filter separately to each color component. In this work we pre-process with a leveling of size 3 and use level 2 of the waterfall hierarchy, as

these parameters result in large contiguous regions. An example of such segmentation is shown in Figure 4.8. To get a second, even simpler and less detailed, representation of the image we compute a second segmentation at level 2 but with a bigger leveling filter of size 7. If even this filter size produces too many (more than 20) regions, we lower the filter size, but increase the level of the Waterfall to 3, producing significantly fewer image regions. We use this simplified representation to analyze each region separately as will be described later in this chapter.

Figure 4.8: Waterfall segmentation example. Left: Original image. Right: Segmented image (each region is colored by its mean color).

4.2 Features

The selection and development of useful image features is an open research topic. For each application, different features are needed to fulfill the task at hand. As already discussed in Chapter 2, the domain of affective classification is a relatively young research field, hence there are only few features that have been developed for this field. We implemented a selection of features, proven to be of use in similar image retrieval [64], analysis [47] or classification tasks [17] [75]. In Chapter 6 we evaluate the impact and usefulness of these features on the particular classification challenge of this work and the used data sets. The goal is to find features which are effective for affective classification.

Further, each section will describe one, or a group of features implemented in the course of this work.

4.3 Color Features

Colors can be (and often are) effectively used by artists to induce emotional effects. However, mapping low-level color features to emotions is a complex task which must consider theories about the use of colors, cognitive models and involve cultural and anthropological backgrounds [37] [15]. In other words, people from different cultures or backgrounds might perceive and interpret the same color pattern quite differently. The emotional impact of color and color combinations has been investigated from

the point of view of artists [37], psychology [67], color scientists [53] and marketing agents. According to Johannes Itten [37], the aesthetic problem of colors in visual art can be viewed from three different view points:

- impressive (perceptive-optical),
- expressive (psychological) and
- constructive (intellectual - symbolical).

For example, in ancient China the color yellow had a strong symbolical character and was reserved only for the emperor, the son of the sun [37], whereas an "impressive" use of this color in another culture would be to use yellow for painting a lemon, just because a lemon is perceived as yellow. The expressive use of a color would be in the sense of the phrase "I'm feeling blue". The masters of art usually used a combination of all of the three concepts to give a deeper meaning to their works of art.
Moreover, artists use color combinations unconsciously and consciously to produce optical and psychological sensations [37]. Warm colors, e.g., attract the eye and concentrate the attention of an observer more than cold colors. Cold colors on a large region can be emphasized by a contrast with a warm color or their effect can be dampened by their coupling with a highly cold tint. In a similar way, small cold color patches can emphasize large warm regions and so on.
Even though colors can be used in so many different ways, for analysis, we first need effective methods to measure colors which occur in an image. The interpretation of these measurements is then a matter of training a classifier or setting rules for the desired values of these features. Of course, better results can be achieved if the features are expressive and specific enough for the given task. In the following sections of this chapter we describe several features we developed to measure color in images.

Saturation and Brightness Statistics

The first and simplest features are the simple mean values of the image colors. We compute the mean of each channel separately, i.e. we get the values for the mean Hue, Saturation and Brightness as our first features. As was shown by Valdez and Mehrabian during their experiments in [67] pleasure relates to brightness and saturation. Darker colors are less pleasant than bright ones and higher saturation is viewed as more pleasant and more arousing.
To measure the average value of the Saturation channel, we simply use the arithmetic mean of the channel in question (X and Y are the width and height of the Saturation channel S of image I):

$$f = \frac{1}{XY} \sum_{x=1}^{X} \sum_{y=1}^{Y} I_S(x, y) \tag{4.4}$$

The value for the Brightness channel is computed correspondingly.
Although Valdez and Mehrabian [67] anticipated a strong relationship between the brightness and saturation of a color and their emotional impact on a person, the magnitudes of their effects that were indicated by the experimental results were surprising.

They found that brightness had a considerably stronger effect than saturation on the pleasure-displeasure reaction scale, whereas saturation had a much stronger effect on the arousal-calming scale. According to the experiments conducted in [67], bright and saturated colors where judged as pleasant by the candidates and less bright and more saturated colors were more arousing. These relationships are expressed by the following equations:

$$Pleasure = \quad 0.69\ Brightness \quad +0.22\ Saturation \tag{4.5}$$

$$Arousal = \quad -0.31\ Brightness \quad +0.60\ Saturation \tag{4.6}$$

$$Dominance = \quad 0.76\ Brightness \quad +0.32\ Saturation \tag{4.7}$$

We take these equations as our next features. Sample images illustrating the influence of these features are shown in Figure 4.9.

Hue Statistics

When working with 3D polar cylindrical color spaces, such as the IHLS color space presented in Section 4.1 and [29], [28], it is important to remember that the hue is an angular value. As is clearly illustrated in Figure 4.10, this presents a problem, particularly for the red and violet color tones. Even though they look similar, the reds and violets are separated by a large discontinuity in their hue values.

In cases when statistical information about the image is needed, one can use standard statistical methods for calculating the mean, standard deviation etc. for brightness and saturation. However, for hue, such calculations don't make sense. E.g. when computing the mean value of hue of an image with purely red and violet tones (such as in Figure 4.10), the result will be about 180 degrees, which represents green-blue tones. Instead, for angular data, such as the hue, the mean direction is that of the resultant vector obtained by adding all unit vectors present in the image each with their direction [46]. A measure of the variation in the directions of the data is given by the length of this vector divided by n (the mean length), which has the following characteristic: the range is [0:1] where values close to 1 mean that the data is less spread out.

Given n values of the hue H_i, the mean direction $\overline{H}$ is calculated as follows:

$$A = \sum_{i=1}^{n} \cos H_i, \qquad B = \sum_{i=1}^{n} \sin H_i, \qquad R^2 = A^2 + B^2 \tag{4.8}$$

$$\overline{H} = \begin{cases} \arctan\left(\frac{B}{A}\right) & \text{if } B > 0, A > 0, \\ \arctan\left(\frac{B}{A}\right) + \pi & \text{if } A < 0, \\ \arctan\left(\frac{B}{A}\right) + 2\pi & \text{if } B < 0, A > 0. \end{cases} \tag{4.9}$$

The mean length $\overline{R}$ is:

$$\overline{R} = \frac{R}{n} \tag{4.10}$$

The previous formulation is standard in the texts on circular statistics, but it ignores the fact that not all hues have the same importance. If we take this into account

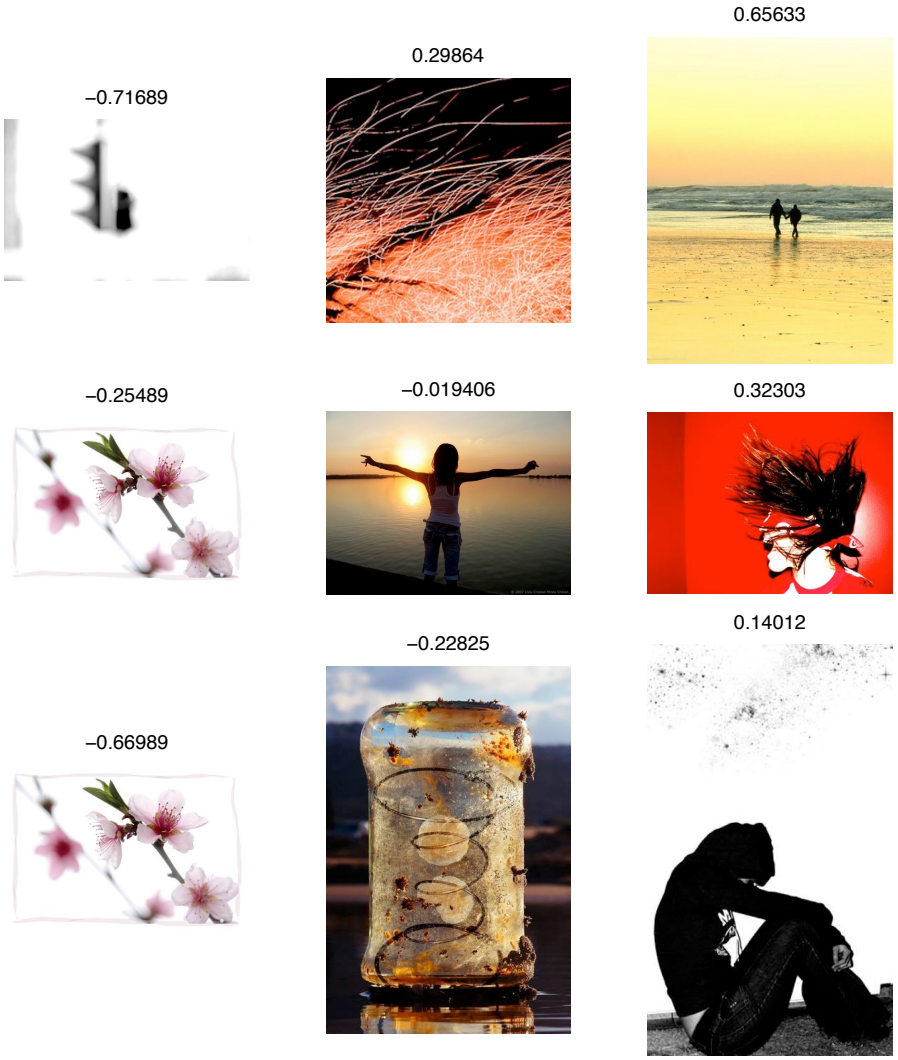

Figure 4.9: Sample images for the emotion equations by Valdez and Mehrabian [67]. Top to Bottom: Pleasure, Arousal and Dominance. Left to Right: Image with minimal, median and maximal value.

Figure 4.10: The hue problem. Left: Color image. Middle: Hue component. Right: Hue histogram.

by weighting the length of each hue vector by the associated saturation value, we get the following formula.

Let S_i be the saturation associated with hue H_i. We replace Formula 4.8 with:

$$A_s = \sum_{i=1}^{n} S_i \cos H_i, \qquad B_s = \sum_{i=1}^{n} S_i \sin H_i, \qquad R^2 = A_s^2 + B_s^2 \tag{4.11}$$

To calculate the saturation-weighted mean direction $\overline{H_s}$, we replace A and B by A_s and B_s in Equation 4.9. Equation 4.10 becomes:

$$\overline{R_s} = \frac{R_s}{\sum_{i=1}^{n} S_i} \tag{4.12}$$

Note that $\overline{R_s}$ remains a measure of the angular dispersion, and does not give information on the mean of the saturation.

Rule of Thirds

A popular rule of thumb when composing pictures is the Rule of Thirds. The rule can be considered as an approximation of the 'golden ratio'. It states that the main element or center of interest in an image should be placed along one of the lines that divide the image into thirds, or even better at its intersections, as illustrated in Figure 4.11. In Figure 4.11 two typical exemplars where this rule was applied are shown. In the case of landscape photography the horizon is placed along one of the dividing lines, and in the case of portraits the eyes are placed along the intersections of the dividing lines. In both cases the main object of interest stretches from the intersections to the center of the image, implying that a large part of the main object lies inside the inner rectangle or on its periphery. Based on these observations, we also compute the average Hue, Saturation and Brightness of the inner rectangle respectively, as was suggested in [17] (X and Y are the width and height of the Saturation channel S of image I):

$$f = \frac{9}{XY} \sum_{x=X/3}^{2X/3} \sum_{y=Y/3}^{2Y/3} I_S(x, y) \tag{4.13}$$

This is computed analogously for the Brightness channel. For the Hue channel we use the vector-based mean function that was mentioned above in Equation 4.11 computed over the inner rectangle.

Figure 4.11: The Rule of Thirds.

Colorfulness

As proposed in [17] we also compute the relative color distribution, distinguishing multi-colored images from monochromatic, sepia or simply low contrast images. We use the Earth Movers Distance (EMD) [59], which is a measure of similarity between any two weighted distributions. As suggested by Datta [17] we divide the RGB color space into 64 cubic blocks with four equal partitions along each dimension, taking each such cube as a sample point. We create two distributions, an "ideal" distribution D_1 and the "real" sample distribution D_2, which is computed from the given image by finding the frequency of occurrence of color within each of the 64 cubes. Distribution D_1 is generated as the color distribution of a hypothetical image such that for each of 64 sample points, the frequency is $\frac{1}{64}$, i.e. each color is present in equally. The EMD measure requires that the pairwise distance between sampling points in the two distributions (D_1 and D_2) be supplied. Since the sampling points in both of them are identical, the distance can be defined as the pairwise Euclidean distances between the geometric centers c_i of each cube i, after conversion to LUV space. We use the LUV color space because of its perceptual uniformity, so the distance between the colors will represent the perceived difference in color. Therefore the colorfulness measure is as follows:

$$colorfulness = emd(D_1, D_2, \{d(a,b)|0 \leqslant a, b \leqslant 63\}), \tag{4.14}$$

$$where \;\; d(a,b) = \|rgb2luv(c_a) - rgb2luv(c_b)\| \tag{4.15}$$

Color Names

Specific colors communicate different meanings. As mentioned before some colors have symbolical characters in some cultures or contexts. There are, however, some common rules of thumb which are often exploited, mainly by producers and designers of commercials, that state how a certain color affects a human observer on the emotional level. According to these, red communicates happiness, dynamism and power.

Orange is the warmest color, it resembles fire and thus communicates glory. Green induces calmness and relaxation and is the color of hope. Blue, a cold but very popular color, improves the dynamism of warm colors or suggests gentleness, fairness, faithfulness and virtue. Purple is a melancholy color and sometimes communicates fear. Brown generally serves as background color for relaxing scenes [15].
From the experiments in [67] we also know that when viewed without context, blue, blue-green, green, pink and purple are perceived as significantly more pleasant than green-yellow, yellow and orange, whereby yellow received the most unpleasant ratings, especially when it was a dark tone. Concerning the arousal ratings, the only noteworthy generalization was that green hues (green-yellow, blue-green and green) elicited the highest arousal reactions from the subjects.
Color histograms were one of the first features used in content-based image retrieval and they have proven to be useful for many tasks in image retrieval. In order to allow a more semantically based approach, we measure the amount of each of the basic colors by using a *color names histogram* based on work by Van de Weijer et al. [68]. In English, eleven basic color terms have been defined based on a linguistic study. These are: black, blue, brown, green, grey, orange, pink, purple, red, white, yellow. Van de Weijer et al. [68] used images downloaded from the Internet to learn the mapping of these color names to RGB coordinates, thereby creating a lookup table of each RGB coordinate to one of eleven color names. An example of the mapping of RGB coordinates to color names is shown in Figure 4.12. However, even slight differences

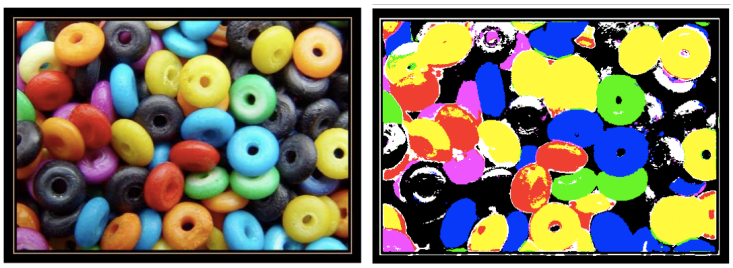

Figure 4.12: Color names. Left: Original image. Right: Mapping to eleven basic color names.

in color, such as can be caused by e.g. lighting changes in photography, can have a noticeable effect on the resulting naming of the colors. This is particularly noticeable for faces. An example of this effect can be seen in Figure 4.13.

Itten Contrasts

Johannes Itten [37], who studied the usage of color in art extensively, introduced a formalism to analyze the use of color in art and the effects that this induces on the viewer's psyche. In his work [37] he used a simplified color model, on which he defined several rules and seven color contrasts used in art images to induce psychological reactions in viewers. He recognizes that color perception is a complex task, where not only the color itself is important, but also its surroundings, position, context, cultural

Figure 4.13: Color names. Examples showing the mapping of color to eleven basic color terms. Images (a) and (c) are original photos, (b) and (d) show the reduced colors by mapping to basic colors.

background etc. are important. Also the famous painter, Picasso confirmed that "actually painters work using only a few colors. Perception gives the idea that there are many if they are in the right position on the canvas" [16].
Itten's color model characterizes colors according to *hue*, *saturation* and *luminance*. Twelve hues are identified as fundamental colors. These are composed of the 3 "primary" colors (red, yellow, blue), 3 "complementary" colors (orange, green, purple) and 6 "tertiary" colors obtained by a linear combination of the "primary" and "complementary" colors (see Figure 4.15). Further, the 12 fundamental hues are varied by five levels of luminance and three levels of saturation. This results in 180 distinct colors, which have been organized into a spherical representation that can be seen in Figure 4.14. The 12 pure colors are located along the equatorial circle, luminance

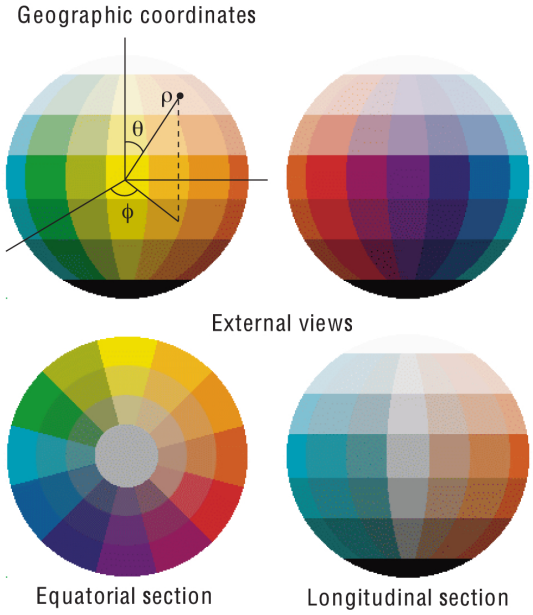

Figure 4.14: Itten's spherical color model. Image from [37] labeled by [15].

varies along the meridians and saturation increases as the radius grows. The center of the sphere is neutral gray, perceptually contrasting colors lie opposite each other with respect to the center. Warm colors lie opposite cold colors, dark colors opposite light colors, etc. Using this polar representation, Itten identified the following seven types of contrasts:

contrast of saturation - formed by different levels of saturation in neighboring regions,

contrast of light and dark - formed by differences in brightness,

contrast of extension - proportion of the sizes of the color patches in relation to the visual "weight" of their color,

contrast of complements - formed by visually complementary colors - colors that are opposite each other on the color wheel,

contrast of hue - formed by differences in hue, the maximal contrast of hues is the complementary contrast,

contrast of warm and cold - formed by the combinations of colors that are considered "warm" and "cold",

simultaneous contrast - when contrasting colors are missing, a special phenomenon is created by the human eye, which causes boundaries of the color patches to "vibrate", creates color illusions or invokes a sense of tension.

These combinations are used unconsciously of consciously by artists in order to produce optical (*impressive*) and psychological (*expressive*) sensations [37], [16].
In [16] an attempt was made to translate several of Itten's theories into a formal language (see our summary of their work in Chapter 2). We adapted some of the formulas from [16] to create features that would give a numerical representation of these concepts. As in [16], the image is segmented into regions (as described above) and some intra-region features in form of membership functions expressing such characteristics as saturation, hue, brightness, warmth, region size and position are computed for each region, transforming the image into the Itten color model. In other words, each region is assigned a given set of membership functions (which is the same for all regions). Each membership function represents one possible characteristic of the region, e.g. the occurrence of dark shadows, and each membership function is assigned a value that expresses the proportion of "member" pixels in the given region that share this characteristic, i.e. if the given region is very light, the membership function for dark shadows will have a very low value, but the membership function of light will have a high value... The use of membership functions here is the same as in the next section, which describes the color features by Wang Wei-ning et al. [75]. In fact, when we were looking for clear definitions of how to assign e.g. warm and cold colors, we found the publication [75], which contains clear definitions of the membership functions we were looking for. While implementing those parts that we needed for the Itten features, we decided to implement also the features as suggested Wang Wei-ning et al. [75] for the sake of comparison of our works. Therefore we describe the Wang

features in the next section, but refer to parts of them in this section.
To measure the *contrast of light and dark* over the whole image, we use standard deviation over the brightness membership functions of all regions, as suggested in [54]. We define the measurement for the *contrast of saturation* in an analogue fashion. For the *contrast of hue* we utilize the vector based measurement of the hue spread as defined in Equation 4.12.

The contrast of complements is measured by computing the differences of hues between the regions. In our implementation, we use both, the average of all differences between all region pairs, and the maximum value among those differences. However, since we have to consider the hue-wheel problem, we use $d = \min(|h_i - h_j|, 360 - |h_i - h_j|)$ as hue difference measure (where h_i is the representative (mean) hue of the region i). In case of a real contrast of complements, the value should be near 180°.

The contrast of warm and cold is defined in [16] according to "temperature" differential of colors. Since each region is assigned three membership functions w_c that express the degree of cold ($t = 1$), neutral ($t = 2$) and warm ($t = 3$) in the region r_i (we chose the same classification as in Figure 4.17, with neutral being $1 - (warm + cold)$), we can define the strength of the contrast between two regions in terms of perceived warmth as:

$$\text{contrast of warmth} = \frac{\sum_{t=1}^{3} w_t(r_1) w_t(r_2)}{\sqrt{\sum_{t=1}^{3} (w_t(r_1))^2 \sum_{t=1}^{3} (w_t(r_2))^2}} \tag{4.16}$$

Again we take the average of all region pair contrasts as measurement of the whole image, as well as the maximum value. To exploit this even more we also measured the total amount of warm and cold area in the image respectively. The simultaneous contrast is basically the absence of contrast of complements, i.e. when the value of the complementary contrast is low. We don't compute the contrast of extension, due to insufficient understanding of its definition and difficulties in finding general rules that could be well formulated in a mathematical sense.

Additionally, Itten [37] discusses the concept of the *harmonic* arrangement of colors in terms of accordance of colors. The combination of hues and tones that generates a stability effect on the human eye is said to be *harmonic*. *Harmony*, in this context, is an objective concept defined as the combination of those colors whose mix is gray. Applied to the spherical model from Figure 4.14, the color accordances that create *harmony* are those color combinations that generate a regular polygon, when their locations on the sphere are connected by lines as shown in Figure 4.15.

To compute *harmony*, we first determine the main hues in the image. For this we implement the Hue count from [40]: Because the human eye doesn't recognize hues in very bright, dark or unsaturated areas and the hue values in these areas are more or less random, we consider only pixels which have a brightness value in the range of $[0.15, 0.95]$ and saturation $s > 0.2$. From the pixels that satisfy these conditions, a 12-bin Hue histogram H is built. However, there usually will still be some noise in the histogram. To reduce it, we use the following threshold:

$$N = \{i \mid H(i) > \alpha m\} \tag{4.17}$$

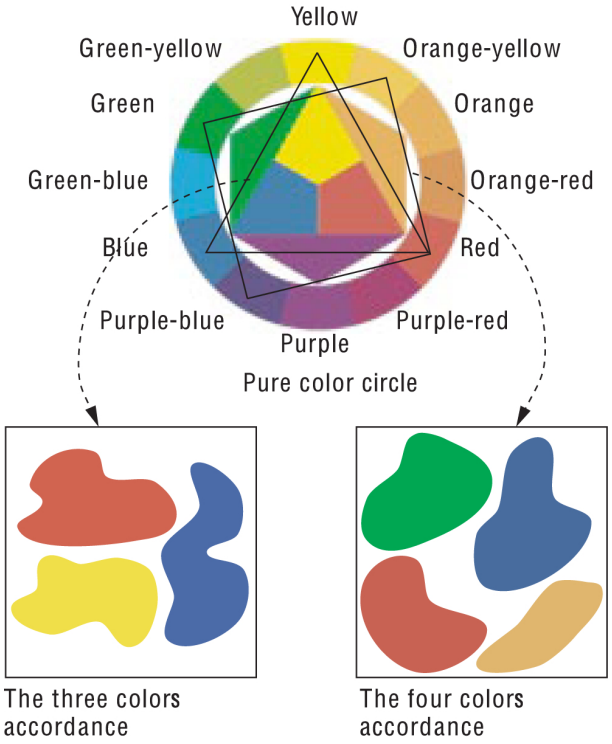

Figure 4.15: The concept of color accordance. Image from [37] labeled by [15].

where N is the resulting set of bins with values higher than αm, m is the maximum value of the original histogram H and α is a control parameter, which we set to $\alpha = 0.05$ as recommended in [40].
We map these n remaining hues onto the Itten color wheel and connect their positions to generate a polygon. Then we compute the internal angles of the generated polygon using the cosine law. Image *harmony* is defined by colors that generate a regular polygon on the sphere. Therefore we compute the regularity of the generated polygon by comparing its internal angles to those of a hypothetical regular polygon that would be generated with the same number of vertices n. The internal angles of a regular polygon are all the same and can be calculated as follows:

$$internalAngle = (n-2)\frac{180}{n} \tag{4.18}$$

The smaller the difference between the real angles to the "ideal" regular angles, the bigger the *harmony*.

Color features by Wang Wei-ning et al. [75]

Based on theory of color, psychology and statistical observations from their test set of images, Wang Wei-ning et al. [75] proposed a set of color features for emotional semantics retrieval. Through factor analysis they transformed their test set of images which were labeled by scores on 12 impression word pairs into a 3-dimensional emotional factor space (we already summarized this paper in Chapter 2). For each of the three emotional factors, they defined a set of low level color features, designed to express the general differences between the images of the three factors well.
To achieve this in a semantically sensible fashion that is usable for humans, semantic words are used to describe the measurement of the features and a fuzzy membership function is assigned to each word to express the fuzziness of the human language.
The features of the first emotional factor should have the ability to describe the lightness and warmth or coolness of images. Five semantic words are used for the lightness function: *very dark, dark, middle, light and very light*. An unsupervised algorithm is used to find fuzzy functions that would reflect the distribution of the data. The resulting function (which we also implemented) is shown in Figure 4.16. For example, if a pixel's lightness value is 68, its membership function for *very dark,*

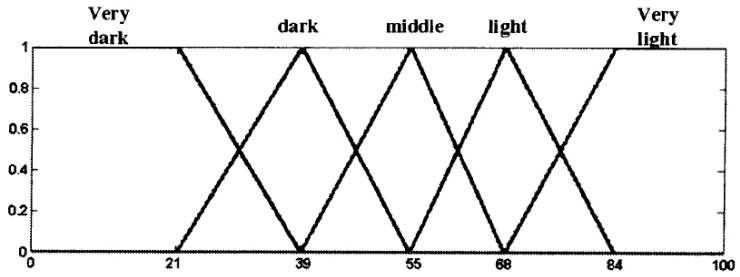

Figure 4.16: Membership function of lightness. Image from [75].

dark, middle, light and very light will be $[0, 0, 0.25, 0.75, 0]$. This membership is computed for every pixel in the image.

Referring to a study in [53], the membership function for *warm-cool* is defined as follows (h stands for a value of the Hue channel):

$$warm = \begin{cases} \cos(h - 50°) & \text{if } 0° \leqslant h < 140° \text{ or } 320° \leqslant h \leqslant 360°, \\ 0 & \text{else} \end{cases} \tag{4.19}$$

$$cold = \begin{cases} \cos(h - 230°) & \text{if } 140° \leqslant h < 320°, \\ 0 & \text{else} \end{cases} \tag{4.20}$$

The membership function for warm-cold is also shown in Figure 4.17, where the solid line stands for the *warm* membership and the dashed line represents *cold*.

To get the feature vector for the first emotional factor, the membership functions are computed for all pixels in the image and combined into a 10-dimensional

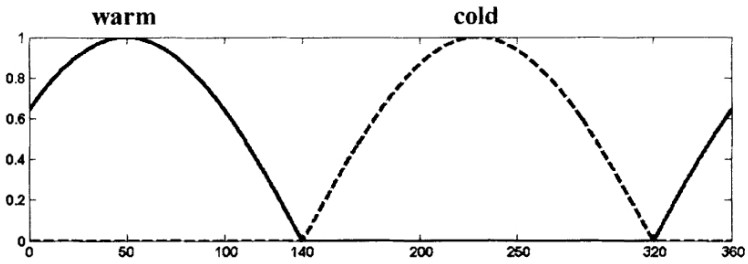

Figure 4.17: Membership function of warm-cold. Image from [75].

histogram with the ability to describe the lightness and warm-cool content of an image by measuring the 10 combinations: “very dark - warm”, “very dark - cold”, “dark - warm”, “dark - cold” etc.

The second emotional factor in [75] was best expressed as a combination of saturation, warm-cold colors and color contrast. Again a fuzzy function was defined to measure saturation, this time using the three expressions *low saturation, middle saturation and high saturation*, which proved to be sufficient as humans are not very sensitive in distinguishing different levels of saturation. The membership function is shown in Figure 4.18.

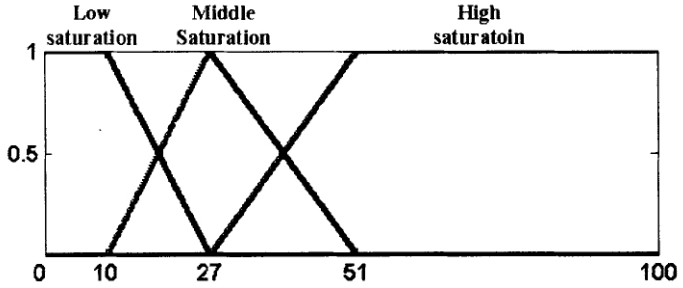

Figure 4.18: Membership function of saturation. Image from [75].

The contrast of an image is defined in [75] in the L*a*b* space (where a_i^* and b_i^* represent the values of a* and b* channels, $\overline{a^*}$ and $\overline{b^*}$ are the average values of the whole image, and N is the number of pixels of the image):

$$contrast = \left[\frac{1}{N-1} \sum_{i=1}^{N} (a_i^* - \overline{a^*})^2 + (b_i^* - \overline{b^*})^2 \right]^{\frac{1}{2}} \tag{4.21}$$

As in the first emotional factor, the saturation, warm-cold (as defined above) and contrast are computed over the whole image and combined to get a 7-dimension vector to express the second emotional factor.

For the third factor, contrast and the sharpness of the edges was found to be a representative vector. To measure sharpness, the average of the top 5% of the gradient values of the image are taken. Together with the contrast as it was defined above, a 2-dimensional vector is the representative of the third emotional factor.
Together, these three histograms give a meaningful representation of the color composition of the image from an affective point of view.

We implemented these features because they seem to measure the color distribution of an image for the purpose of affective analysis and retrieval well and so they might be useful for our purpose, and also for the sake of comparison.

4.4 Texture Features

The textures in images are also important for the emotional expression of an image. Professional photographers and artists usually create pictures which are sharp, or where the main object is sharp with a blurred background. However we observed that also unsharpness in pictures can be used efficiently to achieve a desired expression. Purposefully blurred images were frequently present in the category of art photography images which expressed fear. This technique seems to be effective in this genre and potentially has its origins in the commonly strong fear of the unknown. A dark, blurred image leaves much to the imagination of the observer, who can project his/her own fears onto the unclear image in an atmosphere of suspense. Moreover, certain types of texture in combination with unsaturated or strongly red colors commonly induce disgust. On the other hand, simple pictures with pleasant, smoothly colored regions and only a few clear edges are often perceived as calming and peaceful, and images with many textures may indicate landscape-like pictures which are often shot with a high aperture. Many features for describing texture have been developed [62]. We have chosen some of the commonly used.

Wavelet-based textures

One way to measure spatial smoothness/graininess in images is to use the Daubechies wavelet transform [19]. As suggested in [17], we perform a *three-level* wavelet transform on all three color channels, Hue I_H, Saturation I_S and Brightness I_B. An example of a two-level wavelet transform of a monochrome image can be seen in Figure 4.19, with the horizontal, vertical and diagonal coefficients denoted as H, V and D respectively. Denoting the coefficients in level i for the wavelet transform of the one channel of an image as w_i^h, w_i^v and w_i^d, the wavelet features are defined as follows (where $i = \{1, 2, 3\}$):

$$wavelet_i = \frac{\sum_x \sum_y w_i^h(x,y) + \sum_x \sum_y w_i^v(x,y) + \sum_x \sum_y w_i^d(x,y)}{|w_i^h| + |w_i^v| + |w_i^d|} \tag{4.22}$$

This is computed for every level i and every channel (I_H, I_S and I_B) of the image, i.e. we get 9 wavelet features. Furthermore, we add three more by computing a sum over all three levels for each of the channels I_H, I_S and I_B.

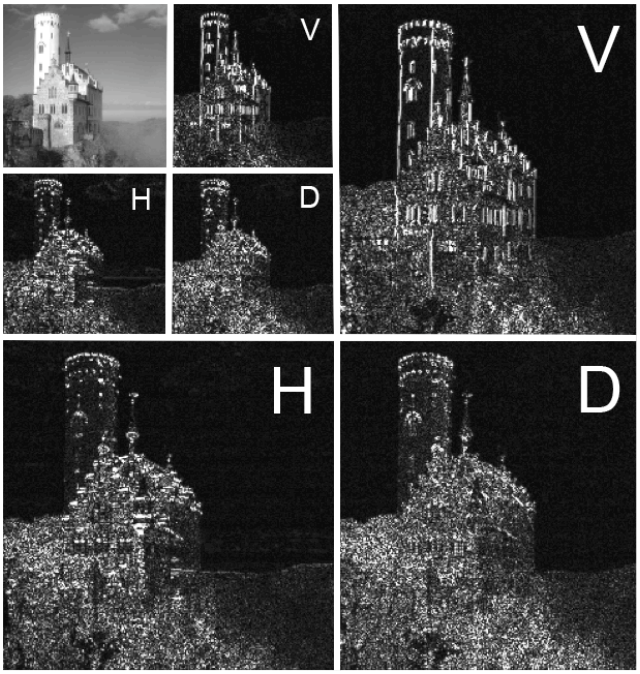

Figure 4.19: Two-level wavelet transform.

Tamura features

Tamura et al. texture features [65] were successfully used in the field of affective image retrieval by Wu et al. [56]. Therefore we decided to use the first three of the Tamura texture features: *coarseness*, *contrast* and *directionality*.
According to [65] **coarseness** is computed in 4 steps:

Step 1: Compute the averages at every point over neighborhoods whose sizes are 2^k, e.g. 1×1, 2×2,..., 32×32. The average over a neighborhood at the point (x, y) is

$$A_k(x,y) = \sum_{i=x-2^{k-1}}^{x+2^{k-1}-1} \sum_{j=y-2^{k-1}}^{y+2^{k-1}-1} f(i,j)/2^{2k} \tag{4.23}$$

where $f(i, j)$ is the gray-level value at (x,y).

Step 2: For each point, take differences between pairs of averages corresponding to non-overlapping neighborhoods, in both, horizontal and vertical directions. For example the horizontal case is

$$E_{k,h}(x,y) = |A_k(x + 2^{k-1}, y) - A_k(x - 2^{k-1}, y)|. \tag{4.24}$$

Step 3: At each point, pick the best size which gives the highest output value:

$$S_{best}(x,y) = 2^k \tag{4.25}$$

where k maximizes E in either direction, i.e.,

$$E_k = E_{max} = max(E_1, E_2, ..., E_L). \tag{4.26}$$

Step 4: Take the average of S_{best} over the picture to be a coarseness measure:

$$coarseness = \frac{1}{m \times n} \sum_{i}^{m} \sum_{j}^{n} S_{best}(i, j) \tag{4.27}$$

where m and n are the width and height of the image, respectively.

Contrast is defined via a measure of polarization, the kurtosis α_4 that can be defined as:

$$\alpha_4 = \frac{\mu_4}{\sigma^4} \tag{4.28}$$

where μ_4 is the forth moment about the mean and σ^2 is the variance about the mean of the gray-level probability distribution. To measure the contrast Tamura combines σ and α_4 as follows:

$$contrast = \frac{\sigma^4}{(\alpha_4)^n} \tag{4.29}$$

where $n = \frac{1}{4}$.

To compute **directionality**, a histogram of local edge probabilities is created first by quantizing the magnitude and direction of the image gradients and then computing the sharpness of the peaks of the edge histogram. In detail, it is defined as follows, where the magnitude of the image gradient is $|\triangle G|$ and the local edge direction is θ:

$$|\triangle G| = \frac{(|\triangle_H| + |\triangle_V|)}{2} \tag{4.30}$$

$$\theta = \tan^{-1}\left(\frac{\triangle_V}{\triangle_H}\right) + \frac{\pi}{2} \tag{4.31}$$

where $\triangle_H$ and $\triangle_V$ are the horizontal and vertical differences measured by the following two 3×3 operators:

$$\begin{pmatrix} -1 & 0 & 1 \\ -1 & 0 & 1 \\ -1 & 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ -1 & -1 & -1 \end{pmatrix} \tag{4.32}$$

The resulting θ is a real number ($0 \leqslant \theta < \pi$). The histogram H_D is then obtained by quantizing θ and counting the points with the magnitude $|\triangle G|$ over the threshold t:

$$H_D(k) = \frac{N_\theta(k)}{\sum_{i=0}^{n-1} N_\theta(i)}, \qquad k = 0, 1, ..., n-1 \tag{4.33}$$

where $N_\theta(k)$ is the number of points at which $(2k-1)\pi/2n \leqslant \theta < (2k+1)/2n$ and $|\triangle G| \geqslant t$. In our case $n = 16$ and $t = 12$.

To measure the directionality quantitatively from H_D we compute the sharpness of the peaks:

$$directionality = 1 - r \cdot n_p \cdot \sum_{p}^{n_p} \sum_{\phi \in w_p} (\phi - \phi_p)^2 \cdot H_D(\phi) \tag{4.34}$$

where n_p is the number of peaks, ϕ_p is the pth peak position of H_D, w_p is the range of the pth peak between valleys, r is the normalizing factor, ϕ is the quantized direction code (cyclically in modulo 180°). We did not consider more than two peaks, i.e. we only test whether n_p is two. $n_p = 2$ if:

$$H_D(v_{12})/H_D(\phi_2) < 0.5, \tag{4.35}$$

$$H_D(v_{21})/H_D(\phi_2) < 0.5, \tag{4.36}$$

$$H_D(\phi_2)/H_D(\phi_1) > 0.2 \tag{4.37}$$

it is $n_p = 1$ otherwise. Where v_{21} and v_{12} are positions of the valleys from the first peak ϕ_1 to the second peak ϕ_2 and vice versa.

Gray-Level Co-occurrence Matrix

Another classical method of measuring texture is by means of the gray-level co-occurrence matrix (GLCM), introduced by Haralick et al. [32], [33]. The GLCM is created by calculating how often a pixel with gray-level (grayscale intensity) value i occurs horizontally adjacent to a pixel with the value j. Hereby the pixel spatial relationships can be specified by the user. Each element (i, j) in the GLCM specifies the number of times that the pixel with value i occurs horizontally adjacent to a pixel with value j. After the GLCM is created, it is normalized so that the sum of its elements is equal to 1. Each element (r, c) in the normalized GLCM is the joint probability occurrence of pixel pairs with a defined spatial relationship having gray level values r and c in the image [66]. From such a normalized GLCM, several statistical properties can be extracted. We chose to extract:

contrast - a measure of the intensity contrast between a pixel and its neighbor over the whole image, defined as $contrast = \sum_{i,j} |i - j|^2 p(i, j)$,

correlation - a measure of how correlated a pixel is to its neighbor over the whole image, defined as $correlation = \sum_{i,j} \frac{(i-\mu_i)(j-\mu_j)p(i,j)}{\sigma_i \sigma_j}$, where μ is mean and σ is standard deviation,

energy - the sum of squared elements in the GLCM, defined as $energy = \sum_{i,j} p(i, j)^2$,

homogeneity - a value that measures the closeness of the distribution of elements in the GLCM to the GLCM diagonal, defined as $homogeneity = \sum_{i,j} \frac{p(i,j)}{1+|i-j|}$.

4.5 Composition Features

We already stated that harmonious composition is essential in a serious work of art and we need to consider it to analyze an image's character. It is by definition the art or practice of so combining the different parts of a work of art as to produce a harmonious whole [4]. There is much potential in exploiting this area of analyzing the spatial relations between the parts of the image. Since such relations tend to be rather complex, we analyze only three aspects of composition, but see this as an area where much improvement could be done in future.

Level of Detail

Simple, elegant shapes, few colors, large monotonous regions or soft textures in a minimalist composition have a calming effect on humans and are mostly perceived as pleasant. On the other hand, busy structures with strong contrasts create tension or excitement. Images with much detail certainly produce a different psychological effect than minimalist compositions. To measure the level of detail of an image we count the number of regions that result from waterfall segmentation with a predefined filter size (filter size 3 and level 2 of the waterfall hierarchy). For simple images, this number will be low whereas busy images might produce even three-digit numbers. Examples are shown in Figure 4.20.

Low Depth of Field

Professional photographers often reduce the depth of field (DOF) for shooting single objects by using larger aperture settings, macro lenses, or telephoto lenses. DOF is the range of distance from a camera that is acceptably sharp in the photograph. By reducing the DOF, the photographer causes the background to blur, thus simplifying the image, reducing the busyness and drawing the attention of the observer to the object of interest, which is sharp. In [17] Datta et al. propose a method to detect low DOF and macro images. The image is divided into 16 rectangular blocks $\{B1, ..., B16\}$, numbered in row-major order, as shown in Figure 4.21. Then, the wavelet coefficients, as described earlier in this chapter, are used. Let $w_3 = w_3^h, w_3^v, w_3^d$ denote the set of wavelet coefficients in the high frequency (level 3 as used in the notation of Equation 4.22) of one of the image channels (I_H, I_S or I_B). Then the *low depth of field indicator* is computed for each channel separately as follows:

$$lowDOF = \frac{\sum_{(x,y)\in B_6\cup B_7\cup B_{10}\cup B_{11}} w_3(x,y)}{\sum_{i=1}^{16}\sum_{(x,y)\in B_i} w_3(x,y)} \tag{4.38}$$

This means that the ratio of the high frequency wavelet coefficients of the 4 inner rectangles against the wavelet coefficients of the whole image is computed. Since the object of interest is usually in sharp focus near the center, while the surrounding is usually out of focus, this essentially means that a large value of the *low DOF indicators* tend to occur for macro shots.
An alternative option to measure the area of focus more precisely using deconvolution is presented by Kovacs and Sziranyi in [42].

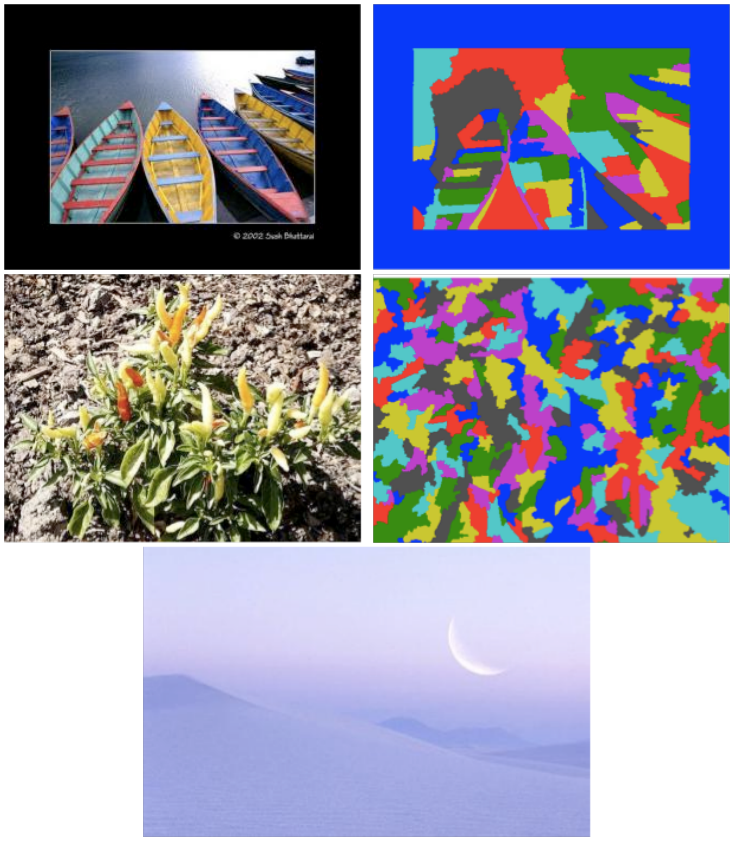

Figure 4.20: Level of Detail. Left: original images. Right: Their segmentations. The *level of detail* of the first row is a two-digit number, second row has a three-digit number as *level of detail*. The image in the third row has *level of detail* = 1.

Dynamics

Studies suggest that lines induce emotional effects [37], [6]. Horizontal lines are always associated with a static horizon and communicate calmness, peacefulness and relaxation, vertical lines are clear and direct and communicate dignity and eternality, slant lines, on the other hand, are unstable and communicate dynamism. Lines with many different directions present chaos, confusion or action. The longer, thicker and more dominant the line the stronger the induced psychological effect. As we already discussed at the beginning of this chapter in the context of automated cropping, detecting significant line slopes in images can be accomplished by using the Hough transform [26]. Hereby, the image is transformed into Hough space, where significant lines form peaks that can be detected (see the example in Figure 4.1). The position of the peak in Hough space gives information about the position and tilt angle of the line slope. We use this information to generate a line slope histogram. An example

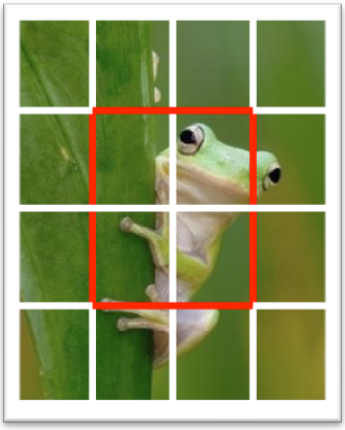

Figure 4.21: The division of the image for the low depth of field indicator.

is shown in Figure 4.22. The found lines are classified into static (horizontal and vertical) or slant according to their tilt angle θ and weighted by their respective lengths. A line is classified as static if $(-15° \leqslant \theta < 15°)$ or if $(75° \leqslant \theta < 105°)$ and as slant otherwise. As a result we get the proportion of static and dynamic lines in the image.

Figure 4.22: Line slopes histogram. Left: Original video frame. Middle: Edge image. Right: Line slopes histogram. Images from [15].

A different way of classifying static vs. dynamic images is presented in [76] and should be considered for applications where the runtime is important, as it potentially performs faster than the Hough transform.

4.6 Content Features

The semantic content of the image has the greatest impact of the emotional influence of any picture. For example, if we consider the two images shown in Figure 4.23, there is no significant difference in colors or textures. The tones of colors on both pictures suggest a pleasant, though cold picture. Even with all the rules of usage of colors in art, differentiating those two is difficult. However, every human would know

at first glance that the picture of the tiger looks anything but peaceful, in contrast to the swimming lady, who looks quite content. This example clearly illustrates that

Figure 4.23: The content of an image is important. Two formally similar images, but with different emotional impact.

algorithms that would recognize the semantic content of a picture would also be of benefit in this area of image retrieval. However, the analysis of semantic content of images is an open research area and the algorithms that are being developed are complex and go beyond the scope of this work. Nevertheless, we included two such algorithms that were available to us.

Human Faces

The presence and size of a human face in the image produces a strong emotional response in many people. Although the expression of the face is very important in order to distinguish between the moods of a picture, algorithms that can recognize the emotional expression of a human face in static images are very hard to find and very complex. However, we can at least detect frontal faces (if there are any) on the picture by using the popular and widely available face detection algorithm by Viola and Jones [70], [71]. For each image we use the number of found faces and the relative size of the biggest face with respect to the image. This way, we can at least distinguish pictures with people from nature, landscape or object photography and portraits from group shots.

Skin

Finally, we attempt to detect human skin. This should be especially valuable if we consider the popular group of pictures labeled as "artistic nudes" and which generally have a specific emotional response. For this we use the algorithm presented in [14], which we adapted to static images. The basic idea is to find the color spectrum that represents skin color in an image. The *YCbCr* color space is best suited for this task as there exists a predefined static model that represents skin color well in many cases.

The conversion from RGB to *YCbCr* is defined as follows:

$$Y = (0.299 * (R - G)) + G + (0.114 * (B - G)) \quad (4.39)$$

$$Cb = (0.564 * (B - Y)) + 128 \quad (4.40)$$

$$Cr = (0.713 * (R - Y)) + 128 \quad (4.41)$$

The static skin color model is defined as a spectrum between the following thresholds: $Cb_{min} = 77$, $Cb_{max} = 127$, $Cr_{min} = 133$, and $Cr_{max} = 173$, the Y component is ignored. In the general case, pixels in this spectrum are counted as skin.

However, the authors of [14] introduced an improvement to this method by including face detection. They utilize the face detection algorithm by Viola and Jones [70], [71] to find faces in the image. If a face is found, the above model is altered to present a spectrum specific to the person found in the image. The new skin model is defined by taking the rectangle defining the position and size of the face, reducing it by 30% from each side to ignore potential hair or parts of the background that might be inside the rectangle, and measuring the average skin pixel color inside the rectangle. This measurement is taken as the new median of skin color. The new model is then defined as a range of 30% of the *Cb* channel and 17,5% of the *Cr* channel around the median skin color.

If more than one face is detected, the present skin color models are combined.

We compute the area of skin (i.e. the number of pixels in skin color) and the proportion of the "skin area" to the size of the detected face as features.

4.7 Summary

The following table summarizes the features we compute and which are all described in this chapter. The column # gives the length of the feature vector for each feature.

category	short name	#	short description
color	*Saturation, Brightness*	2	mean saturation and brightness
	Pleasure, Arousal, Dominance	3	approx. emotional coordinates based on brightness and saturation
	Hue	4	vector based mean hue and angular dispersion, saturation weighted and without saturation
	Rule of Thirds	3	mean saturation, brightness and hue of the inner rectangle
	Colorfulness	1	colorfulness measure based on EMD
	Color Names	11	amount of black, blue, brown, green, grey, orange, pink, purple, red, white, yellow

	Itten	20	average *contrast of brightness, contrast of saturation, contrast of hue, contrast of complements, contrast of warmth, harmony, hue count, hue spread, area of warm, area of cold,...*and the maximum of each
	Wang	19	features (histograms) by Wang Wei-ning et al. [73] (*factors 1 (10), factor 2 (7) and factor 3 (2)*)
	Area statistics	10	based on Wang features: *area of very dark, area of dark, area of middle, area of...light, very light, high saturation, middle saturation, low saturation, warm, cold*
textures	*Tamura*	3	features by Tamura et al [65].: *coarseness, contrast, directionality*
	Wavelet textures	12	wavelet textures for each channel (Hue, Saturation, Brightness) and each level (1-3), sum of all levels for each channel
	GLCM-features	12	features based on the GLCM: *contrast, correlation, energy, homogeneity* for Hue, Saturation and Brightness channel
composition	*Level of Detail*	1	number of segments after waterfall segmentation
	Low Depth of Field (DOF)	3	low depth of field indicator; ratio of wavelet coefficients of inner rectangle vs. whole image (for Hue, Saturation and Brightness channel)
	Dynamics	6	Line slopes: static, dynamic (absolute and relative), lengths of static lines, lengths of dynamic lines
content	*Faces*	2	*number of frontal faces, relative size of the biggest face*
	Skin	2	number of skin pixels, relative amount of skin with respect to the size of faces

CHAPTER 5

Data sets

It is not easy to get representative test data sets for the affective classification field, as there are few images labeled in terms of the emotions they generate. We chose three different data sets for this work. The first is a general photography set which is available for all researchers, the second is set of art photography which includes images made for the purpose of expressing emotion, and the third is an abstract paintings set which was chosen to investigate the measure of influence of the context versus context-free colors and textures. They are all described in detail in this section.

5.1 IAPS

The International Affective Picture System (IAPS) [43] is a common stimulus set widely used in emotion research. It consists of 716 natural color pictures taken by professional photographers. They depict complex scenes containing portraits, puppies, babies, animals, landscapes, scenes of poverty, pollution, mutilation, illness, accidents, insects, snakes, attack scenes and others. A selection of 396 of these pictures was categorized into discreet emotional categories in a study by Mikels et al. [49]. The categorization was done in two steps: A pilot study (20 subjects) with an open answering format has revealed eight frequently named types of emotions. In the main study the 60 participants had to label each picture in terms of these eight categories on a seven-point scale. Using this method 396 pictures were labeled either as one specific emotion or as a mixture of several emotions [49]. The resulting eight emotional categories were *Amusement, Awe, Contentment* and *Excitement* representing emotions that are *positive* on the Valence scale, and *Anger, Disgust, Fear* and *Sad* as *negative* emotions. Some examples are shown in Figure 5.1.

The images in this set are mostly documentary-style photographs that were not made as works of art or with the goal to express a feeling, but purely to capture and document a scene as we see it in real life, as realistically as possible. Colors and textures that appear on the photograph are not by the choice of an artist, but random,

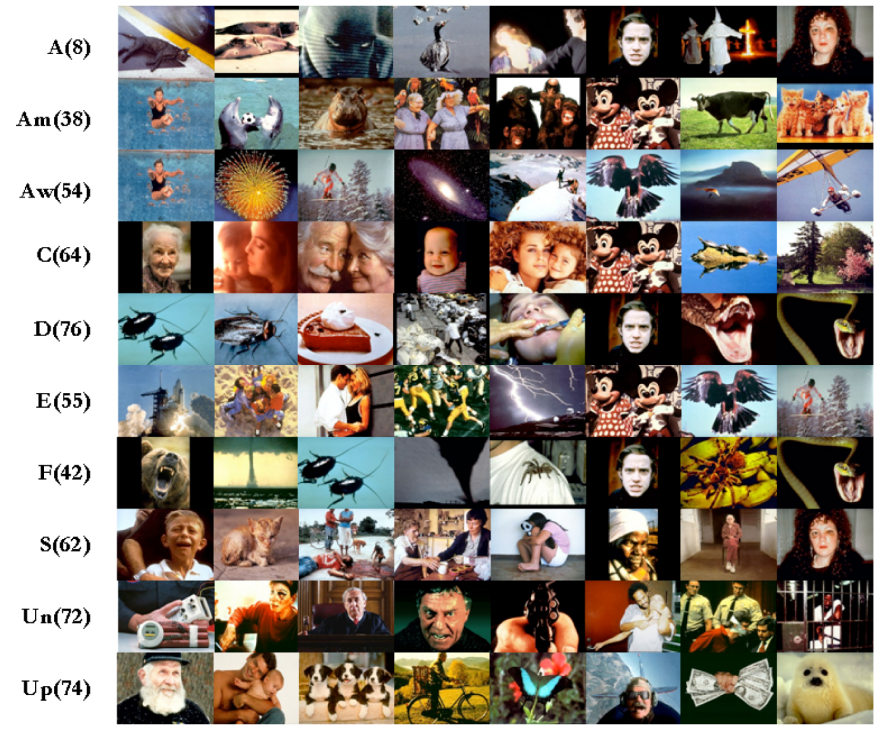

Figure 5.1: IAPS. Examples of the categories: A - anger, Am - amusement, Aw - awe, C - contentment, D - disgust, E - excitement, F - fear, S - sadness, Un - undifferentiated negative, Up - undifferentiated positive. Image from [78].

based on "what was there". The emotions that these pictures induce are based solely on the semantic content of the depicted scene and how the person who is viewing the image relates to it.
We chose this data set because it is available to anyone who does research in this field along with the same "emotional labeling" that has been done in a proper psychological experiment. Additionally, this set of pictures has already been used by other scientists in this field of work [78]. Using this set makes this work and its results directly comparable to the work of others.

The distribution of the classes in this set is shown in Figure 5.2.

5.2 Art photography

Since the IAPS data set contains mostly documentary-style photographs, but we wanted to test our features for "artistic" images, we created another data set. For this data set we looked for photographs where the intention of the artist is to express an emotion or induce a feeling in the observer of the image, and where the artist also has the opportunity to influence the appearance of the work of art as a whole and each

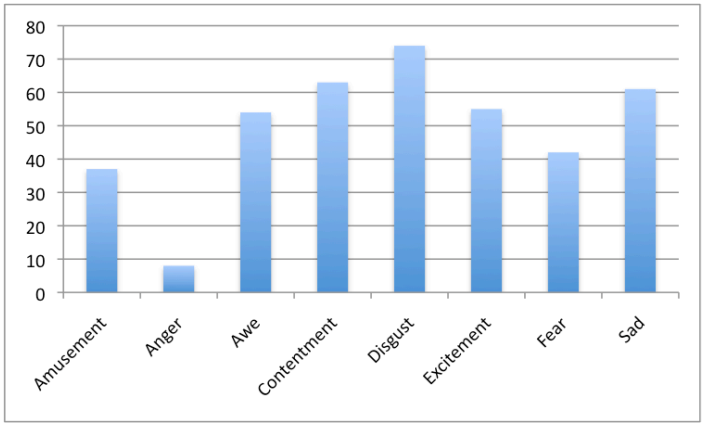

Figure 5.2: IAPS - the distribution of pictures per emotional class.

object in it. With this set we want to investigate whether the conscious use of colors and textures displayed by the artists improves the automatic affective image classification. To find such pictures, we browsed several photo-sharing web sites and chose the deviantArt.com [1] platform, since this platform concentrated on sharing selected art images instead of making public whole image collections and holiday snapshots (as e.g. on Flickr [2]), and had a relatively good tagging and search engine (as opposed to Photo.net [3] or others). To select the images for our data set, we simply entered the relevant affective words (i.e. the categories from the previous set, plus some synonyms of those words, e.g. anger, angry, sad, sadness, fear, content, peaceful, calm, excitement,...) into the search engine of the site, and downloaded the relevant pictures (omitting pictures that used the search word in a negated form, or pictures which simply displayed the word itself), from the first few pages of results. This resulted in a set of 807 images, which where sorted into categories by the intent of the artist expressed in the title, caption or labels that he or she attached to the image when uploading it to the website. Many of the illustrations in this document are examples taken from this "art images" set. The distribution of the classes in this set is shown in Figure 5.3.

5.3 Abstract paintings

To investigate the influence of certain image attributes (like colors and textures) on pictures without contextual content, we created another set of images. For this set we collected random abstract paintings from the popular art sharing platform deviantArt.com [1]. We were interested in how people would perceive and label such images without a clear content or scene. Our assumption is that in such cases the user would have to decide solely on the combinations of color and textures and their own personal taste and associations, i.e. in the first two cases the decision process might be more similar to that of a computer with an affective image classification software. Therefore we set up a website, where we asked users to label the images with the

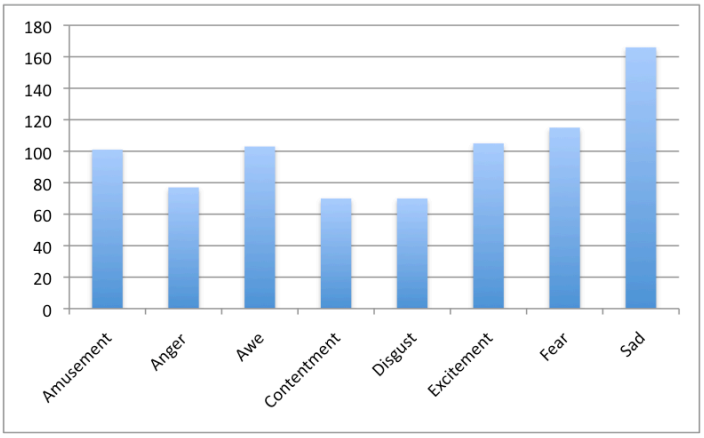

Figure 5.3: Art photography - the distribution of pictures per emotional class.

given emotional labels. We requested them to answer questions to measure the Valence and Arousal (VA), as well as to choose one of the words used by Mikels et al. [49] to annotate the IAPS set. For the VA analysis we used the Self-Assessment Mannequin (SAM), that was used in [43]. A screenshot of the labeling form is shown in Figure 5.4. We collected a set of 280 images. To avoid fatigue each user rated only 20

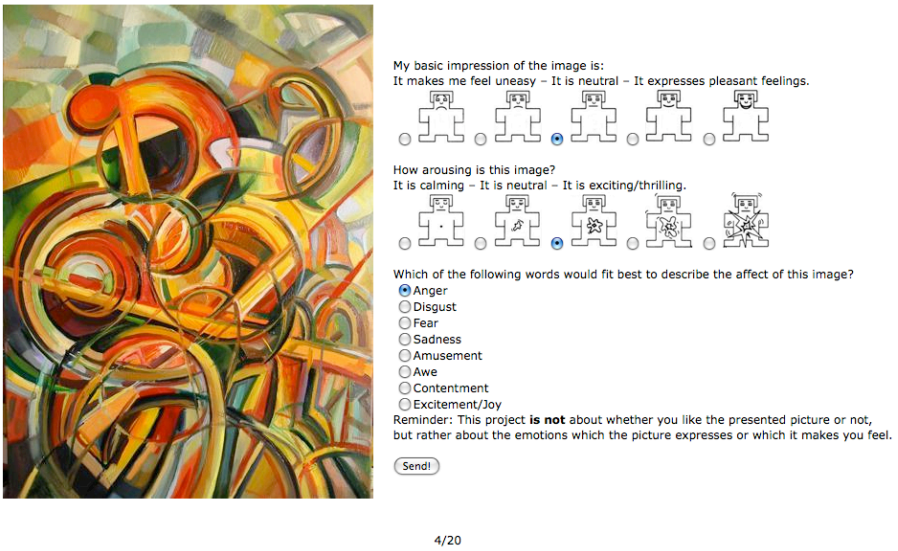

Figure 5.4: Screenshot of the image rating form.

randomly drawn images in a row. Each user could repeat the experiment any time, as often as he or she wanted (or until all images were rated). 232 users took part in the

Male	**Female**	unknown
122	101	9

Table 5.1: Number of male/female users.

Age	**Number of users**
< 20	24
20 - 25	99
25 - 30	52
30 - 40	29
> 40	19
unknown	9

Table 5.2: The age of users.

labeling, and they labeled the images approximately 4000 times, i.e. each of the 280 images was rated about 14 times.

Tables 5.1 and 5.2 provide some statistics about the users who took the challenge to label the abstract images. The gender statistics can be seen in Table 5.1. 44 of these received an art education, the field of work of the rest of the users was not art related. The youngest user was 9 years old, the oldest was 64. Table 5.2 presents the age distribution of the users. The recruited users were some students and staff of the Vienna University of Technology, their families and friends and the families, friends and colleagues of some of those families. Most of the participants were from Central Europe, but a number of participants where from other parts of the world.

The labeling results in the form of distribution of the classes in this set is shown in Figure 5.5. When processing the resulting labels, we observed that with these images, people did not very often come to a common rating of the image. Most images were labeled with several different emotions. In many cases, these were milder differences in labels where at least the Valence dimension was the same, the users chose different words and Arousal level, e.g. the image was mostly labeled as expressing “Amusement” and “Contentment”. However, there were also images where the majority of the users voted e.g. either for “Contentment” or “Disgust”, i.e. two very different feelings. There where also cases where so many different labels were used, that no clear “winner” could be found. Due to all these “dual” or unclear cases we had to reduce the set to 228 images, leaving out the cases where it could not be decided which label had most votes. However, it is clear that even those images left in the data set should receive more votes to clear up small uncertainties on how the image is perceived commonly. Clearly, when the context is missing, even humans have difficulty deciding on the emotional impact of an image. Moreover, it seems to be difficult and rare to induce strong emotions, such as anger just by using colors and textures. In our data set, only three images had a majority voting on “Anger”, making the machine learning task for this category practically impossible.

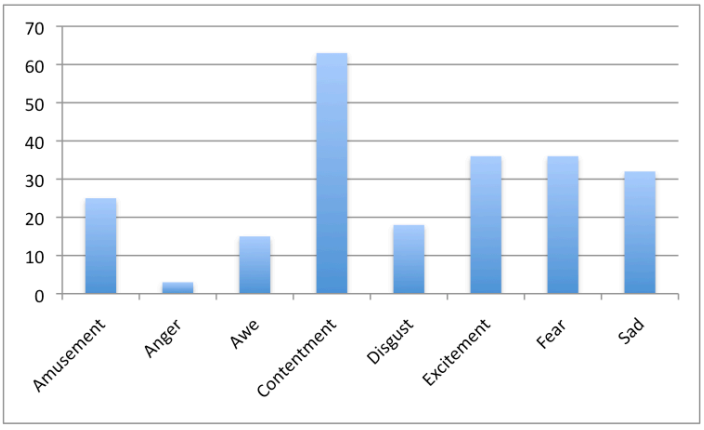

Figure 5.5: Abstract paintings - the distribution of pictures per emotional class.

CHAPTER 6

Evaluation and Results

For evaluation, we conducted several experiments to measure the performance of our features and compare the results with that of Yanulevskaya et al [78] and also to Wang Wei-ning et al [75].

In [78] the IAPS picture set was used for evaluation, as well as the same emotional categories as in this work, so the results are directly comparable. Similar to the work in [78], we will perform some feature extraction and compare their results for their feature subsets to ours. Following Yanulevskaya et al. we will then select our best feature subsets for each category and compare their performance to the best results in [78].

Wang Wei-ning [75], however, used an unknown data set as well as different categories. However, when we look at his *three-dimensional emotional factor space*, we can see that the dimensions can be roughly associated with the three dimensions of the emotion space discussed in Chapter 3. Namely, some of the impression words in factor F1 ((exhilarated - depressive, warm - cool, happy - sad, light - heavy, hard - soft, brilliant - gloomy, lively - tedious) are evocative of Valence (pleasantness - uneasiness), the factors F2 (magnificent - modest, vibrant - desolate, showy - elegant) and F3 (clear -fuzzy, fanciful - realistic) seem to be a mixture of Arousal (calming - thrilling) and Dominance (subdued - masterful). The categories used by Wang are mostly on the *expressive level* (as used by [15], mentioned in Chapter 2). We want to see how their features perform on our categories from the *emotional level*.

6.1 Experiments

Since we have 8 categories, we conducted several different experimental set-ups to analyze the performance of our framework.

The classification in all of the experimental set-ups is done by a *Naive Bayes* classifier [38], a classic Machine Learning algorithm implemented by the Weka Data Mining Software [77], in combination with K-fold Cross Validation. We used K=5 for

cross validation, because there are only 8 images in the category *Anger* of the IAPS data set. We used a *stratified* version of K-fold Cross Validation, i.e. ensuring that the distribution of the data in the test and training set was approximately the same as in the original data set (as opposed to a *non-stratified* version, where the division into training and test set is simply random without consideration of class distributions) [77].

All emotions

First we evaluated the obvious case where the whole data set is used as it is, with all 8 emotional categories, distributed as they are in the retrieved set. The baseline for this evaluation is 12.5% for a random classification.

One - versus - All

Here we use the standard "one-versus-all" paradigm to transform the classification problem into a binary classification by choosing one category and compare it to all others. Hence we get a very unbalanced set, with one huge category and a small one. To ensure a fair evaluation and avoid having to subsample the set, we use the evaluation measure described in Section 6.3. The baseline for this experiment is therefore 50%.

Each - versus - Each

To explore the separability between the classes, we create 28 data sets, each containing two of the 8 classes, hereby classifying each class against each. Again, the baseline for each set is 50%.

Dimensions

Finally, we also test the classification performance along two of the dimensions of the VAD-emotional space, the Valence axis (positive versus negative emotions) and the Arousal axis (calming versus exciting/thrilling), although the separation in the second case is not so clear.

6.2 Training and Test sets

We tested our features on all three data sets presented in Chapter 5, plus we combined all the datasets to form one "general" image set, containing *documentaristic*, realistic straight-out-of-the-camera photographs, as well as *artistic* expressive photographs and *abstract* art paintings. As already mentioned above, we separate the data into a training and test set using 5-fold Cross Validation (i.e. separating the data into 80% training and 20% for testing, repeating it 5 times each time taking a different portion of the data for testing and training).

6.3 Evaluation measure

Since we do not have a balanced data set, but the probabilities for the categories should theoretically be the same for all, we optimize for the average *error per class*, or rather average *true positive rate per class*, instead of the correct rate over all samples. This procedure is independent of the number of positive and negative samples. Hence, we do not have to subsample the classes. We will also call this measure an *unbiased correct rate*.

$$\textit{unbiased correct rate} = \frac{\sum_i (\text{percentage of true positives for category } i)}{i} \tag{6.1}$$

6.4 Feature Selection

Additionally some feature selection is performed for each experiment. The first employs a Subset extraction using using the *ClassifierSubsetEval* algorithm in Weka [77] with *GreedyStepwise* search method (further, we will call the subsets extracted with this method *Subset 1*), and the second uses wrapper-based subset evaluation [41] with a genetic search algorithm [24] (implemented in Weka [77] as *WrapperSubsetEval* and *GeneticSearch*, we will refer to it as *Wrapper subset*). For comparison, we also performed Principal Component Analysis (PCA) on the original datasets to reduce the features. In the cases of the eight one-versus-all experiments, we also performed "manual" feature selection, by performing the classifications of the data sets with only one feature at a time, for each feature and then selecting only those features for each data set which resulted in a classification which had the *unbiased correct rate* value higher than 0.51 (with only the one selected feature). We will refer to these subsets as *manual* subsets in the Results section. For complete lists of features in each subset, see the tables in the Appendix.

6.5 Results - IAPS data set

All emotions

First we evaluated the obvious case where the whole data set is used as it is, with all 8 emotional categories, distributed as they are in the data set. The baseline for this evaluation is 12.5% for a random classification. As can be seen in Figure 6.1, Subset 1, the feature subset selected by the *ClassifierSubsetEval* algorithm with *GreedyStepwise* search yields the best performance and is 10% better than a random classifier. The selected features were in this case only *Rule of Thirds - Brightness of Center*, *Number of frontal faces*, *Tamura - direction*, *Texture -Hue- correlation*, *Texture -Saturation- entropy*, *Area-warm*, and *Emotion Color equation- std Pleasure*. The classification performance values in the diagrams are calculated as described above in section 6.3. As can be seen in the confusion matrix in Table 6.1, all classes are most often classified as *Disgust*. This is due to the fact that *Disgust* is with its 74 samples the most frequent class of the data set. *Anger* samples never get classified correctly. However, it should be noted, that there are only 8 pictures in the category *Anger*, which makes

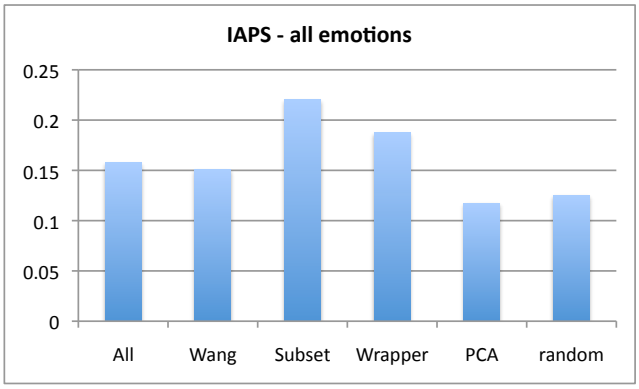

Figure 6.1: Classification performance for the IAPS with all 8 categories.

machine learning very difficult. Furthermore due to the double labeling (several images received two class labels) the data set is not completely separable, which further dampens the performance.

↓ classified as →	Am	An	Aw	Co	Di	Ex	Fe	Sa	sum
Amusement (Am)	7	0	6	7	10	0	1	6	37
Anger (An)	0	0	1	1	1	1	2	2	8
Awe (Aw)	1	0	22	3	16	5	4	3	54
Contentment (Co)	7	0	16	8	21	3	1	7	63
Disgust (Di)	1	0	7	2	54	1	1	8	74
Excitement (Ex)	3	0	24	3	17	2	1	5	55
Fear (Fe)	2	0	6	2	20	0	1	11	42
Sad (Sa)	2	0	8	5	23	6	4	13	61

Table 6.1: IAPS - confusion matrix.

One - versus - All

The experiments in this section were done by taking one category and trying to distinguish it from all others. This was done for each of the 8 emotional categories, leading to a two-class classification for each. In Figure 6.2 the classification performance for each feature subset is shown. For comparison, we also added the performances of the Weibull, Gabor and combined Weibull + Gabor features computed by Yanulevskaya et al [78] (the detailed performances for these sets are from a presentation kindly sent to us by Yanulevskaya).

Interestingly we notice that the manually selected features for the category *Amusement*, namely the *number of faces* and the *size of the biggest face*, drastically improve the performance on this class. It seems that in this data set, the presence of a big face points to an amusing image. We assume / hypothesize that this is because images with

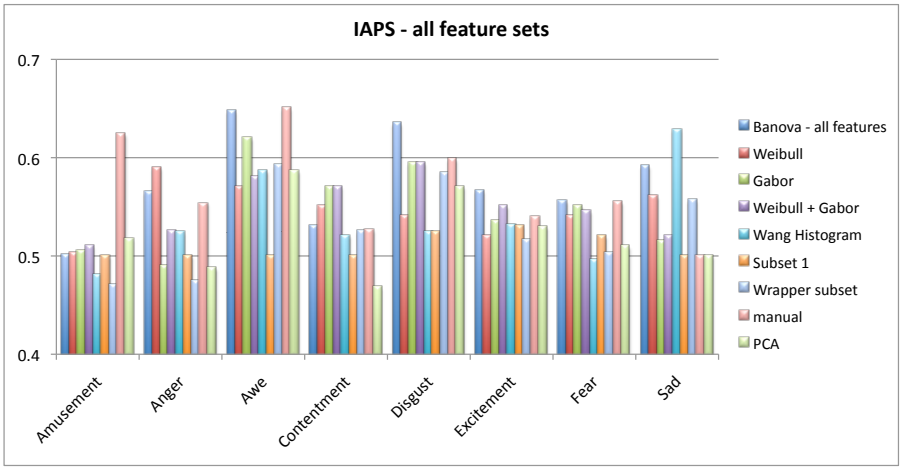

Figure 6.2: Classification performance for each feature set of the one-vs-all experiments. Weibull, Gabor and combined Weibull + Gabor set are from Yanulevskaya [78].

big faces are mostly portraits and most commonly portraits show a somehow idealized view on the person leading to a positively perceived image.

For category *Awe*, taking all features leads to the same results as using only the *manual* subset consisting of *Itten - area of cold colors, Itten - contrast of hue, blue, Low depth of Field of the Saturation channel, Rule of Thirds - mean Hue of center*, and *Wang Histogram - factor 1*. *Sad* images were best detected by the histogram developed by Wang et al [75]. Surprisingly, the automated feature selection methods didn't offer any benefit in form of classification performance boosting for this data set.

In Figure 6.3 we take the best performing features for each class and compare them to the best performing features from [78] as well as to the performance of the features suggested in [75]. As can be deduced from the diagram, our feature sets outperform both the results by Yanulevskaya and the pure Wang histogram features on this data set for 5 of 8 categories (Figure 6.2) or, if we take the Wang histogram as our selection for the Sad category, our results outperform Yanulevskaya's numbers in 6 of 8 categories (Figure 6.3).

Each - versus - Each

To explore the separability between the classes, we create 28 data sets, each containing two of the 8 classes, hereby classifying each class against each. Again, the baseline for each set is 50%. For the classification results in Table 6.2 all features were used, although the results of the per-feature classification suggest that selecting adequate subsets for each of the category-pairs could boost the performance of certain classifications. Table 6.2 shows what we already expected: classes with "similar" emotions, like *Fear vs. Anger*, *Sad vs. Anger* or *Excitement vs. Awe* don't separate well, whereas

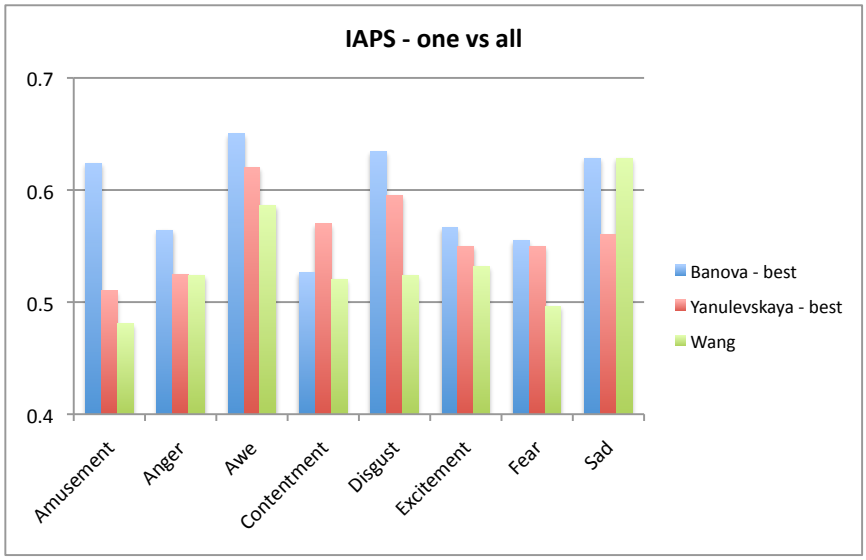

Figure 6.3: Classification performance taking our best features for each category compared against best features from Yanulevskaya [78] and the feature set as described in [75].

very "distinct" emotions, like *Sad vs. Awe* or *Disgust vs. Awe* separate better.

	Am	An	Aw	Co	Di	Ex	Fe	Sa
Amusement (Am)								
Anger (An)	0.56							
Awe (Aw)	0.70	0.57						
Contentment (Co)	0.42	0.63	0.63					
Disgust (Di)	0.58	0.63	0.69	0.60				
Excitement (Ex)	0.54	0.57	0.46	0.53	0.68			
Fear (Fe)	0.54	0.47	0.62	0.60	0.49	0.67		
Sad (Sa)	0.56	0.47	0.70	0.54	0.62	0.58	0.59	

Table 6.2: Classification performance of each-versus-each experiments from the IAPS.

Dimensions

Finally, we also test the classification performance along two of the dimensions of the VAD-emotional space, the *Valence* axis (*pleasant* versus *unpleasant* emotions) and the *Arousal* axis (*calming* versus *exciting/thrilling*). The diagram in Figure 6.4 depicts the classification performance on the *Valence* axis, separating *positive* from *negative* emotions. The best result is achieved with all features, measuring 73% average of

true positive rates per class. The measurements for the classification of *Arousal* were, however, not better than random.

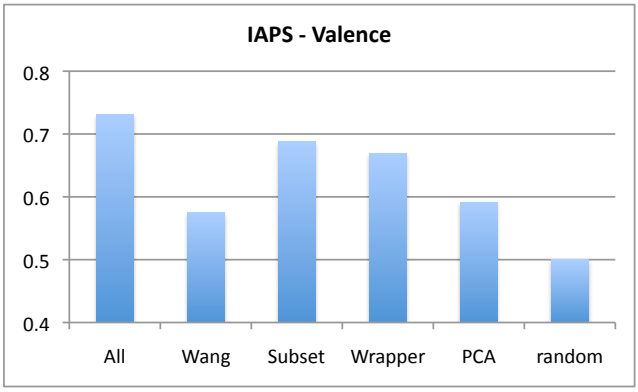

Figure 6.4: Classification performance for the IAPS along the *Valence* axis, separating *pleasant* images from *unpleasant*.

6.6 Results - art data set

All emotions

For the *art* image set we get generally the best results for the eight class classification when compared to the other data sets. As can be seen in Figure 6.5, all classifications are better than random and the best classification performance, 28,4%, is achieved with the Wrapper subset, the feature subset selected by the *WrapperSubsetEval* algorithm with *GeneticSearch*. This is more than 15% better than random.

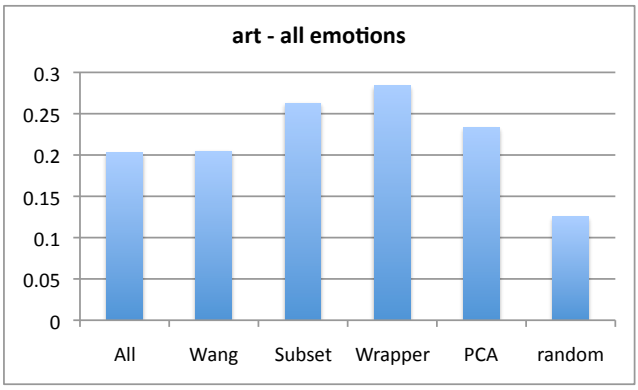

Figure 6.5: Classification performance for the *art* image set with all 8 categories.

As can be seen in the confusion matrix in Table 6.3, all classes are most often falsely classified as *Disgust*, followed by *Fear* and *Amusement*. However, *positive* classes are mostly classified as *Amusement* or *Content*, and *negative* classes are preferably classified as *Disgust* of *Fear*, which makes sense, since these classes are often confused even by humans as they emotionally "overlap".

↓ classified as →	Am	An	Aw	Co	Di	Ex	Fe	Sa	sum
Amusement (Am)	41	1	7	12	19	8	8	5	101
Anger (An)	6	7	4	10	24	1	19	6	77
Awe (Aw)	25	2	4	24	16	4	16	12	103
Contentment (Co)	7	0	4	22	12	4	6	15	70
Disgust (Di)	13	6	1	4	19	3	20	4	70
Excitement (Ex)	31	0	12	9	25	15	7	6	105
Fear (Fe)	13	3	3	11	22	0	47	16	115
Sad (Sa)	19	1	12	20	28	1	42	43	166

Table 6.3: *art* image set - confusion matrix.

One - versus - All

As with the IAPS data set , the experiments in this section were done by taking one category and trying to distinguish it from all others. Again we show the classification performances for all feature subsets created by feature selection in Figure 6.6. Since this data was not used by any other scientific paper, we can't compare our results on this data set.

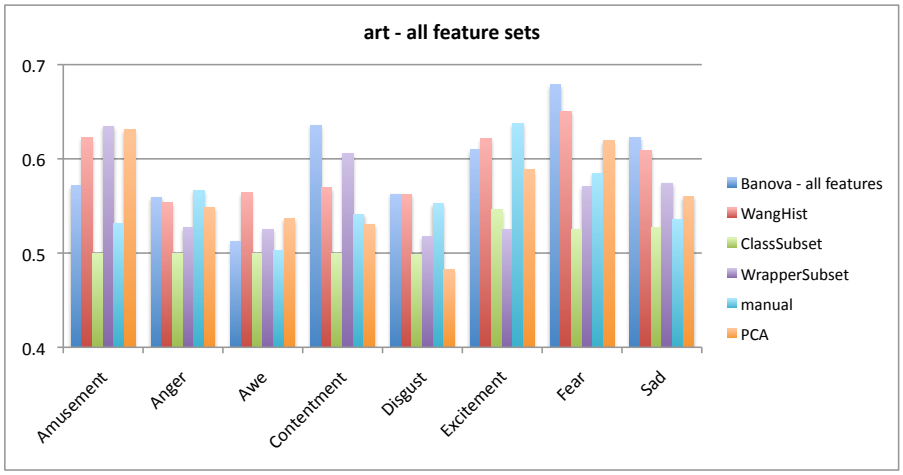

Figure 6.6: Classification performance for each feature set of the one-vs-all experiments.

The best performance on the *Amusement* class is achieved by adding *Level of Detail*, *Low Depth of Field*, *Line slopes*, *Itten contrasts*, *Emotion color equations* and *Saturation and Hue statistics* to the *Wang histogram* features. *Anger* is best expressed by *Wavelet textures*, *Texture entropy, contrast, mean range and mean standard deviation on the Hue channel* and *Itten Saturation contrast*. *Excitement* classifies best by measuring various Saturation statistics, texture of the Saturation channel, *Itten Saturation and Hue contrast*, *Itten - Area of cold colors*, brightness contrast, the *Emotion color equations of Arousal* and the amount of *pink* and *orange*. All others are best classified using either all features or the *Wang histogram* feature subset.

In Figure 6.7 we show the performance for our best performing features for each class.

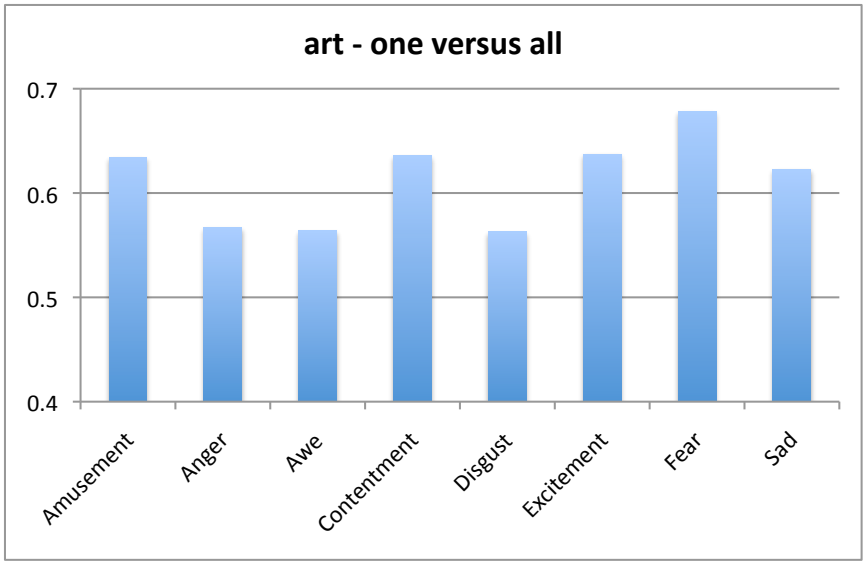

Figure 6.7: Classification performance for the *art* image set taking our best features for each category.

Each - versus - Each

When classifying each class against each, we get the classification results presented in Table 6.4. All features were used, although the results of the per-feature classification suggest that selecting adequate subsets for each of the category-pairs could boost the performance of certain classifications. As Table 6.4 shows the best performance in distinguishing between the two selected classes is achieved between *Contentment vs. Anger*, *Disgust vs. Contentment*, *Sad vs. Excitement*, and *Fear vs. Contentment*. The Naive Bayes classifier had the biggest difficulty separating *Disgust vs. Anger*, *Excitement vs. Amusement* and *Excitement vs. Awe*.

	Am	An	Aw	Co	Di	Ex	Fe	Sa
Amusement (Am)								
Anger (An)	0.66							
Awe (Aw)	0.55	0.55						
Contentment (Co)	0.58	0.71	0.63					
Disgust (Di)	0.62	0.47	0.57	0.71				
Excitement (Ex)	0.49	0.67	0.51	0.64	0.62			
Fear (Fe)	0.69	0.59	0.63	0.70	0.57	0.73		
Sad (Sa)	0.61	0.60	0.62	0.65	0.63	0.70	0.62	

Table 6.4: Classification performance of each-versus-each experiments from the *art* image set.

Dimensions

The classification performance along two of the dimensions of the VAD-emotional space, the *Valence* axis and the *Arousal* axis were similar to the IAPS data set. The classification of Valence performed slightly worse than on IAPS, but Arousal was slightly better, having values between 52-56% for the various subsets. The performance for classification of Valence can be seen in Figure 6.8.

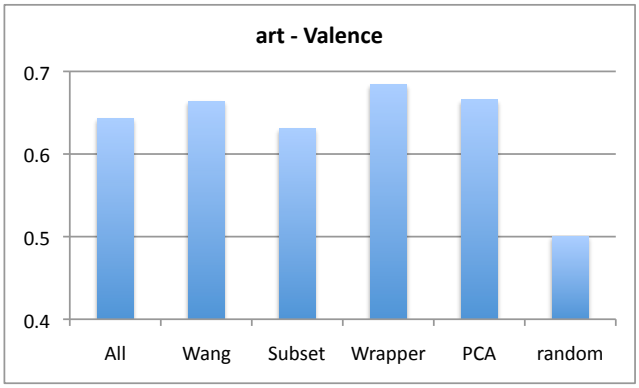

Figure 6.8: Classification performance for the *art* image set along the *Valence* axis, separating *pleasant* images from *unpleasant*.

6.7 Results - *ABSTRACT* data set

All emotions

The classification performance for the *ABSTRACT* image set with all emotions is barely better than random. Particularly the *Wang histogram* feature set gives a bad classification statistic, which is surprising since it was developed for art paintings [75] and one would expect it to work best for this set of images. Table 6.5 shows that

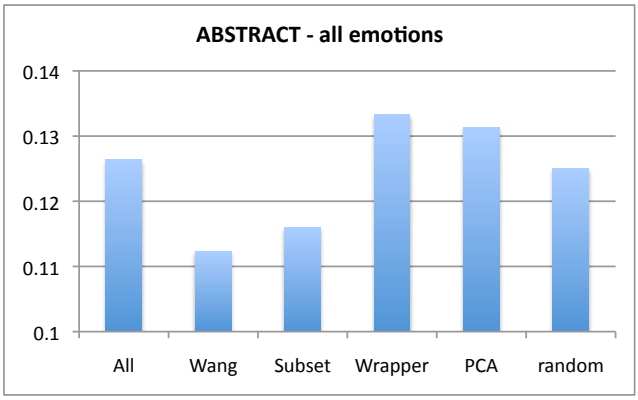

Figure 6.9: Classification performance for the *ABSTRACT* image set with all 8 categories.

most samples are falsely classified as *Excitement* or *Awe*, with only one resp. seven images classified as *Anger* or *Disgust*, most of them falsely. However, we must note that the category Anger has only 3 samples and it is therefore practically impossible for a machine learning algorithm to learn to classify them correctly. As was already noted before, not even humans agree on the labeling of such images, so the generally bad results for the automatic classification are not unexpected.

↓ classified as →	Am	An	Aw	Co	Di	Ex	Fe	Sa	sum
Amusement (Am)	1	1	4	6	0	7	4	2	25
Anger (An)	1	0	0	1	0	0	0	1	3
Awe (Aw)	1	0	5	4	1	1	1	2	15
Contentment (Co)	4	0	12	15	1	19	5	7	63
Disgust (Di)	1	0	5	5	2	3	0	2	18
Excitement (Ex)	0	0	7	11	1	9	2	6	36
Fear (Fe)	2	0	6	4	2	8	8	6	36
Sad (Sa)	1	0	6	6	2	7	4	6	32

Table 6.5: *ABSTRACT* image set - confusion matrix.

One - versus - All

The statistics for distinguishing one category from all others are shown for all feature subsets in Figure 6.10. Again, the best results for *Amusement* are for the feature subset selected by *manual* feature selection, which consists only of texture measurements (entropy, contrast, mean and standard deviation) of the Hue channel and the area of light regions. Clearly the best features for *Excitement* are the average and median of the Saturation channel, *Low Depth of Field Indicator* of the Hue channel, the amount of orange and purple, texture measurements (homogeneity, mean, standard deviation)

of the Saturation channel, and the area of dark regions. For the first time, features created by the Principal Components Analysis lead to the best separation of the *Fear* and *Sad* categories.

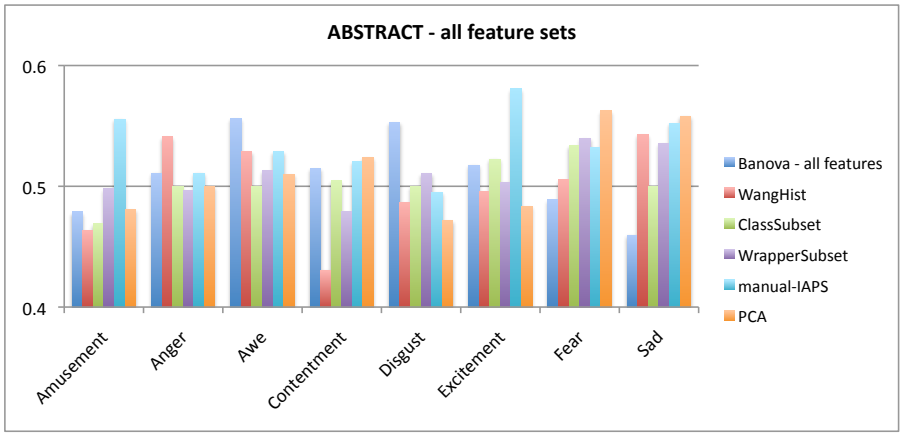

Figure 6.10: Classification performance for each feature set of the one-vs-all experiments.

In Figure 6.11 we summarize the performance for our best performing features for each class.

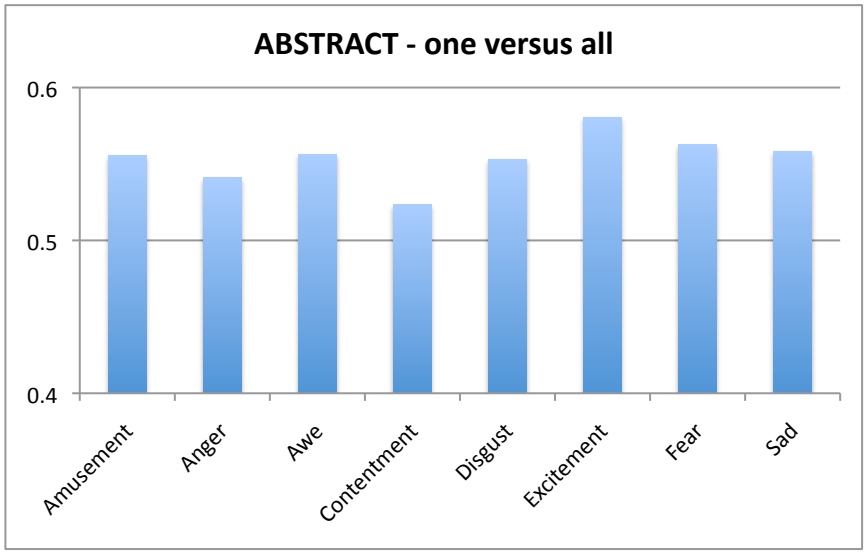

Figure 6.11: Classification performance of the *ABSTRACT* image set taking our best features for each category.

Each - versus - Each

When classifying each class against each, we get the classification results presented in Table 6.6. All features were used, although the results of the per-feature classification suggest that selecting adequate subsets for each of the category-pairs could boost the performance of certain classifications. In Table 6.2 it can be seen that only the pairs *Fear vs Awe*, *Disgust vs. Sad* and *Disgust vs. Awe* have a success rate above 60%.

	Am	An	Aw	Co	Di	Ex	Fe	Sa
Amusement (Am)								
Anger (An)	0.50							
Awe (Aw)	0.35	0.50						
Contentment (Co)	0.44	0.50	0.53					
Disgust (Di)	0.59	0.45	0.62	0.55				
Excitement (Ex)	0.50	0.50	0.53	0.51	0.47			
Fear (Fe)	0.46	0.50	0.68	0.56	0.57	0.57		
Sad (Sa)	0.46	0.50	0.58	0.49	0.64	0.57	0.46	

Table 6.6: Classification performance of each-versus-each experiments from the *ABSTRACT* image set.

Dimensions

During classification performance along two of the dimensions of the VAD-emotional space, the *Valence* axis and the *Arousal* axis only the feature subset created by *ClassifierSubsetEval* algorithm with *GreedyStepwise* search method scored more than 60%. The subset in question was composed of the relative amount of green, pink and red in the picture, the length of the horizontal and vertical lines, the correlation of texture in the Hue and Saturation channel, and a part of the first and second factor of the Wang histogram [75]. More so, the *ABSTRACT* set is the only image set where the classification of *Arousal* reaches the 60% level. Again this is achieved by the Subset 1, this time made up of median Brightness, some Hue statistics, the *Rule of Thirds* measures, the amount of black, *Itten contrasts* of Saturation and Hue and *Tamura texture* direction as well as texture contrast of Brightness, and mean and standard deviation of Hue texture. The performance for classification of Valence can be seen in Figure 6.12.

6.8 Results - combined data set

All emotions

As can be seen in Figure 6.13 the best results for the eight class classification for the *combined* image set, 22%, was achieved by the subset chosen by the *WrapperSubsetEval* algorithm with *GeneticSearch*, but with only relatively small differences in performance between the various subsets. The subset in question comprises 40 values representing various texture measurements (*Wavelet*, *Tamura*, contrast, entropy, correlation, energy,...), all main *Itten contrasts*, the area of dark, middle gray and light

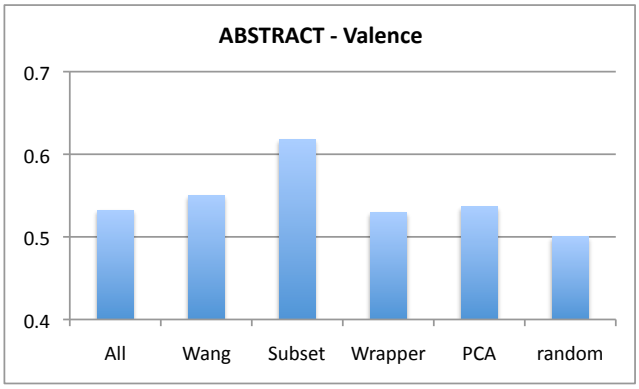

Figure 6.12: Classification performance for the *ABSTRACT* image set along the *Valence* axis, separating *pleasant* images from *unpleasant*.

regions, area of cold, median of Saturation, average Hue (vector), Brightness of center, the *Emotional color equation* for *Dominance*, *Level of Detail*, and the size and number of faces.

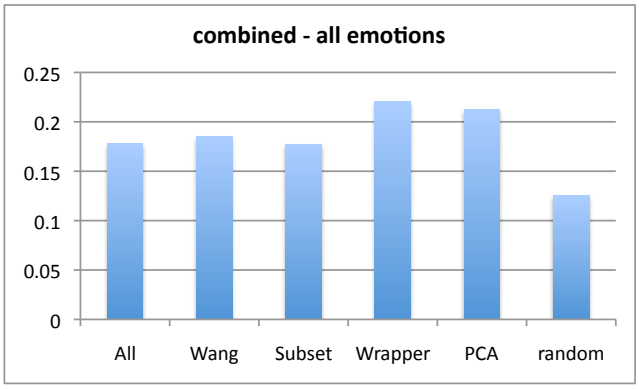

Figure 6.13: Classification performance for the *Combined* image set with all 8 categories.

From the confusion matrix in Table 6.7 we can observe, again, that the highest number of samples were falsely classified as *Disgust*, followed by *Awe*. In the combined data set we have no class that would clearly stand out as being classified correctly most often.

One - versus - All

For the combined image set, Figure 6.14 suggests, that the best features are either to take all features, the *Wang histogram* feature set or features created from our features

↓ classified as →	Am	An	Aw	Co	Di	Ex	Fe	Sa	sum
Amusement (Am)	47	9	21	30	30	11	1	14	163
Anger (An)	5	14	17	5	19	4	12	12	88
Awe (Aw)	18	18	55	22	22	10	9	18	172
Contentment (Co)	29	11	38	36	42	18	5	17	196
Disgust (Di)	19	24	14	24	48	7	17	9	162
Excitement (Ex)	29	5	47	27	42	25	9	12	196
Fear (Fe)	26	32	20	19	35	4	29	28	193
Sad (Sa)	21	24	41	22	51	9	34	57	259

Table 6.7: *Combined* image set - confusion matrix.

by transforming them with PCA, i.e. many features are needed to separate such a diverse image set as the combination of general photos, artistic photos and abstract art paintings.

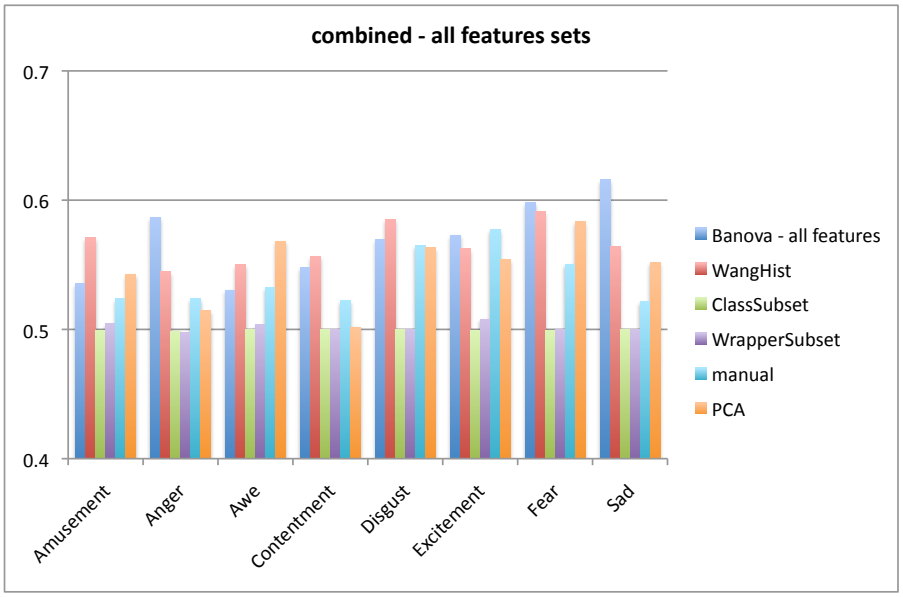

Figure 6.14: Classification performance for each feature set of the one-vs-all experiments.

In Figure 6.15 we summarize the performance when taking the best performing features for each category.

Each - versus - Each

When classifying each class against each, we get the classification results presented in Table 6.8. All features were used, although the results of the per-feature classification

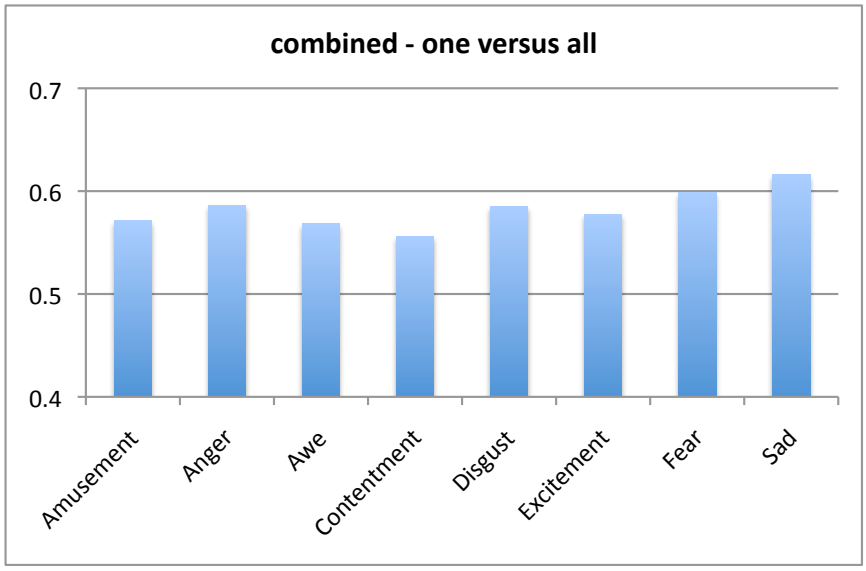

Figure 6.15: Classification performance of the combined image set taking our best features for each category.

suggest that selecting adequate subsets for each of the category-pairs could boost the performance of certain classifications. As Table 6.8 shows the best performance in distinguishing between the two selected classes is achieved between *Fear vs. Amusement*, *Fear vs. Excitement*, and *Excitement vs. Anger*. The Naive Bayes classifier had the biggest difficulty separating *Excitement vs. Amusement*.

	Am	An	Aw	Co	Di	Ex	Fe	Sa
Amusement (Am)								
Anger (An)	0.62							
Awe (Aw)	0.60	0.55						
Content (Co)	0.54	0.60	0.53					
Disgust (Di)	0.61	0.57	0.63	0.58				
Excitement (Ex)	0.52	0.64	0.54	0.53	0.63			
Fear (Fe)	0.65	0.56	0.61	0.61	0.55	0.65		
Sad (Sa)	0.58	0.53	0.60	0.56	0.55	0.62	0.59	

Table 6.8: Classification performance of each-versus-each experiments from the *Combined* image set.

Dimensions

In the classification of *Valence* there is no clear winner among the various subsets. All perform about 60-64%. *Arousal* had again discouraging values with the maximum

classification success at 53%. The performance for classification of Valence can be seen in Figure 6.16.

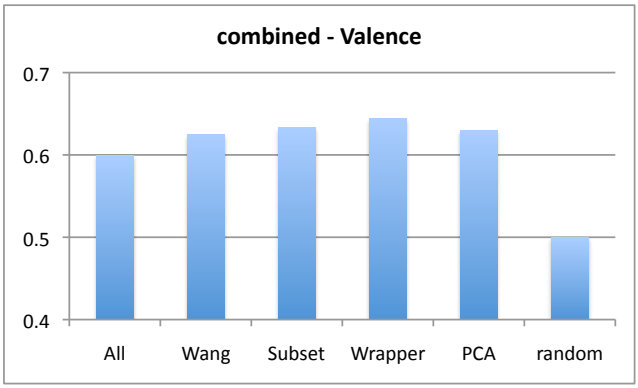

Figure 6.16: Classification performance for the *Combined* image set along the *Valence* axis, separating *pleasant* images from *unpleasant*.

Comparison

Figure 6.17 is a comparison between the data sets used in this work. The classification of the ABSTRACT data set has the worst performance.

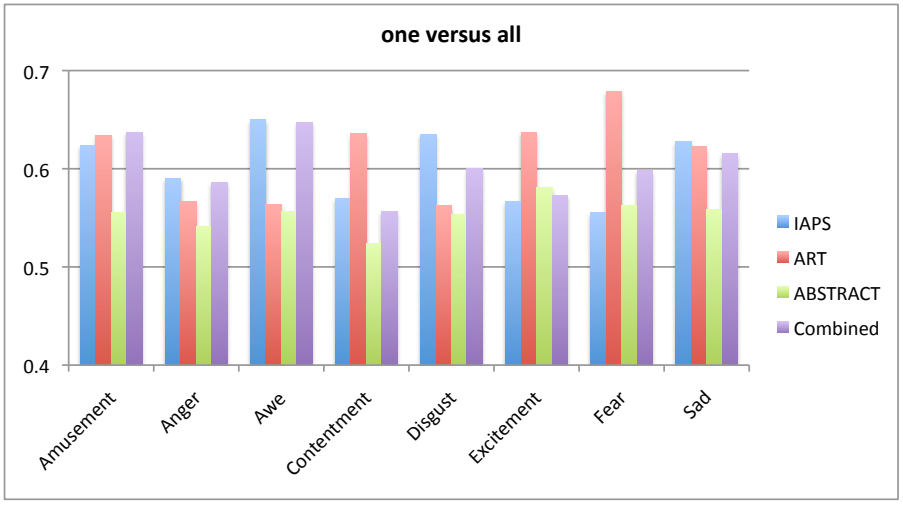

Figure 6.17: Comparison of the classification performance for all image sets used in this work. The results are from the best feature selections implemented during this work (i.e. previously referred to as "Banova-best").

CHAPTER 7

Conclusions and Future Work

As mentioned in the introduction, the field of affective content analysis of images is a new field and much more work has to be done to make this kind of system applicable in the real world. At this stage of research, given the character of the problem in question, we can consider classification performance that is better than random as a progress in research. In our work, we made that progress. Moreover, our system performed better on the same data set than the one in the publication by Yanulevskaya et al. [78]. This indicates a potential to accomplish even better results in this field.

However, much has to be done to achieve the goal of creating a system that could effectively aid humans in their creative tasks and improve image retrieval significantly. As already summarized in [74], the future research trends lie in individualizing emotion models, integrating affective content with text information on the web, integrating affect with image aesthetics or building general systems.

Individualizing emotion models means a deviation from the hypothesis of a common emotion and towards building an individual and subjective emotion model that is tailored to the specific user or user group, their personality and needs. The general idea is to let the user create his own emotion profile and train the system to his or her preferences and common emotions.

Integrating with text search on the web basically means adding linguistic information expressed in the image, i.e., some kind of labeling that would make the image searchable for text-based search mechanisms or integrating specific features and their possible interpretations directly into the search algorithms.

There is a strong relationship between image aesthetics and the emotional response to an image, indicating that combining the two should be of advantage for the research of emotional semantics. We already used many features that have also been used for aesthetic classification in [17] and several of them proved to be effective in this context.

Since most of the current systems are built for specialized practical purposes or in specific domains, such as textile, craft objects, patterns, art paintings, or landscapes,

the authors of [74] see the biggest challenge in creating a general purpose system. In our work, we tested our system on three different data sets, containing general photographs of various topics and scenes, artistic photography work and abstract paintings. The combined data set of all of these three subsets can be viewed as a general purpose image database. This means that we tested our framework for performance in the general purpose case and it performed better than random, which can be viewed as a good result in the context of the problem at hand, although it is far from being useful for real world applications.

However, all these emphases have a common factor. They first have to solve the key issues presented in the introduction (Chapter 1): the issue of feature extraction and classification algorithms, the issue of selecting the emotional categories and the collection of test data.

The collection and labeling of the data combined along with the selection of emotional categories should be best solved by psychologists, as they have the proper tools and knowledge on how to conduct experiments with humans as well as analyze and name their emotions. The IAPS [43] and the follow-up work on the IAPS by Mikels et al. [49] are the first available sets that were labeled during rigorous psychological experiments. The disadvantage is that this set is relatively small (only 396 images are labeled). For a well trained machine learning classifier a much higher number of instances is desirable. It would be useful if larger or more such image sets existed. Using such "standard" sets would also make the various publications in this topic comparable, as it has made our work comparable to that in [78]. Our experience with collecting and labeling the abstract paintings data set also showed that a large number of participants is needed to create a representative ground truth. For future work, we should expand the abstract data set to contain more images, but particularly more participants to label the images and create clearer statistics on the opinion of the majority or the "common user". But from the reactions of the participants, we saw that not even humans can decide what they feel when looking at an abstract image, and if they decide, the ratings of two distinct people can be very different, even two ratings from the same person on a different day vary. The next question poses itself: If not even people can decide how an image affects them, what kind of performance in this area should we expect from a computer?

More research in the direction of what makes an image look sad/happy/angry/... is needed. Many of the features have an implicit relationship to the affective response of the image. When using such "implicit" or "common" features, often more features are needed than when using specified features which directly reflect an attribute that influences affect, but in particular the results tend to be worse. In the topic of affective content analysis, this was specifically addressed in [75] and it also can be illustrated in our comparison with [78]. However, at this point there are still many theories about composing art and the psychology of art, which are yet unexploited, but could be of good use in this topic. Theories about the use and impact of shapes, lines and proportions by Arnheim [6] are among such potentially effective principles. In this work, we focused more on the influence of color and color features (although even this is not exploited as far as it could be), but the analysis of shape, symmetries,

proportions and "leading lines" appears to have a high potential. The combinations of those two, like e.g. analyzing neighboring regions' color contrasts, or identifying the color of the main focus of attention, even more so. Except for the abstract paintings category, the semantic content of the image helps much in deciding how humans feel about it. The occurrence of insects, e.g. was almost always accompanied by the label "Disgust". If we detected a smile on the image, it would most probably be viewed as a pleasing picture. In our work, we were surprised by how much simply the detection and measurement of the size of a frontal face in the image improved the classification results. Therefore we are positive that semantic content analysis could improve classification results.

Finally, the machine learning algorithm is of great importance. The choice of a particular algorithm, its training and fine-tuning are critical for the final result. Optimizations, like a good feature selection or weighting of the classifiers can improve the results dramatically, as e.g. Datta et al. showed in their work on image aesthetics [17] [18].

Apart from the image classification itself, the question of a proper user interface for such an image retrieval system is of importance. We suggest a browser-based approach as discussed at the end of Chapter 3, where the user could search images by specifying an emotional subspace and he/she could get clusters of thumbnails of images that map to that specific area, as well as their emotional labels. Developing a proper user interface could exploit the main advantages of the combined approach to emotion classification and enhance the usability of an application using affective image retrieval.

As the demand for effective content based image retrieval systems increases, affective image classification or similar emotional semantic image retrieval systems represent a promising and challenging research area with a prospect of various fields of application in the near future. This work represents a promising initial step towards a system capable of performing sufficiently well for practical use.

Acknowledgements

I would like to thank my supervisor Dr. Allan Hanbury for his patience and guidance and Julian Stöttinger for his help with many details and technical problems. Also I thank the rest of the colleagues at PRIP for kind support.

The images shown in this work are mostly from deviantArt.com and I acknowledge them as original artworks of their creators. If anyone would be interested in acquiring any of the images, please contact me. If they are to be modified, I can provide information and possibly contact to the artists who made them. As I understand it, it falls under "fair use" of the copyright statement to use them for scientific purposes, so if there is interest in obtaining the whole set I used for such purposes, I can make it available.

Special thanks go to my parents who made my whole studies, including this work, possible and whose kind support and love I can't appreciate enough.

References

[1] deviantart. www.deviantart.com.

[2] Flickr. www.flickr.com.

[3] photo.net. www.photo.net.

[4] Wordnet - a lexical database for the english language. http://wordnet.princeton.edu/.

[5] J. Angulo and J. Serra. Color segmentation by ordered mergings. *Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on*, 2:II–125–8 vol.3, Sept. 2003.

[6] Rudolf Arnheim. *Art and Visual Perception: A Psychology of the Creative Eye*. University of California Press, 2004.

[7] John Bates. Colorless green photographs. http://theonlinephotographer.typepad.com/the_online_photographer/2009/05/colorless-green-photographs.html.

[8] S. Beucher. The watershed transformation applied to image segmentation. In *Scanning Microscopy International*, pages 299–314, 1991.

[9] N. Bianchi, L. Berthouze, and T. Kato. Towards a comprehensive integration of subjective parameters in database browsing. pages 870–874, Aug 1998.

[10] Nadia Bianchi-Berthouze. K-dime: an affective image filtering system. *Multimedia, IEEE*, Volume 10(Issue 3):103 – 106, July-Sept. 2003.

[11] A. C. Bovik, M. Clark, and W. S. Geisler. Multichannel texture analysis using localized spatial filters. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 12(1):55–73, 1990.

[12] M. M. Bradley and P. J. Lang. International affective digitized sounds (IADS): Stimuli, instruction manual and affective ratings (Tech. Rep. No. B-2). Technical report, Gainesville, FL: The Center for Research in Psychophysiology, University of Florida, 1999.

[13] Sung-Bae Cho. Emotional image and musical information retrieval with interactive genetic algorithm. *Proceedings of the IEEE*, 92(4):702–711, Apr 2004.

[14] Martin Kampel Christian Liensberger, Julian Stöttinger. Color-based and context-aware skin detection for online video annotation. In *Proceedings of the IEEE 2009 International Workshop on Multimedia Signal Processing*, Oct. 2009.

[15] C. Colombo, A. Del Bimbo, and P. Pala. Semantics in visual information retrieval. *Multimedia, IEEE*, 6(3):38–53, Jul-Sep 1999.

[16] J. M. Corridoni, A. Del Bimbo, and P. Pala. Image retrieval by color semantics. *Multimedia Syst.*, 7(3):175–183, 1999.

[17] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Ze Wang. Studying aesthetics in photographic images using a computational approach. In *ECCV (3)*, pages 288–301, 2006.

[18] Ritendra Datta, Jia Li, and James Z. Wang. Learning the consensus on visual quality for next-generation image management. In *MULTIMEDIA '07: Proceedings of the 15th international conference on Multimedia*, pages 533–536, New York, NY, USA, 2007. ACM.

[19] Ingrid Daubechies. *Ten Lectures on Wavelets*. Regional Conference Series in Applied Mathematics. Soc for Industrial & Applied Math, December 1992.

[20] R. Dietz and A. Lang. Affective agents: Effects of agent affect on arousal, attention, liking and learning. 1999.

[21] Paul Ekman, Wallace V. Friesen, Maureen O'Sullivan, Anthony Chan, Irene Diacoyanni-Tarlatzis, Karl Heider, Rainer Krause, William Ayhan LeCompte, Tom Pitcairn, Pio E. Ricci-Bitti, Klaus Scherer, Masatoshi Tomita, and Athanase Tzavaras. Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of Personality and Social Psychology*, 53(Issue 4):712–717, Oct 1987.

[22] J. M. Geusebroek. Compact object descriptors from local colour invariant histograms. In *British Machine Vision Conference*, volume 3, pages 1029–1038, 2006.

[23] David E. Goldberg. *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1989.

[24] David E. Goldberg. *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley Professional, January 1989.

[25] Stuart Golodetz. Watersheds and waterfalls. http://accu.org/index.php/journals/1469.

[26] Rafael C. Gonzales and Richard E. Woods. *Digital Image Processing*. Number ISBN 0-201-18075-8 in S. 587ff. Prentice Hall, New Jersey, 2002.

[27] M. K. Greenwald, E. W. Cook, and P. J. Lang. Affective judgement and psychophysiological response: dimensional covariation in the evaluation of pictorial stimuli. *Journal of Psychophysiology*, 3:51–64, 1989.

[28] Allan Hanbury. A 3d-polar coordinate colour well adapted to image analysis. In *In Proceedings of the Scandinavian Conference on Image Analysis (SCIA)*, pages 804–811, 2002.

[29] Allan Hanbury. Constructing cylindrical coordinate colour spaces. *Pattern Recogn. Lett.*, 29(4):494–500, 2008.

[30] A. Hanjalic. Extracting moods from pictures and sounds: towards truly personalized TV. *Signal Processing Magazine, IEEE*, 23(2):90–100, 2006.

[31] Alan Hanjalic and Li-Qun Xu. Affective video content representation and modeling. *IEEE Transactions on Multimedia*, 7(1):143–154, 2005.

[32] R. M. Haralick, Dinstein, and K. Shanmugam. Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-3:610–621, November 1973.

[33] Robert Haralick and Linda Shapiro. *Computer and Robot Vision*. Addison-Wesley Longman Publishing Co., Inc., 1992.

[34] T. Hayashi and M. Hagiwara. Image query by impression words-the IQI system. *Consumer Electronics, IEEE Transactions on*, 44(2):347–352, May 1998.

[35] Ming-Kuei Hu. Visual pattern recognition by moment invariants. *Information Theory, IEEE Transactions on*, 8(2):179–187, 1962.

[36] Jing Huang, S.R. Kumar, M. Mitra, Wei-Jing Zhu, and R. Zabih. Image indexing using color correlograms. *Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on*, pages 762–768, Jun 1997.

[37] Johannes Itten. *The art of color : the subjective experience and objective rationale of color*. John Wiley, New York :, 1973.

[38] George H. John and Pat Langley. Estimating Continuous Distributions in Bayesian Classifiers. In *In Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 338–345. Morgan Kaufmann, 1995.

[39] Hang-Bong Kang. Affective content detection using HMMs. In *MULTIMEDIA '03: Proceedings of the eleventh ACM international conference on Multimedia*, pages 259–262, New York, NY, USA, 2003. ACM.

[40] Yan Ke, Xiaoou Tang, and Feng Jing. The design of high-level features for photo quality assessment. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1:419–426, June 2006.

[41] Ron Kohavi and George H. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97(1-2):273–324, 1997.

[42] L. Kovacs and T. Sziranyi. Focus area extraction by blind deconvolution for defining regions of interest. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(6):1080–1085, June 2007.

[43] P.J. Lang, M.M. Bradley, and B.N. Cuthbert. International affective picture system (IAPS): Affective ratings of pictures and instruction manual. Technical report, University of Florida, Gainesville, FL., 2008.

[44] Michael S. Lew, Nicu Sebe, Chabane Djeraba, and Ramesh Jain. Content-based multimedia information retrieval: State of the art and challenges. *ACM Trans. Multimedia Comput. Commun. Appl.*, 2(1):1–19, February 2006.

[45] Beatriz Marcotegui and Serge Beucher. Fast implementation of waterfall based on graphs. volume 30 of *Computational Imaging and Vision*, pages 177–186. Springer-Verlag, Dordrecht, 2005.

[46] Kanti V. Mardia and Peter E. Jupp. *Directional Statistics*. Wiley, 1972.

[47] Kresimir Matkovic, Denis Gracanin, Wolfgang Freiler, Jana Banova, and Helwig Hauser. Large image collections - comprehension and familiarization by interactive visual analysis. In *Smart Graphics*, pages 15–26, 2009.

[48] Fernand Meyer. Levelings, image simplification filters for segmentation. *J. Math. Imaging Vis.*, 20(1-2):59–72, 2004.

[49] Joseph A. Mikels, Barbara L. Fredrickson, Gregory R. Larkin, Casey M. Lindberg, Sam J. Maglio, and Patricia A. Reuter-Lorenz. Emotional category data on images from the international affective picture system. *Behavior Research Methods*, 37(4):626–630, November 2005.

[50] A. Ortony and T. J. Turner. What's basic about basic emotions? *Psychol Rev*, 97(3):315–331, July 1990.

[51] Andrew Ortony, Gerald L. Clore, and Allan Collins. *The Cognitive Structure of Emotions*. Cambridge University Press, July 1988.

[52] C. E. Osgood, G. Suci, and P. Tannenbaum. *The measurement of meaning*. University of Illinois Press, Urbana, IL, 1957.

[53] Li-Chen Ou, M. Ronnier Luo, Andrée Woodcock, and Angela Wright. A study of colour emotion and colour preference. Part I: Colour emotions for single colours. *Color Research and Application*, 29(Issue 3):232 – 240, June 2004.

[54] Eli Peli. Contrast in complex images. *Journal of the Optical Society of America*, 7(10):2032–2040, Oct. 1990.

[55] Rosalind W. Picard. *Affective Computing*. The MIT Press, Cambridge, September 1997.

[56] Chaonan Wang Qingfeng Wu, Changle Zhou. Content-based affective image classification and retrieval using support vector machines. *Affective Computing and Intelligent Interaction*, 3784:239–247, 2005.

[57] Z. Rasheed, Y. Sheikh, and M. Shah. On the use of computable features for film classification. *Circuits and Systems for Video Technology, IEEE Transactions on*, 15(1):52–64, Jan. 2005.

[58] Jia Li Ritendra Datta, Dhiraj Joshi and James Z. Wang. Acquine. http://acquine.alipr.com/.

[59] Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, V40(2):99–121, November 2000.

[60] Andrew Salway and Mike Graham. Extracting information about emotions in films. In *MULTIMEDIA '03: Proceedings of the eleventh ACM international conference on Multimedia*, pages 299–302, New York, NY, USA, 2003. ACM.

[61] Pierre Soille. *Morphological Image Analysis: Principles and Applications*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2003.

[62] Milan Sonka, Vaclav Hlavac, and Roger Boyle. *Image Processing: Analysis and Machine Vision*. Thomson-Engineering, September 1998.

[63] Charles Spearman. "general intelligence," objectively determined and measured. *American Journal of Psychology*, 15:201–293, 1904.

[64] Julian Stöttinger, Jana Banova, Thomas Pönitz, Allan Hanbury, and Nicu Sebe. Translating journalists' requirements into features for image search. *International Conference on Virtual Systems and Multimedia (VSMM)*, 2009.

[65] Hideyuki Tamura, Shunji Mori, and Takashi Yamawaki. Textural features corresponding to visual perception. *IEEE Transactions on Systems, Man and Cybernetics*, 8(Issue 6):460–473, June 1978.

[66] Inc. The MathWorks. Matlab, 1994-2009.

[67] P Valdez and A Mehrabian. Effects of color on emotions. *Journal of Experimental Psychology: General*, (123):394–409, 1994.

[68] J. van de Weijer, C. Schmid, and J. Verbeek. Learning color names from real-world images. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, June 2007.

[69] Jan C. van Gemert, Jan-Mark Geusebroek, Cor J. Veenman, Cees G. M. Snoek, and Arnold W. M. Smeulders. Robust scene categorization by learning image statistics in context. In *CVPRW '06: Proceedings of the 2006 Conference on Computer Vision and Pattern Recognition Workshop*, Washington, DC, USA, 2006. IEEE Computer Society.

[70] Paul Viola and Michael Jones. Robust real-time object detection. *International Journal of Computer Vision*, 2001.

[71] Paul Viola and Michael Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, 2004.

[72] Hee Lin Wang and Loong-Fah Cheong. Affective understanding in film. *Circuits and Systems for Video Technology, IEEE Transactions on*, 16(6):689–704, June 2006.

[73] Wei-Ning Wang and Ying-Lin Yu. Image emotional semantic query based on color semantic description. In *Machine Learning and Cybernetics, 2005. Proceedings of 2005 International Conference on*, volume 7, pages 4571–4576 Vol. 7, 2005.

[74] Weining Wang and Qianhua He. A survey on emotional semantic image retrieval. *15th IEEE International Conference on Image Processing*, pages 117–120, Oct. 2008.

[75] Jiang Sheng-ming Wang Wei-ning, Yu Ying-lin. Image retrieval by emotional semantics: A study of emotional space and feature extraction. *IEEE International Conference on Systems, Man and Cybernetics*, 4(Issue 8-11):3534 – 3539, Oct. 2006.

[76] Wang Wei-ning, Yu Ying-lin, and Zhang Jian-chao. Image emotional classification: static vs. dynamic. *Systems, Man and Cybernetics, 2004 IEEE International Conference on*, 7:6407–6411 vol.7, Oct. 2004.

[77] Ian H. Witten and Eibe Frank. *Data Mining: Practical machine learning tools and techniques, 2nd Edition*. Morgan Kaufmann, San Francisco, 2005.

[78] V. Yanulevskaya, J. C. van Gemert, K. Roth, A. K. Herbold, N. Sebe, and J. M. Geusebroek. Emotional valence categorization using holistic image features. In *IEEE International Conference on Image Processing*, 2008.

[79] Yun Zhai, Zeeshan Rasheed, and Mubarak Shah. A framework for semantic classification of scenes using finite state machines. In *International Conference on Image and Video Retrieval Machines*, pages 21–23, 2004.

Appendix

Feature set - manual selection

IAPS

Amusement	Number of frontal faces, Size of biggest face, Amount of skin - relative to face size	3
Anger	Low DOF Indicator - Hue	1
Awe	Rule of Thirds - Hue, Low DOF Indicator - Saturation, blue, Itten Colors - Hue Contrast, Itten Colors - Area cold, WangHist - f1	14
Content	yellow, Size of biggest face, Tamura - direction, Area-highsat	4
Disgust	Wavelet texture -Hue -level 1, orange, pink, red, Amount of skin - relative to face size, Texture -H- contrast, Texture -H- homogeneity, Texture -H- mean std, Texture -S- mean std, std Arousal	10
Excitement	Rule of Thirds - Hue, white, Amount of skin - relative to face size, Itten Colors - Area cold, WangHist - f1, Area-verylight	15
Fear	green, Itten Colors - Hue Contrast weighted, Texture -H- contrast, Texture -Y- energy	4
Sad	blue	1

Art image set

Amusement	Average Saturation, Median Saturation, Rule of Thirds - Saturation, pink, white, yellow, Itten Colors - Area warm, Texture -H- energy, Texture -H- homogeneity, Texture -S- contrast, Texture -Y- mean std, WangHist - f2, WangHist - f3, contrast	21
Anger	Wavelet texture -Hue -level 1, Wavelet texture -Hue -level 3, Itten Colors - Saturation Contrast, Texture -H- entropy, Texture -H- contrast, Texture -H- mean range, Texture -H- mean std	7
Awe	Itten Colors - Area cold	1

Content	pink, Texture -H- energy, Texture -S- contrast, Wavelet texture -Saturation -level 1, blue, Texture -Y- mean range, Itten Colors - Area cold	7
Disgust	Texture -H- contrast, Texture -H- mean range, Texture -H- mean std, Amount of skin - relative to face size, Texture -Y- correlation	5
Excitement	Itten Colors - Area cold, Average Saturation, Median Saturation, Rule of Thirds - Saturation, WangHist - f2, WangHist - f3, contrast, Wavelet texture -Saturation-level 3, orange, red, Line Slopes - dynamic, Itten Colors - Saturation Contrast weighted, Itten Colors - Hue Contrast, Texture -S- mean std, WangHist - f1, Arousal	32
Fear	Texture -H- contrast, Texture -S- contrast, Texture -Y- mean range, Texture -Y- mean std, Colorfulness, Wavelet texture - Lightness -level 1, Wavelet texture -Saturation -sum, Low DOF Indicator - Saturation, black, Line Slopes - static, Line Slopes - dynamic, Itten Colors - Hue count, Tamura - coarseness, Texture -S- homogeneity, Texture -S- mean range, Texture -Y- entropy, Texture -Y- homogeneity, Area-light, sharpness, mean Pleasure	20
Sad	Texture -S- contrast, Colorfulness, Wavelet texture -Saturation -sum, Low DOF Indicator - Saturation, Line Slopes - static, Line Slopes - dynamic, Itten Colors - Hue count, Tamura - coarseness, Texture -S- homogeneity, Texture -S- mean range, Texture -Y- entropy, Texture -Y- homogeneity, Area-light, sharpness, red, grey, purple, Tamura - contrast, Area-highsat	19

Abstract image set

Amusement	Texture -H- entropy, Texture -H- contrast, Texture -H- mean std, Area-light	4
Anger	Texture -H- entropy, Line Slopes - dynamic, Tamura - coarseness, Texture -S- contrast	4
Awe	Low DOF Indicator - Lightness	1
Content	brown, Amount of skin - relative to face size	2
Disgust	brown, Amount of skin - relative to face size, Tamura - coarseness, black, pink, white, Texture -H- homogeneity, Texture -Y- mean std, Pleasure	9
Excitement	Average Saturation, Median Saturation, Low DOF Indicator - Hue, orange, purple, Texture -S- homogeneity, Texture -S- mean std, Area-dark, Texture -Y- entropy, Texture -Y- contrast, Texture -Y- correlation, Texture -Y- energy, Texture -Y- homogeneity	12

Fear	Low DOF Indicator - Hue, Texture -S- mean std, pink, Texture -H- contrast, Texture -H- mean std, Wavelet texture -Hue -level 1, red, Tamura - direction, Texture -S- energy, Texture -S- mean range, WangHist - f1, Tamura - contrast, Tamura - direction, Texture -H- entropy, Texture -H- correlation, Texture -H- mean std, WangHist - f2, WangHist - f3, Area-verydark	35
Sad	Texture -H- homogeneity, Wavelet texture -Lightness -level 1, blue, green, Size of biggest face, Itten Colors - Area cold, Area-cold, Area-verydark, Itten Colors - Hue maximum range, Itten Colors - Hue count, WangHist - f2	17

Combined image set

Amusement	orange	1
Anger	Wavelet texture-S-level 3, Line Slopes - static	2
Awe	Itten Colors - Hue Contrast weighted	1
Content	Texture -S- entropy	1
Disgust	Rule of Thirds - Hue, Itten- Warmth Contrast, Pleasure	3
Excitement	Itten - Hue Contrast, Average Light, Median Light, Wavelet texture -S -level 2, Line Slopes - dynamic, Texture -H- contrast, Texture -S- correlation, Texture -S- energy, Texture -S- homogeneity, Texture -S- mean range, Texture -S- mean std, Texture -Y- entropy, Texture -Y- contrast, Texture -Y- correlation, Texture -Y- energy, Texture -Y- homogeneity	16
Fear	Texture -S- entropy, Wavelet texture -S -level 3 , Average Hue Vector Lenght - Saturation weighted, Wavelet texture -H -sum, Low DOF - Hue, blue, brown, Itten - Hue Contrast, Itten - Hue spread, Itten - Area warm, Itten - Harmony, Tamura - contrast, Tamura - direction, Texture -H- entropy, Texture -H- correlation, Texture -H- std, WangHist- f2, WangHist - f3, Area-verydark	21
Sad	Average Hue Lenght - Saturation weighted, Low DOF - Hue, Itten - Hue Contrast, Itten - Harmony, Tamura - direction, Texture -H- entropy, Texture -H- correlation, Texture -H- mean std, WangHist - f2, WangHist - f3, Area-verydark, Itten - Hue maximum range, Itten - Hue count	20

Feature set - Wrapper-based selection

IAPS

Amusement	Average Saturation, Median Light, Wavelet texture -Saturation -level 1, Low DOF Indicator - Hue, Low DOF Indicator - Saturation, Line Slopes - dynamic, Itten Colors - Saturation Contrast weighted, Texture -S- homogeneity, WangHist - f2, Area-middlesat, Area-highsat	13
Anger	Average Hue, Median Hue, Average Hue Vector Direction, Average Hue Vector Lenght - Saturation weighted, Rule of Thirds - Lightness, Wavelet texture -Hue -level 1, Wavelet texture -Hue -level 3, Wavelet texture -Saturation -level 1, Wavelet texture -Saturation -level 2, Wavelet texture -Saturation -level 3, Wavelet texture -Lightness -level 1, Wavelet texture -Lightness -level 3, Wavelet texture -Hue -sum, Wavelet texture -Saturation -sum, Low DOF Indicator - Saturation, Low DOF Indicator - Lightness, grey , pink , purple , Line Slopes - dynamic, Line Slopes - dynamic, Size of biggest face, Amount of skin - relative, Itten Colors - Saturation Contrast, Itten Colors - Hue Contrast, Itten Colors - Hue Contrast, Itten Colors - Warmth Contrast average, Itten Colors - Warmth Contrast max, Itten Colors - Warmth Contrast Strength average, Itten Colors - Area warm, Itten Colors - DisHarmony, Tamura - coarseness, Texture -H- entropy, Texture -H- contrast, Texture -H- energy, Texture -H- homogeneity, Texture -H- mean range, Texture -S- energy, Texture -Y- correlation, Texture -Y- energy, WangHist - f1, WangHist - f2, WangHist - f3, Area-dark, Area-middle-grey, Area-light, Area-warm, mean Arousal, Dominance	53
Awe	Wavelet texture -Hue -level 1, Wavelet texture -Hue -level 2, Wavelet texture -Lightness -level 1, Wavelet texture -Lightness -level 2, Wavelet texture -Lightness -sum, pink , purple , Itten Colors - Brightness Contrast, Itten Colors - Hue Contrast, Itten Colors - Warmth Contrast max, Tamura - direction, Texture -H- entropy, Texture -H- mean std, Texture -S- energy, Texture -S- mean std, Texture -Y- energy, Texture -Y- homogeneity, WangHist - f1, WangHist - f2, WangHist - f3, Area-lowsat, Area-highsat, mean Pleasure, mean Arousal, Dominance	27
Content	Average Light, Wavelet texture -Lightness -level 2, Low DOF Indicator - Saturation, pink , Line Slopes - static, Line Slopes - dynamic, Number of frontal faces, Size of biggest face, Itten Colors - Brightness Contrast weighted, Itten Colors - Hue Contrast, Itten Colors - Hue spread, Itten Colors - Warmth Contrast average, Itten Colors - Area cold, Tamura - coarseness, Texture -H- correlation, Texture -S- homogeneity, WangHist - f1, WangHist - f2, sharpness, mean Pleasure, mean Dominance	21

Disgust	Median Light, Average Hue Vector Lenght - Saturation weighted, Wavelet texture -Hue -level 1, Wavelet texture -Hue -level 2, Wavelet texture -Lightness -level 1, Wavelet texture - Lightness -level 2, Wavelet texture -Saturation -sum, grey , pink , Itten Colors - Brightness Contrast weighted, Itten Colors - Saturation Contrast weighted, Itten Colors - Hue Contrast weighted, Tamura - direction, Texture -H- homogeneity, WangHist - f1, WangHist - f2, mean Dominance, std Arousal	19
Excitement	Wavelet texture -Hue -level 1, Wavelet texture -Hue -level 2, Wavelet texture -Lightness -level 1, Wavelet texture -Lightness -level 2, Wavelet texture -Saturation -sum, grey , Itten Colors - Brightness Contrast weighted, Itten Colors - Warmth Contrast Strength average, Tamura - coarseness, Tamura - direction, Texture -H- energy, Texture -H- mean std, Texture -S- contrast, WangHist - f1, WangHist - f2, Area-light, Area-cold, sharpness, Arousal	20
Fear	Average Hue Vector Direction, Average Hue Vector Direction - Saturation weighted, Wavelet texture -Hue -level 1, Wavelet texture -Lightness -level 1, Wavelet texture -Lightness -level 2, Wavelet texture -Saturation -sum, Itten Colors - Brightness Contrast weighted, Itten Colors - Warmth Contrast average, Tamura - coarseness, Texture -H- mean std, Texture -S- contrast, Texture -S- mean range, Texture -Y- mean range, WangHist - f1, WangHist - f2, Area-verydark, Area-light, mean Pleasure, mean Arousal, Pleasure	22
Sad	Average Light, Average Hue Vector Direction, Average Hue Vector Lenght - Saturation weighted, Colorfulness, Wavelet texture -Saturation -level 3, Wavelet texture -Lightness -level 2, Level of Detail, Low DOF Indicator - Lightness, black , brown , white , Line Slopes - static, Line Slopes - dynamic, Line Slopes - dynamic, Itten Colors - Hue Contrast, Itten Colors - Hue spread, Itten Colors - Hue maximum complementary colors, Itten Colors - Warmth Contrast Strength average, Itten Colors - Area warm, Itten Colors - DisHarmony, Tamura - contrast, Texture -H- homogeneity, Texture -H- mean range, Texture -S- mean std, Texture -Y- correlation, Texture -Y- energy, Texture -Y- mean range, Texture -Y- mean std, WangHist - f1, WangHist - f2, WangHist - f3, Area-verylight, Area-warm, mean Pleasure, mean Arousal, Pleasure, Arousal, Dominance	39

Art image set

Amusement	Average Saturation, Average Hue, Average Hue Vector Direction, Average Hue Vector Lenght, Average Hue Vector Lenght - Saturation weighted, Wavelet texture -Hue -level 1, Wavelet texture -Hue -sum, Level of Detail, Low DOF Indicator - Hue, Low DOF Indicator - Lightness, orange , pink , Line Slopes - dynamic, Line Slopes - dynamic, Itten Colors - Saturation Contrast, Itten Colors - Saturation Contrast weighted, Itten Colors - Hue Contrast, Itten Colors - Hue Contrast, Itten Colors - Warmth Contrast average, Itten Colors - Area warm, Itten Colors - Area cold, Itten Colors - DisHarmony, Texture -H- entropy, Texture -Y- energy, WangHist - f1, WangHist - f2, Area-verydark, Area-verylight, Area-warm, Area-cold, Area-lowsat, mean Pleasure, mean Arousal	42
Anger	Average Light, Average Saturation, Average Hue Vector Direction, Wavelet texture -Hue -level 3, Wavelet texture -Hue -sum, Level of Detail, Low DOF Indicator - Lightness, pink , Line Slopes - dynamic, Line Slopes - static, Size of biggest face, Itten Colors - Brightness Contrast, Itten Colors - Saturation Contrast, Itten Colors - Hue Contrast, Itten Colors - Hue Contrast, Itten Colors - Warmth Contrast Strength average, Tamura - contrast, Texture -H- correlation, WangHist - f2, Area-cold, Area-lowsat, mean Arousal, mean Dominance, std Arousal	25
Awe	Average Saturation, Wavelet texture -Hue -level 3, Wavelet texture -Saturation -level 3, Wavelet texture -Lightness -level 2, Low DOF Indicator - Lightness, pink , Line Slopes - dynamic, Itten Colors - Brightness Contrast, Itten Colors - Hue Contrast, Tamura - contrast, Tamura - direction, Texture -H- energy, WangHist - f1, WangHist - f2, WangHist - f3, Area-lowsat, mean Arousal	18
Content	Average Hue, Wavelet texture -Saturation -level 2, Wavelet texture -Hue -sum, black , blue , brown , grey , pink , Line Slopes - dynamic, Line Slopes - static, Amount of skin - relative, Itten Colors - Brightness Contrast, Itten Colors - Hue Contrast, Itten Colors - Hue spread, Itten Colors - Hue complementary colors, Itten Colors - Hue maximum complementary colors, Itten Colors - Warmth Contrast average, Itten Colors - Warmth Contrast max, Itten Colors - Warmth Contrast Strenght max, Itten Colors - Area warm, Itten Colors - Area cold, Itten Colors - DisHarmony, Tamura - direction, Texture -H- entropy, Texture -H- contrast, Texture -H- mean range, Texture -H- mean std, Texture -S- correlation, Texture -S- mean std, Texture -Y- correlation, WangHist - f1, WangHist - f2, Area-middle-grey, Area-light, Area-verylight, Area-warm, Area-lowsat, Area-middlesat, mean Pleasure, mean Dominance, Pleasure, Arousal	46

Disgust	Average Saturation, Median Light, Wavelet texture -Saturation -level 1, Low DOF Indicator - Hue, Low DOF Indicator - Saturation, yellow , Line Slopes - dynamic, Amount of skin - relative, WangHist - f1, WangHist - f2	11
Excitement	Average Light, Median Hue, Average Hue Vector Direction, Average Hue Vector Direction - Saturation weighted, Average Hue Vector Lenght - Saturation weighted, Rule of Thirds - Lightness, Wavelet texture -Hue -level 3, Wavelet texture -Saturation -level 1, Wavelet texture -Lightness -level 1, Wavelet texture - Lightness -level 3, Wavelet texture -Hue -sum, Low DOF Indicator - Hue, pink , purple , Line Slopes - dynamic, Amount of skin - relative, Itten Colors - DisHarmony, Texture -S- contrast, WangHist - f1, WangHist - f2, Pleasure	24
Fear	Median Saturation, Wavelet texture -Hue -level 1, Wavelet texture -Hue -level 2, Wavelet texture -Lightness -level 1, Wavelet texture -Lightness -level 2, Wavelet texture -Lightness -level 3, Wavelet texture -Saturation -sum, grey , Line Slopes - dynamic, Itten Colors - Brightness Contrast weighted, Texture -H- energy, Texture -H- homogeneity, Texture -H- mean std, Texture -S- contrast, Texture -S- mean range, Texture -Y- mean range, WangHist - f1, WangHist - f2, Area-light, Area-cold, sharpness, Pleasure, Arousal	25
Sad	Average Hue Vector Lenght - Saturation weighted, Colorfulness, Wavelet texture -Hue -level 2, Wavelet texture -Lightness -level 1, Wavelet texture -Lightness -level 2, Wavelet texture - Saturation -sum, Wavelet texture -Lightness -sum, grey , purple , white , Line Slopes - dynamic, Size of biggest face, Itten Colors - Brightness Contrast, Itten Colors - Brightness Contrast weighted, Itten Colors - Saturation Contrast, Itten Colors - Warmth Contrast Strength average, Itten Colors - Warmth Contrast Strenght max, Tamura - coarseness, Texture -H- energy, Texture -S- contrast, Texture -S- mean range, Texture -Y- entropy, Texture -Y- correlation, Texture -Y- mean std, WangHist - f1, WangHist - f2, Area-light, Area-cold, sharpness, mean Pleasure, mean Dominance, Pleasure, Arousal	36

Abstract image set

Amusement	Average Light, Median Hue, Average Hue Vector Lenght - Saturation weighted, Rule of Thirds - Lightness, Rule of Thirds - Saturation, Rule of Thirds - Hue, Wavelet texture -Saturation -level 2, Wavelet texture -Saturation -level 3, Wavelet texture - Lightness -level 1, Wavelet texture -Lightness -level 2, black , Line Slopes - dynamic, Line Slopes - dynamic, Line Slopes - static, Itten Colors - Hue spread, Itten Colors - Warmth Contrast Strength average, Itten Colors - DisHarmony, Tamura - coarseness, Tamura - contrast, Texture -H- energy, Texture -S- entropy, Texture -Y- mean std, WangHist - f1, WangHist - f2, WangHist - f3, Area-cold, Area-lowsat, mean Arousal	30
Anger	Average Light, Average Hue, Median Light, Wavelet texture - Saturation -level 1, Wavelet texture -Saturation -level 3, Low DOF Indicator - Hue, Low DOF Indicator - Saturation, black , red , yellow , Line Slopes - dynamic, Line Slopes - static, Amount of skin - relative, Itten Colors - Brightness Contrast weighted, Itten Colors - Hue Contrast weighted, Itten Colors - Warmth Contrast max, Itten Colors - Area warm, WangHist - f1, WangHist - f2	21
Awe	Average Saturation, Average Hue, Average Hue Vector Direction - Saturation weighted, Average Hue Vector Lenght - Saturation weighted, Rule of Thirds - Saturation, Rule of Thirds - Hue, Wavelet texture -Saturation -level 1, Wavelet texture -Saturation -level 2, Wavelet texture -Saturation -level 3, Wavelet texture - Lightness -level 2, Level of Detail, green , purple , Line Slopes - static, Line Slopes - static, Itten Colors - Brightness Contrast, Itten Colors - Saturation Contrast, Itten Colors - Hue Contrast, Itten Colors - Warmth Contrast Strength average, Itten Colors - Area cold, Itten Colors - DisHarmony, Texture -H- contrast, Texture -S- correlation, Texture -Y- homogeneity, WangHist - f1, WangHist - f2, Area-light, Area-warm, Area-lowsat, contrast, mean Dominance, Dominance	36
Content	Average Saturation, Average Hue Vector Lenght, Rule of Thirds - Saturation, Wavelet texture -Hue -sum, Low DOF Indicator - Lightness, green , Line Slopes - dynamic, Itten Colors - Brightness Contrast, Itten Colors - Hue Contrast, Itten Colors - Warmth Contrast Strenght max, Itten Colors - Area cold, Tamura - contrast, Texture -H- contrast, Texture -H- correlation, WangHist - f1, WangHist - f2, Area-middle-grey, Area-cold, mean Arousal	22

Disgust	Average Light, Average Hue Vector Direction - Saturation weighted, Rule of Thirds - Hue, Wavelet texture -Hue -level 1, Wavelet texture -Hue -level 2, Wavelet texture -Saturation -level 2, Wavelet texture -Lightness -sum, Low DOF Indicator - Hue, grey , purple , Line Slopes - static, Line Slopes - static, Line Slopes - dynamic, Amount of skin - relative to face size, Itten Colors - Brightness Contrast, Itten Colors - Saturation Contrast, Itten Colors - Hue Contrast, Itten Colors - Hue spread, Itten Colors - Warmth Contrast max, Itten Colors - Warmth Contrast Strength average, Itten Colors - Warmth Contrast Strenght max, Tamura - contrast, Texture -H- entropy, Texture -H- mean range, Texture -S- contrast, Texture -S- energy, Texture -Y- entropy, Texture -Y- contrast, Texture -Y- correlation, Texture -Y- homogeneity, WangHist - f1, WangHist - f2, Area-light, Area-warm, Area-lowsat, sharpness, mean Pleasure, mean Arousal, Arousal, Dominance	45
Excitement	Average Light, Median Hue, Average Hue Vector Lenght, Average Hue Vector Lenght - Saturation weighted, Colorfulness, Rule of Thirds - Lightness, Wavelet texture -Hue -level 1, Wavelet texture -Saturation -level 1, Wavelet texture -Lightness -sum, Level of Detail, orange , pink , Line Slopes - static, Line Slopes - dynamic, Line Slopes - static, Line Slopes - static, Line Slopes - dynamic, Itten Colors - Brightness Contrast, Itten Colors - Saturation Contrast, Itten Colors - Hue Contrast, Itten Colors - Hue spread, Itten Colors - Hue maximum range, Itten Colors - Hue count, Tamura - coarseness, Tamura - contrast, Tamura - direction, Texture -Y- correlation, WangHist - f1, WangHist - f2, Area-cold, Area-lowsat, mean Arousal, Pleasure	37
Fear	Average Hue, Median Saturation, Median Hue, Average Hue Vector Direction - Saturation weighted, Colorfulness, Rule of Thirds - Saturation, Rule of Thirds - Hue, Wavelet texture - Saturation -level 1, Wavelet texture -Lightness -level 2, Wavelet texture -Hue -sum, Low DOF Indicator - Hue, blue , green , pink , Line Slopes - static, Number of frontal faces, Amount of skin - relative, Itten Colors - Brightness Contrast weighted, Itten Colors - Saturation Contrast, Itten Colors - Hue Contrast, Itten Colors - Hue Contrast weighted, Itten Colors - Hue Contrast, Itten Colors - Hue maximum range, Itten Colors - Warmth Contrast max, Itten Colors - Warmth Contrast Strength average, Itten Colors - Warmth Contrast Strenght max, Itten Colors - Area cold, Tamura - contrast, Texture -H- entropy, Texture -S- correlation, WangHist - f1, WangHist - f2, Area-light, mean Dominance	38

Sad	Average Saturation, Median Saturation, Average Hue Vector Direction, Average Hue Vector Lenght - Saturation weighted, Rule of Thirds - Saturation, Wavelet texture -Hue -level 1, Wavelet texture -Hue -level 2, blue , grey , pink , Line Slopes - dynamic, Line Slopes - static, Itten Colors - Hue Contrast, Itten Colors - Area cold, Texture -H- correlation, Texture -H- mean range, Texture -H- mean std, Texture -S- correlation, Texture -Y- energy, WangHist - f1, WangHist - f2, Area-cold, Area-lowsat, Area-highsat, contrast, mean Arousal, mean Dominance	32

Combined image set

Amusement	Wavelet texture -Saturation -level 1, Line Slopes - static, Texture -S- energy, WangHist - f1, WangHist - f2, Area-verylight	6
Anger	Itten Colors - Hue maximum range, Texture -Y- mean range, WangHist - f1, WangHist - f2, WangHist - f3, Area-lowsat	7
Awe	Wavelet texture -Saturation -level 1, Wavelet texture -Lightness -level 2, black , Line Slopes - static, Tamura - coarseness, Texture -S- contrast, Texture -Y- energy, WangHist - f1, WangHist - f3, Area-verylight	10
Content	Rule of Thirds - Hue, Line Slopes - static, WangHist - f2, Area-verylight	4
Disgust	Rule of Thirds - Hue, Line Slopes - static, WangHist - f2, Area-verylight	4
Excitement	Average Hue Vector Lenght - Saturation weighted, Rule of Thirds - Hue, Line Slopes - static, Itten Colors - Warmth Contrast Strenght max, Texture -H- correlation, Texture -Y- correlation, WangHist - f1, WangHist - f2, Area-verylight	9
Fear	Rule of Thirds - Hue, Line Slopes - static, WangHist - f2, Area-verylight	4
Sad	Average Hue Vector Lenght - Saturation weighted, Rule of Thirds - Hue, Wavelet texture -Saturation -level 1, Level of Detail, Line Slopes - static, Itten Colors - Warmth Contrast Strenght max, Area-verylight, Pleasure	8