Affective Image Classification using Features Inspired by Psychology and Art Theory

Jana Machajdik
Institute of Computer Aided Automation, Vienna
University of Technology
Favoritenstr. 9/183, 1040 Vienna, Austria
jana@caa.tuwien.ac.at

Allan Hanbury
Information Retrieval Facility
Tech Gate Vienna, Donau City Straße 1,
1220 Vienna, Austria
a.hanbury@ir-facility.org

ABSTRACT

Images 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. We investigate and develop methods to extract and combine low-level features that represent the emotional content of an image, and use these for image emotion classification. 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 testing and training, we use three data sets: the International Affective Picture System (IAPS); a set of artistic photography 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 and ratings on pictures without contextual content. Improved classification results are obtained on the International Affective Picture System (IAPS), compared to state of the art work.

Categories and Subject Descriptors

H.3.1 [**Information storage and retrieval**]: Content Analysis and Indexing; I.4.7 [**Image processing and computer vision**]: Feature Measurement

General Terms

Algorithms, Experimentation, Human Factors, Performance

Keywords

image affect, image classification, image features, art theory, psychology, emotional semantic image retrieval

1. INTRODUCTION

In recent years, with the increasing use of digital photography technology by the general public, the number 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 commercial systems use textual indexing to find the relevant images. To escape the limits of manual tagging, content-based image retrieval (CBIR) systems 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 [30]. Therefore, there is a current trend towards dealing with a higher-level of multimedia semantics. In this context, two levels are recognized [11]: *Cognitive level* and *Affective level*.

While in the 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 savanna 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" [11]. However there are situations when it would be helpful if images could be retrieved based on their emotional content. As an example, let's mention a magazine editor searching for illustrative photos for an extensive article on the topic of "depression". There is no specific subject to look for, but the images should have a "sad" atmosphere.

In [30], analysis and retrieval of images at the *affective level* is 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 [4]. 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 [30].

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" [11].

In this work, we concentrate on determining the affect of still images. The main goal of this work is to study features that are specific to the task of affective image classification.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
MM'10, October 25–29, 2010, Firenze, Italy.
Copyright 2010 ACM 978-1-60558-933-6/10/10 ...$10.00.

We exploit theoretical and empirical concepts from psychology and art theory to define image features that are specific to the domain of artworks with emotional expression. Specifically, the results of psychological experiments on emotional response to color [26], as well as work on color in art [15] are used. There are generally two approaches to modelling emotion: the *dimensional approach* [22] represents emotions as coordinates in a two or three dimensional space, while the *category approach* assigns descriptive words to regions of this space. As our experiments are done in a classification framework, we adopt the latter approach, where classification is done into eight emotional categories defined in a psychological study [21]. Experiments are performed on the International Affective Picture System (IAPS) [16], as well as on two further datasets collected for this study, which we make available to the research community.

We begin in the next section with a short overview of the state-of-the-art and a critique of existing work leading to a presentation of the contributions of this paper. Section 3 gives an overview of the framework used in this study. A detailed presentation of the features is given in Section 4. Finally, the experimental evaluation of the proposed features is presented in Section 5.

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 [17], but where affective retrieval and classification of visual or acoustic signals (digital media) is concerned, the publications are few, but the growing number of recent publications on the topic shows that the interest in this field is high.

One of the first ESIR systems is K-DIME [3], which builds an individual model for each user using a neural network. Another early work was done by Colombo et al. [5]: Based on Itten's color contrast theory [15] they define the features for hue, saturation, warmth and luminance contrast, and color harmony between distinct image regions. A fuzzy representation model is used to describe and store each property 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. Wang Wei-ning et al. [32] also developed features specific for affective image classification. However, other works, such as [14] and [33] were done using generic image processing features such as e.g. color histograms. In [34] a scene categorization algorithm using Gabor and Wiccest features was adapted and combined with machine learning to perform emotional valence categorization. Sung-Bae Cho [4] developed a human-computer interface for the purpose of aiding humans in the creative process of fields such as architecture, art, music and design. Work on the affective content analysis in movies is presented in [11] and [29].

2.1 Critique

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

generic features — Designing features that specifically encode knowledge about the effect of certain image characteristics on the emotion of the observer should lead to better separation of the emotion classes in feature space and hence better classification performance. Although some work suggests that art and impression specific features are of advantage [32, 7], some of the works in this field use common or generic features that are designed for example for object categorisation (e.g. in [3, 33, 34]).

arbitrary emotional categories — Often, the emotional categories used as output of the classification are ad hoc. As was shown in [29] and many psychological studies [22, 21], 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 categorization due to the use of the dimensional approach (e.g. [12]) to 35 "impression words" (in [14]). Furthermore, the output categories are on different *levels of significance* (according to [5]). Most of the "*kansei* impression words" as used in [14, 33, 32] or [3] are at the *expression level*, whereas emotional adjectives as used e.g. in [34] or [29] are at the higher, *emotional level*. These differences in categories makes result comparison difficult.

unpublished data sets — The data sets are in most cases unknown (unpublished). In many cases, no information is given on how the images were selected, for example if there was a manual filtering process that could potentially be biased. An exception is [34], in which the IAPS is used, and with which we compare our classification results.

missing or unclear evaluation - Another problem is presented by the poorly described evaluation measures (e.g. in [34]), incomplete description (e.g. [4]) or even lack of evaluation (e.g. in [12], [3]) of the presented methods.

Due to these factors, the results presented in the papers are not comparable. In this work we overcome these problems by:

1. choosing emotional categories defined in a rigorous psychological study [21].
2. using an image data set which is freely available for research purposes, the International Affective Picture System (IAPS) [16]. We also make the other two data sets used in this work available for research purposes[1].
3. describing our evaluation measures extensively and showing detailed results.

3. SYSTEM FLOW OF THE FRAMEWORK

The system flow of the processing framework used in this work is described here (and is also illustrated in Figure 1). First, the image database and corresponding ground truth labels is set up. Then, some preprocessing is done, which involves resizing the images, cropping away borders, converting the images from RGB to a cylindrical coordinate color space [10], and segmentation of each image into continuous regions by the waterfall segmentation algorithm [19].

[1] `http://www.imageemotion.org`

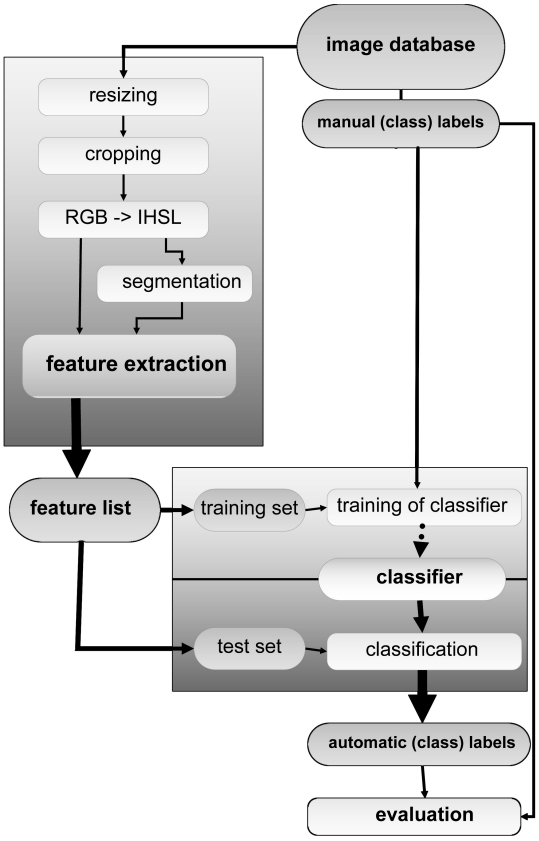

Figure 1: System flow.

The segmentation result is stored along with the original image. Both are input to the feature extraction process, which presents the core of the framework. During feature extraction all features are computed for each image and saved in a feature vector. After all features of all images have been computed, 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. 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.

As classifier, we chose the Naive Bayes classifier. We also evaluated other popular classifiers, such as Support Vector Machines (SVM), Random Forest or the C4.5 tree classifier, but in our case the Naive Bayes proved to be the best, both in performance and speed.

3.1 Preprocessing

At first all the images are resized to 200 000 pixels in total (or as close as possible while maintaining the aspect ratio). Images of this size contain enough detail, but allow faster calculation and consistent output. As the next step, "digital framing" or single-color borders added to the image are cropped away automatically by a combination of the Hough transform and the Canny edge detector. Furthermore, a conversion from RGB to a cylindrical coordinate color space without saturation normalization [10] is performed. This color space expresses an intuitive definition of colors by defining a well separated Hue (H), Saturation (S) and Brightness (Y) channel (see Figure 2). This color space has no dependence between the Saturation and Brightness and is therefore especially well suited for image analysis.

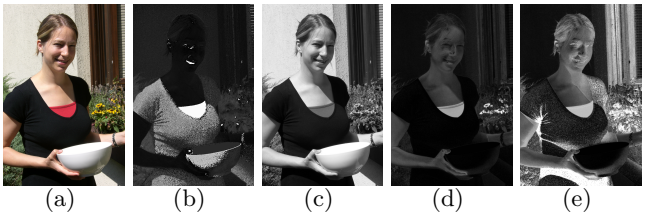

Figure 2: Cylindrical coordinate color space. (a) Color image; (b) hue; (c) brightness; (d) non-normalized saturation. For comparison (e) saturation using the HSV conversion formulae.

Finally, the images are segmented to characterize the spatial organization of the image, i.e. the composition. This is done using waterfall segmentation [19] on the images in the cylindrical coordinate color space. The waterfall segmentation algorithm takes both color and spatial information into account and results in an image separated into contiguous regions.

4. 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. We introduce features based on the experimentally-determined relation between color saturation and brightness, and emotion dimensions [26], as well as features based on relations between color combinations and induced emotional effects from art theory [15]. We complement these features by a selection of features, some of which are shown to be of use in similar image retrieval [24] and classification tasks [7, 32]. In this work, features representing the color, texture, composition and content were implemented. In the following we discuss the features in each group. All the features are summarized in Table 1.

4.1 Color

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 [15, 5]. 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 [15], psychology [26], color scientists [23] and marketing agents. 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. For measuring colors the following features were implemented:

Saturation and Brightness statistics were computed because saturation and brightness can have direct influence

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
	Colorfulness	1	colorfulness measure based on EMD
	Color Names	11	amount of black, blue, brown, green, gray, 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. [31] (*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*
texture	*Tamura*	3	features by Tamura et al [25]: *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
	Rule of Thirds	3	mean saturation, brightness and hue of the inner rectangle
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

Table 1: Summary of all features. The column '#' indicates the feature vector length for each type of feature.

on *pleasure, arousal* and *dominance*, the three axes of the emotion space according to the dimensional approach to emotions [22]. Hence in addition to the mean and standard deviation of the Saturation and Brightness channels, the direct relationships to *pleasure, arousal* and *dominance* are computed according to the formulae determined by Valdez and Mehrabian from their psychological experiments [26]. The experiments were conducted in a controlled environment, where 250 people were shown series of single color patches and rated them on a standardized emotional scale (Pleasure-Arousal-Dominance) to describe how they feel about the color. Several sessions of this kind of experiment were conducted, investigating different aspects of color and their relationship to emotions. The results of the analysis show a significant relationship between the brightness and saturation of color and their emotional impact as expressed in the following equations:

$$Pleasure = 0.69\ Y + 0.22\ S \quad (1)$$

$$Arousal = -0.31\ Y + 0.60\ S \quad (2)$$

$$Dominance = 0.76\ Y + 0.32\ S \quad (3)$$

Hue statistics are also computed, as the tone of the image is important. However, since Hue is measured in a circular way (in degrees), vector-based, circular statistics [20] must be used to compute measurements like mean, hue spread, etc.

Colorfulness is measured using the Earth Mover's Distance (EMD) between the histogram of an image and the histogram having a uniform color distribution, according to an algorithm suggested by Datta [7].

Color Names — each color has a special meaning and is used in certain ways by artists. We count how many pixels of each of the 11 basic colors (black, blue, brown, green, gray, orange, pink, purple, red, white, yellow) are present on the image using the algorithm of van de Weijer et al. [27].

Itten contrasts [15] are a powerful concept in art theory. Itten studied the usage of color in art extensively, and by his contrasts he formalized concepts for combining colors to induce an emotional effect in the observer and to achieve a harmonious image. Itten's color model characterizes colors according to hue, saturation and luminance. Twelve hues are identified as fundamental colors. These 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. The 12 pure colors are located along the equatorial circle, luminance 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 contrast: *contrast of saturation*, *contrast of light and dark*, *contrast of extension*, *contrast of complements*, *contrast of hue*, *contrast of warm and cold* and the *simultaneous contrast*. Additionally he formalized color combinations that look harmonious: the color accordances (see Figure 3). We translate these concepts into mathematical formulae, partly based on the work in [6].

First we transform the image into a simpler collection of colored patches by taking the regions created by the waterfall segmentation and computing the average Hue, Saturation and Brightness of each region. Further simplification is applied by translating the region's average values into the *Itten color model* [15] and so describing each region as being e.g. "dark", with "low saturation" and "green hue". The saturation and brightness in this model is encoded using the fuzzy membership functions defined in [32]. We do all further analysis of Itten contrasts on these regions, where each has an *Itten color* and a size. To measure the *contrast of light and dark* we use the standard deviation over the Brightness membership functions of all regions weighted by their relative size, and define the *contrast of saturation* in an analogue fashion. For the *contrast of hue* we use a vector based measurement of the hue spread. The contrast of complements is measured by computing the differences of hues between the segmented regions of the image. 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). If the contrast of complements is present, the value should be close to 180°. The contrast of warm and cold is defined in [6]. Each region is assigned three membership functions w_t that express the degree of cold ($t = 1$), neutral ($t = 2$) and warm ($t = 3$) in the region r_i (we chose the same definition of warm and cold as in [32], with neutral being $1 - (warm + cold)$). The strength of the *warm-cold contrast* between two regions is defined as: $\frac{\sum_{t=1}^{3} w_t(r_1) w_t(r_2)}{\sqrt{\sum_{t=1}^{3}(w_t(r_1))^2 \sum_{t=1}^{3}(w_t(r_2))^2}}$. To exploit this further we also measured the total amount of warm and cold area in the image. 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.

A combination of hues and tones that generates a stability effect on the human eye is said to be *harmonious*. *Harmony*, in this context, is an objective concept defined as the combination of those colors whose mix is gray. Applied to the spherical model, 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 3. To compute *harmony*, first we determine the main hues occurring in the image by creating a hue histogram with 12 bins (representing the 12 main colors on the Itten wheel in Figure 3) from the image, ignoring those bins which have less than 5% support. Usually this results in the 3–4 main hues that occur in the current image. We map these main hues onto the Itten color wheel and connect their positions to generate a polygon. *Harmony* is measured as the difference between the internal angles of the

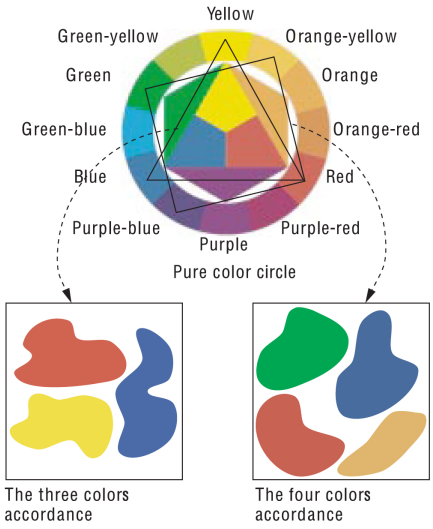

Figure 3: The concept of color accordance. Image from [15] labeled by [5].

generated polygon and the internal angles of a hypothetical regular polygon with the same number of vertices.

Color features by Wang Wei-ning et al. [32] form a specialized histogram designed to express the emotional impact of color in images. We implemented them in this work for the sake of comparison.

4.2 Texture

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 blur 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. Many features for describing texture have been developed. We have chosen some of those commonly used.

Wavelet-based features are introduced to measure spatial smoothness/graininess in images using the Daubechies wavelet transform [8]. As suggested in [7], we perform a *three-level* wavelet transform on all three color channels, Hue H, Saturation S and Brightness Y. Denoting the coefficients in level i for the wavelet transform of one channel of an image as w_i^h, w_i^v and w_i^d, the wavelet features are defined as follows (where $i = \{1, 2, 3\}$):

$$f_i = \frac{\sum_{x,y} w_i^h(x,y) + \sum_{x,y} w_i^v(x,y) + \sum_{x,y} w_i^d(x,y)}{|w_i^h| + |w_i^v| + |w_i^d|} \quad (4)$$

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

Tamura texture features [25] were successfully used in the field of affective image retrieval [33]. Therefore we decided to use the first three of the Tamura texture features: *coarseness*, *contrast* and *directionality*.

Gray-Level Co-occurrence Matrix (GLCM) [13] is another classic method of measuring texture. By means of

Figure 4: Rule of Thirds.

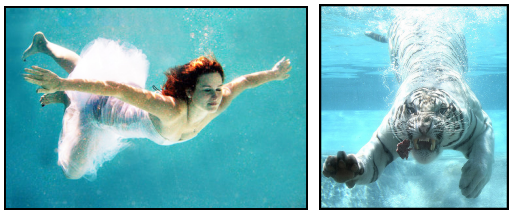

Figure 5: The content of an image is important. Two formally similar images, but with different emotional impact.

the GLCM we compute *contrast*, *correlation*, *energy*, and *homogeneity* of an image.

4.3 Composition

Harmonious composition is essential in a work of art and we need to consider it to analyze an image's character. 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 few aspects of composition, but we see this as an area where much improvement could be made in future.

Level of Detail expresses the observation that images with much detail generally 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 Alternating Sequential Filter size (filter size 3 and level 2 of the waterfall hierarchy). For simple images, this number will be low, and will be high for "busy" or cluttered images.

Low Depth of Field (DOF) is used by professional photographers to blur the background, thus simplifying the image, reducing the "busyness" and drawing the attention of the observer to the object of interest, which is sharp. In [7] a method is proposed to detect low DOF and macro images by computing a ratio of the wavelet coefficients in the high frequency (level 3 as used in the notation of Equation 4) of the inner part of the image against the whole image.

Dynamics - Studies suggest that lines in an image induce emotional effects [15, 2]. Horizontal lines are associated with a static horizon and communicate calmness, peacefulness and relaxation; vertical lines are clear and direct and communicate dignity and eternity; 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. We detect significant line slopes in images by using the Hough transform. 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 $(-15^\circ < \theta < 15^\circ)$ or if $(75^\circ < \theta < 105^\circ)$ and as slant otherwise. As a result we get the proportion of static and dynamic lines in the image.

Rule of Thirds represents a rule of thumb for good composition. Hereby the image is divided into thirds (see Figure 4). If it is heeded the main object lies on the inside or periphery of the inner rectangle. We measure the color statistics for the inner rectangle.

4.4 Content

The semantic content of the image has the greatest impact on the emotional influence of any picture. For example, if we consider the two images shown in Figure 5, 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 these two images 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 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 being developed go beyond the scope of this work. Nevertheless, we included two such algorithms that were available to us.

Human Faces in an image strongly draw the attention of human observers. Although the expression of the face is very important in order to distinguish between the moods of a picture, algorithms that can effectively recognize the emotional expression of a human face in static images are not yet fully mature. However, we can at least detect frontal faces (if there are any) in the picture by using the widely available face detection algorithm by Viola and Jones [28]. For each image the number of faces found and the relative size of the biggest face with respect to the image is used. This way, we can at least distinguish pictures with people from nature, landscape or object photography, and portraits from group shots.

Skin or rather its amount in the image can be used to detect "artistic nudes" images, which usually have a very specific emotional response. For this we use the algorithm presented in [18], 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 well suited to this task as there exists a predefined static model (thresholds on the Cb and Cr channels) that represents skin color well in many cases. However, the authors of [18] introduced an improvement to this method by including face detection. They use the face detection algorithm by Viola and Jones [28] to find faces in the image. If a face is found, the thresholds in the above model are altered to present a spectrum specific to the person found in the image. If more than one face is detected, the 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.

5. EVALUATION AND RESULTS

In this section, we present the experimental results on the classification of images into emotional categories. To generate discrete output categories, we use the emotional word list defined by Mikels et al. [21] in a psychological

study on affective images. Our emotional output categories are: *Amusement, Awe, Contentment, Excitement* as positive emotions, and *Anger, Disgust, Fear, Sad* to represent negative emotions. These emotional categories also correspond to the Ekman 1972 list of basic emotions [9], which are Anger, Disgust, Fear, Happiness, Sadness and Surprise. Surprise is omitted, while Happiness, the only "positive" Ekman emotion is split into four categories, resulting in an equal number of positive and negative emotions.

5.1 Data sets

For testing and training, we use three data sets: (1) the *International Affective Picture System (IAPS)* [16], (2) a set of 807 *artistic photographs* downloaded from an art sharing site [1], (3) a set of 228 peer rated *abstract paintings*. The fourth data set is a *combined set* of all of the above.

The *IAPS* is a common stimulus set widely used in emotion research. It consists of documentary-style natural color photos depicting complex scenes containing portraits, puppies, babies, animals, landscapes, scenes of poverty, pollution, mutilation, illness, accidents, insects, snakes, attack scenes and others. A selection of 394 of these pictures was categorized into the above discreet emotional categories in a study [21] (this dataset is also used by Yanulevskaya et al. [34] in a similar work).

The *artistic photographs* were obtained by using the emotion categories as search terms in the art sharing site, so the emotion category was determined by the artist who uploaded the photo. These photos are taken by people who attempt to evoke a certain emotion in the viewer of the photograph through the concious manipulation of the image composition, lighting, colors, etc. This dataset therefore allows us to investigate whether the conscious use of colors and textures by the artists improves the classification.

The *abstract paintings* consist only of combinations of color and texture, without any recognisable objects. They should therefore be well suited to classification by the proposed system, as it also does not attempt to model how specific objects evoke emotions. This dataset is very different to the IAPS dataset, as in the latter dataset, emotions are often evoked due to the presence of a certain object in the image. To obtain ground truth for the abstract paintings dataset, the images were peer rated in a web-survey where the participants could select the best fitting emotional category from the ones mentioned above for 20 images per session. 280 images were rated by approximately 230 people, where each image was rated about 14 times. For each image the category with the most votes was selected as the ground truth. Images where the human votes where inconclusive were removed from the set, resulting in 228 images.

Table 2 shows the number of images per data set and emotional category. At the same time it shows the main drawback of these data sets, namely that all the sets are unbalanced in terms of the number of examples per class and therefore the choice of classifier and evaluation measure must be considered carefully in order to obtain valid results.

5.2 Experiments

For evaluation, we conducted several experiments to measure the performance of our features and compare the results with that of Yanulevskaya et al. [34] and also to Wang Wei-ning et al. [32].

The experimental setup was as follows: each category was separated against all others in turn, in other words a "one category against all" setup. We separate the data into a training and test set using K-fold Cross Validation ($K = 5$). Since we do not have a balanced data set in terms of the number of examples per category, but the probabilities for the categories should be the same for all, we optimize for the *true positive rate per class* averaged over the positive and negative classes, 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 sub-sample the classes. This measure is also used as the evaluation measure for the experimental results.

In [34] 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 [34], we perform dimensionality reduction on the feature vectors, for which we use three algorithms. Two feature selection algorithms are used: a wrapper-based subset evaluation with a genetic search algorithm (referred to as the *wrapper-based method*), and an algorithm based on the classification performance of a single feature at a time, selecting only those features which resulted in an average true positive rate per class higher than 0.51 (referred to as the *single feature method*). We also use a feature extraction algorithm, *principal component analysis* (PCA). Results using all features were also computed. Following Yanulevskaya's approach we select our best feature subsets for each category (as the best performing subset from the above feature selection and extraction algorithms) and compare their performance to the best results in [34].

Wang Wei-ning et al. [32], in contrast, used an unknown data set as well as different categories. To get comparable results, we reimplemented the features from [32] (further referred to as "Wang Histogram") and used them in our classification scheme.

5.3 Results

Table 3 shows our resulting best feature selection for each data set and each category, along with the dimensionality reduction method producing each feature set listed in the last row. For the IAPS set, the Yanulevskaya et al. features give the best performance for the *Anger* and *Contentment* classes. For these classes, we show the best selection from the features implemented in this study. It is clear that the best feature set is dependent on both the category and the data set. The occurrence and size of human faces was clearly the strongest feature for the *Amusement* category of the IAPS image set. The categories in the IAPS set are strongly content related, e.g. *Fear* and *Disgust* images often show snakes, insects or injuries, whereas *Amusement* images often contain portraits of happy people. The classifier has the best performance when it basically detects the portraits. However, there is no such strong connection between faces and categories in the artistic sets. Instead the colors become much more important for the *artistic photos*. We see that features based on art theories (Itten colors, Wang Wei-ning histograms) are indeed most often selected for the *artistic photos* set. A large number of Itten features are selected for the *Amusement* and *Excitement* classes. The color features developed by Wang Wei-ning et al. [32] are also effective for this task — these features on their own had the best performance for the *Awe* and *Disgust* classes, and were also selected by the feature selection algorithms for the *Amusement* and *Excitement* classes.

	Amusement	Anger	Awe	Contentment	Disgust	Excitement	Fear	Sad	sum
IAPS	37	8	54	63	74	55	42	61	394
Art photo	101	77	103	70	70	105	115	166	807
Abstract paintings	25	3	15	63	18	36	36	32	228
Combined	163	88	172	196	162	196	193	259	1429

Table 2: Number of images per data set and emotion category. As can be seen the image sets are unbalanced.

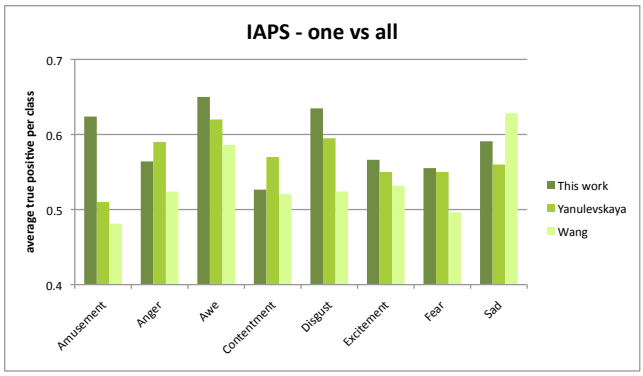

Figure 6: Classification performance for IAPS taking our best features for each category compared against the best features from [34] (Yanulevskaya) and the feature set described in [32] (Wang).

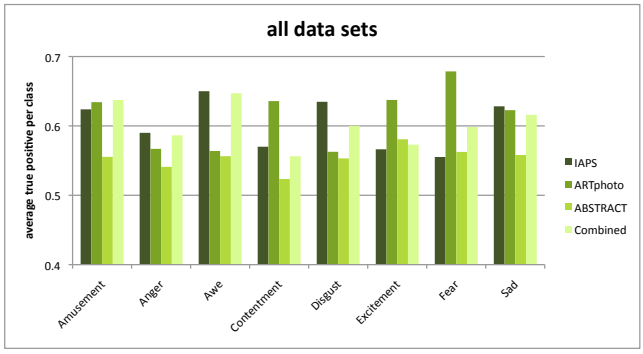

Figure 7: Classification performance for all image sets used in this work. The results are from the best feature selections implemented during this work.

In Figure 6 the results for the IAPS are shown, taking the best performing features for each class, and comparing them to the best performing features from [34] as well as to the performance of the features suggested in [32]. As can be deduced from the diagram, our feature sets outperform both the results by Yanulevskaya and the Wang Wei-ning features on the IAPS data set for five of eight categories. This shows that features created specifically for the task of affective image classification perform well, outperforming the more generic features of Yanulevskaya et al. for the majority of the classes. Figure 7 shows classification results for the best features for every data set used in this work. It shows no clear difference in results for the IAPS and the *artistic photo* image sets. For half of the classes the results are better for the IAPS set, for the other half for the *artistic photo* set.

The classification of the *abstract paintings* data set has

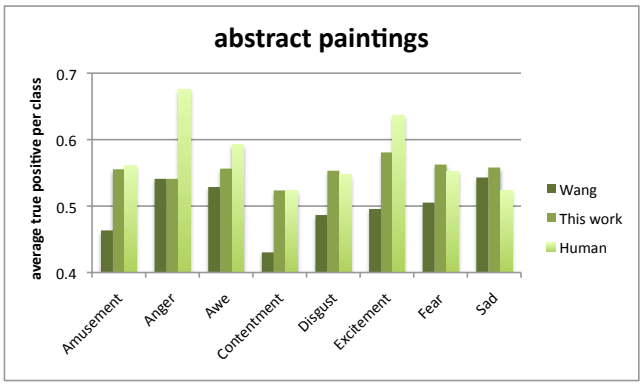

Figure 8: Classification performance of the *abstract paintings* image set taking our best features for each category.

the worst performance. In Figure 8 we compare the performance of our best performing features for each class for the abstract paintings data set with the results achieved with the features from [32], but also with humans, based on the multiple responses in the web survey. The bars labeled "Human" in Figure 8 where calculated by taking the number of votes for the category with the most votes and dividing it by the overall number of votes that the picture received. This gives us a measure of agreement among the survey participants, which can be interpreted as confidence in the winning emotion category. There is generally good correlation between the level of agreement of the humans and the results of the proposed algorithm. The largest difference is in the anger class, but this is likely due to the small number of images in this class, which makes the classification results less reliable.

6. CONCLUSIONS

We used a selection of features specific to the problem of affective image classification and achieved results that are better than comparable state of the art. There is nevertheless potential for improvement, both in developing better features and better classification. Especially semantic-based features, such as the recognition of the emotional expression of faces, or certain common symbols (e.g. hearts) could be of advantage. For the classification, rather than forcing each image to be in a single emotional category, an "emotional histogram" showing a distribution over the categories could be produced.

Furthermore, more reliable ground truth from a larger number of people is required — this will set an upper limit on the classification accuracy that should be achieved by such approaches. A further potential development is learning the preferences of individual people, instead of the consensus-based approach adopted in this work.

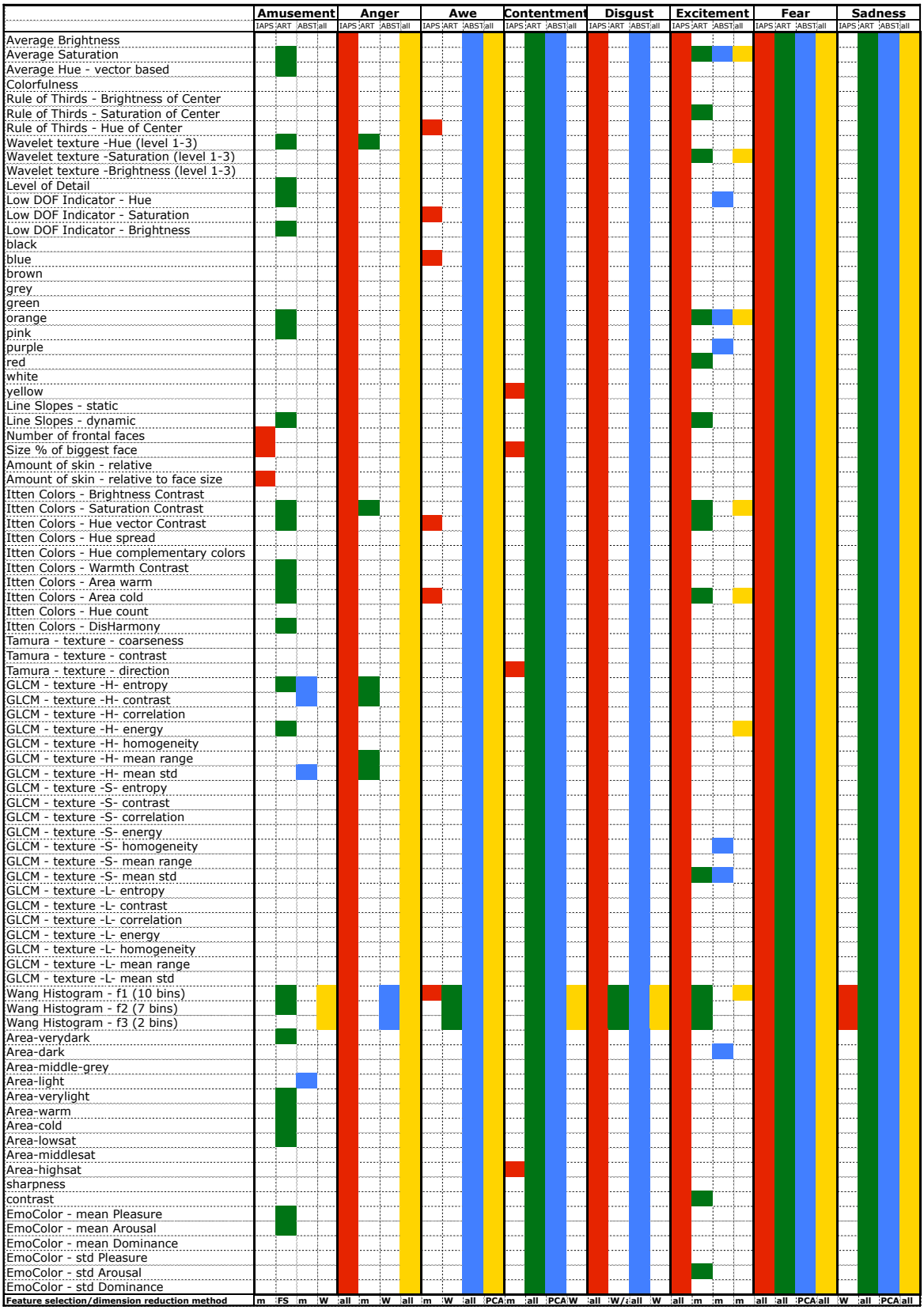

Table 3: Feature selection results for each image set and each emotion category. (data sets: IAPS (red), ART=artistic photos (green), ABST= abstract paintings (blue), all=all image sets combined (yellow), dimensionality reduction methods: m=single feature method, FS=wrapper-based method, W= Wang[32] features, PCA= linear combinations of features created by PCA, all= using all features/no selection)

7. REFERENCES

[1] deviantart. www.deviantart.com.

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

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

[4] S.-B. Cho. Emotional image and musical information retrieval with interactive genetic algorithm. *Proc. of the IEEE*, 92(4):702–711, 2004.

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

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

[7] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Studying aesthetics in photographic images using a computational approach. In *ECCV (3)*, pages 288–301, 2006.

[8] I. Daubechies. *Ten Lectures on Wavelets*. Regional Conf. Series in Applied Mathematics. Soc for Industrial & Applied Math, December 1992.

[9] P. Ekman, W. V. Friesen, M. O'Sullivan, A. Chan, I. Diacoyanni-Tarlatzis, K. Heider, R. Krause, W. A. LeCompte, T. Pitcairn, P. E. Ricci-Bitti, K. Scherer, M. Tomita, and A. 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.

[10] A. Hanbury. Constructing cylindrical coordinate colour spaces. *Pat. Rec. Lett.*, 29(4):494–500, 2008.

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

[12] A. Hanjalic and L.-Q. Xu. Affective video content representation and modeling. *IEEE Transactions on Multimedia*, 7(1):143–154, 2005.

[13] R. Haralick and L. Shapiro. *Computer and Robot Vision*. Addison-Wesley Longman, 1992.

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

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

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

[17] M. S. Lew, N. Sebe, C. Djeraba, and R. Jain. Content-based multimedia information retrieval: State of the art and challenges. *ACM Trans. Multimedia Comput. Commun. Appl.*, 2(1):1–19, 2006.

[18] C. Liensberger, J. Stöttinger, and M. Kampel. Color-based and context-aware skin detection for online video annotation. In *Proc. IEEE 2009 Int. Workshop on Multimedia Signal Processing*, 2009.

[19] B. Marcotegui and S. Beucher. Fast implementation of waterfall based on graphs. In C. Ronse, L. Najman, and E. Decencière, editors, *Mathematical Morphology: 40 Years on*, volume 30 of *Computational Imaging and Vision*, pages 177–186. Springer-Verlag, Dordrecht, 2005.

[20] K. V. Mardia and P. E. Jupp. *Directional Statistics*. Wiley, 1972.

[21] J. A. Mikels, B. L. Fredrickson, G. R. Larkin, C. M. Lindberg, S. J. Maglio, and P. A. Reuter-Lorenz. Emotional category data on images from the international affective picture system. *Behavior Research Methods*, 37(4):626–630, 2005.

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

[23] L.-C. Ou, M. R. Luo, A. Woodcock, and A. 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.

[24] J. Stöttinger, J. Banova, T. Pönitz, A. Hanbury, and N. Sebe. Translating journalists' requirements into features for image search. *Int. Conf. on Virtual Systems and Multimedia*, 2009.

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

[26] P. Valdez and A. Mehrabian. Effects of color on emotions. *Journal of Experimental Psychology: General*, 123(4):394–409, 1994.

[27] J. van de Weijer, C. Schmid, and J. Verbeek. Learning color names from real-world images. *IEEE CVPR*, pages 1–8, 2007.

[28] P. Viola and M. Jones. Robust real-time face detection. *Int. Journal of Computer Vision*, 57(2):137–154, 2004.

[29] H. L. Wang and L.-F. Cheong. Affective understanding in film. *Circuits and Systems for Video Technology, IEEE Transactions on*, 16(6):689–704, 2006.

[30] W. Wang and Q. He. A survey on emotional semantic image retrieval. *15th IEEE Int. Conf. on Image Processing*, pages 117–120, 2008.

[31] W.-N. Wang and Y.-L. Yu. Image emotional semantic query based on color semantic description. In *Machine Learning and Cybernetics. Proc. 2005 Int. Conf. on*, volume 7, pages 4571–4576, 2005.

[32] W. Wei-ning, Y. Ying-lin, and J. Sheng-ming. Image retrieval by emotional semantics: A study of emotional space and feature extraction. *IEEE Int. Conf. on Systems, Man and Cybernetics*, 4(Issue 8-11):3534 – 3539, Oct. 2006.

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

[34] 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 Int. Conf. on Image Processing*, 2008.