Color Emotions for Multi-Colored Images

Martin Solli,* Reiner Lenz

Department of Science and Technology (ITN), Linköping University, SE-60174 Norrköping, Sweden

Received 5 September 2009; revised 15 October 2009; accepted 29 October 2009

Abstract: We investigate the emotional response to colors in ordinary multicolored images. In psychophysical experiments, using both category scaling and interval scaling, observers are asked to judge images using three emotion factors: activity, weight, and heat. The color emotion metric was originally developed for single colors, and later extended to include pairs of colors. The same metric was recently used in image retrieval. The results show that people in general perceive color emotions for multi-colored images in similar ways, and that observer judgments correlate with the recently proposed method used in image retrieval. The intended usage is in retrieval systems publicly available on the Internet, where both the user and the viewing environment is unknown, which requires novel ways of conducting the psychophysical experiments. © 2010 Wiley Periodicals, Inc. Col Res Appl, 36, 210–221, 2011; Published online 29 April 2010 in Wiley Online Library (wileyonlinelibrary.com). DOI 10.1002/col.20604

Key words: color emotion; color image retrieval; color perception

INTRODUCTION

Most of us live in a world full of colors, usually perceived in an infinite number of multicolored combinations. Our colorful environment affects us in many ways. It is known that the relationship between colors and human emotions has a strong influence on how we perceive our environment. Naturally, the same holds for our perception of images. All of us are in some way emotionally affected when looking at a photograph or an image. One can often relate some of the emotional response to the context, or to particular objects in the scene, like familiar faces, etc. Simultaneously, as part of the color perception process, also the color content of the image will affect us emotionally. Such emotions, generally called color emotions, can be described as emotional feelings evoked by a single color or color combinations. They are typically expressed with semantic words, such as "warm," "soft," "active," etc. Color emotions, together with color memory, color harmony, color meaning, etc., belong to the cognitive aspects of colors. The original motivation for this research was to include high level semantic information, such as emotional feelings, in image classification and image retrieval systems. Emotional responses based on objects, faces, etc. are often highly individual, and therefore one has to be careful when including them in the management of general image databases. However, the emotional response evoked by color content is more universal. Consequently, the question dealt within this article is whether the emotional response related to colors in ordinary multicolored images are similar between persons, like previous research has shown for color emotions related to single colors and two-color combinations. If so, one can combine color emotions with Content Based Image Retrieval and discover new ways of searching for images using semantic properties of colors. The intended use of the method is primarily as a tool for sorting or preselecting images. For instance, a popular keyword-query may result in a huge set of images that are impossible for the user to review. New methods that can help the user by selecting a subset of the query result, or grade the images found, are therefore highly desirable.

In Solli and Lenz,[1] we have proposed a method using color emotions in image retrieval. The color emotion metric used there is obtained from studies made by Ou *et al.*[2–4] From psychophysical experiments Ou *et al.* derived color emotion models for single colors and two-color combinations. By factor analysis they identified three color-emotion factors: activity, weight, and heat. We demonstrated how to use Ou's models together with ordinary RGB-histograms of images to obtain compact but efficient image descriptors. As the method only involves transformations on ordinary RGB-histograms, usually already computed in many Content Based Image

*Correspondence to: Martin Solli (e-mail: Martin.Solli@itn.liu.se).

Contract grant sponsor: Knowledge Foundation, Sweden (*Visuella Världar*).

Retrieval systems, the method is very time efficient, which is essential when dealing with very large image databases containing millions or billions of images. The earlier method describing how to use color emotions in Content Based Image Retrieval is summarized in one of the following sections. A comprehensive description can be found in Solli and Lenz.[1]

Research on color emotions for single colors and two-color combinations has received considerable attention in the past. However, similar investigations for multicolored images are less common, which probably is one of the explanations why color emotions are seldom used in image retrieval systems. In this article, we use methods from psychophysical scaling to evaluate the retrieval methods used in Solli and Lenz.[1] At the same time, the evaluation will also indicate if people perceive color emotions for multicolored images in similar ways, or if judgments are highly individual. To investigate people's emotional responses a set of psychophysical experiments are conducted. Observers are judging images on different scales or factors related to color emotions. Both category scaling and interval scaling methods are used. Similarities and dissimilarities among observer judgments are presented and discussed, together with results and statistics for the correlation between user judgments and the proposed method used in image retrieval. The findings show that people do perceive color emotions for multi-colored images in similar ways, and that the recently proposed method for color emotions in image retrieval can be a useful tool in Content Based Image Retrieval and image indexing. Since the findings of this study will be used in retrieval systems typically accessible on a public web page, some unusual conditions are present. Basically, we have unknown users in unknown environments. A discussion on how to handle such conditions is included in the "Limitations" section.

The article is organized as follows. Related work is reviewed in the next section, followed by a section describing the special limitations and conditions that are present in this study. The subsequent section describes the predictive model used for estimation color emotions for images. Next, the psychophysical experiments are explained, and the results are discussed. Then the intended application in Content Based Image Retrieval is illustrated. Finally, conclusions are drawn, and some ideas about future work are presented.

RELATED WORK

Research on color emotions for single colors and two-color combinations is by now a well established research area. In a series of papers, Ou *et al*.[2–4] investigated the relationship between color emotions and color preference. Color emotion models for single colors and two-color combinations are derived from psychophysical experiments. Observers were asked to assess single colors on 10 color emotion scales. It is then shown that factor analysis can reduce the number of color emotions scales to only three categories, or color emotion factors: activity, weight and heat. Ou *et al*. conclude that the three factors agree with studies done by others, for instance Kobayashi[5] and Sato *et al*.[6] In this article, we will use those emotion factors when investigating color emotions for multicolored images. In another study of human's emotional response to colors, Gao and Xin[7] selected twelve pairs, or scales, of color emotion words which were considered fundamental. They also show that most of the variance in the data can be represented by fewer factors than twelve. By maximum likelihood factor analysis they group scales into three categories or indexes, called activity, potency, and definition. One of their conclusions is that color emotion connotations are mainly connected to lightness and chroma, and less connected with hue. One important question is whether color emotions are influenced by different regional or cultural backgrounds. In an extensive study by Gao *et al*.[8] it was concluded that the influence of cultural background on color emotions is very limited. In psychophysical experiments totally 214 color samples were evaluated on 12 emotion variables by subjects from seven different regions worldwide. Using factor analysis they show that a smaller number of factors are needed for the representation, which corresponds well to other studies. Another conclusion, common with others studies, is that lightness and chroma are the most important factors in color emotions, whereas the influence of hue is limited. Similar results about regional and cultural backgrounds were earlier found in cross-regional comparisons by Xin *et al*.[9,10] Also age-related differences were investigated. A recent example is Beke *et al*.,[11] where color preference of aged observers are compared to young observers. The results indicate important differences, both depending on neuro-physiological changes, and other aspects such as cultural implications.

Related to the problem of color emotion is the concept of color harmony. These research areas often share methods and ideas in their attempt to perform psychophysical experiments and also when creating predictive models. As an example, we mention the extensive study by Ou and Luo,[12] where harmony judgments of two-color combinations are investigated in order to develop a quantitative model for the prediction of harmony.

There are few articles addressing the problem of including color emotions in image retrieval. The methods presented are often focusing on semantic image retrieval in a more general way. Wang and Yu[13] propose an emotional semantic query model based on image color semantic descriptors. Images are segmented by clustering in the CIELAB color space. Then images are converted to the CIELCh color space (the cylindrical version of CIELUV), and segmented regions are converted to semantic terms through a fuzzy clustering algorithm. Both regional and global semantic descriptors are extracted. The user is able to query the image database with emotional semantic words, like "sad" and "warm," and also with more complex sentences. One interesting detail to notice is that they

use Gaussian low-pass filtering to remove edges before segmentation, with the motivation that the capability of the human visual system to distinguish different colors drops rapidly for high spatial frequencies. However, what is not discussed in the article is that one should probably be careful with the amount of filtering. As an example, we can consider an image containing stripes, where the semantic response may change rapidly with different amounts of low-pass filtering. Also Corridoni *et al.*[14] use clustering in the color space to segment images into regions with homogenous colors. Then Itten's formalism together with fuzzy sets are used to represent intraregion properties (warmth, hue, luminance, etc.) and inter-region properties (hue, saturation, luminance contrast, etc.). Properties are gathered to a color description language based on color semantics. Querying the database is accomplished through a rather complex user interface including sketches and dialog boxes.

In a paper by Hong and Choi[15] a search scheme called FMV (Fuzzy Membership Value) Indexing is presented. It allows the user to retrieve images based on high-level semantic concepts, using keywords such as "cool," "soft," "romantic," etc. Emotion concepts are derived from color values in the HSI color space. Cho and Lee[16] developed an image retrieval system based on human preference and emotion by using an interactive genetic algorithm (IGA). Image features are created from average colors and wavelet coefficients. Yoo[17] proposes an emotion-based image retrieval method using descriptors called query color code and query gray code. The descriptors are based on human evaluation of color patterns on 13 emotion pairs or scales, most of them related to color. The image database is queried with one of the emotions, and a feedback method is utilized for dynamically updating the search result.

We conclude with two recent papers discussing emotion-based image retrieval. Wang *et al.*[18] use a three-dimensional emotional space (with some similarities to the emotion space used in this article) for annotating images and perform semantic queries. The space is based on psychological experiments with 12 pairs of emotion words. Image properties are described with different kinds of histograms, and from histogram features emotional factors are predicted using a support vector machine. They create a search interface where the user can create semantic queries based on one of the emotion words. A disadvantage with the presented work is that emotion scales are derived from category scaling without, for instance, anchor images, which is not the most reliable scaling method (discussed later in this article). The method was developed and evaluated for paintings only. In Lee *et al.*[19] the authors show how rough set theory can be used to build an emotion-based color image retrieval system. Emotion data is extracted by letting people observe different random color patterns in category scaling experiments. Three different emotion scales are incorporated: warm–cool, dynamic–static, and heavy–light. The primarily field of application seems to be different color patterns, like wall papers etc. But the authors mention that the method can also be applied to image retrieval.

Common for the image retrieval methods mentioned earlier is that the evaluations are limited and user studies are missing. It is hard to know whether the models agree with observer judgments. Unique for the retrieval method related to this article is that, by this study, a comprehensive user study is incorporated in the overall presentation.

LIMITATIONS

The main goal of our research is to develop methods that can be implemented in large scale image indexing, especially search engines for large image databases publicly available on the Internet. In other words, the findings of this study will be used in a situation rather different from the situation encountered in traditional color emotion experiments. The main concern is that, since results will be used on the Internet, we have no control over the user environment. An unknown adaptation is present, which may influence the result in many ways. For instance, users may have very different monitor settings. Also the viewing environment, like illumination settings, background colors, etc. may vary significantly. Moreover, the psychological state of the user is unknown. As related research about color emotion models usually is carried out in controlled environments, it is of course questionable if the same models can be applied in an uncontrolled environment. However, we believe that if the entire evaluation process is based on user tests on the Internet, where we have no control over environmental settings, and still the results are successful, then we can also apply the models in an uncontrolled environment.

Another fact to be aware of is that image content in general, not only the color content, will affect the user's emotional response. Some images may even have strong emotional content closely related to those emotions evoked by the color content. For instance, some may experience that a photo of a sun-drenched beach brings a warm feeling, even though the color of the water can be perceived as rather cold. And observers familiar with cold climates may perceive a photo containing snow and ice as cold, even if the actual color content is rather neutral. But as long as the user is aware of that the emotional search is purely based on color content, we believe most users will both notice and accept images containing other emotional content. Finally, since tests are carried out anonymously on public web pages, we have to face the risk that some users will deliberately submit answers disagreeing with their own emotional feelings.

PREDICTIVE MODEL

This section describes the model used for predicting the emotional response of ordinary multicolored images. The model was earlier presented in Solli and Lenz.[1] We refer readers to the original paper for more details.

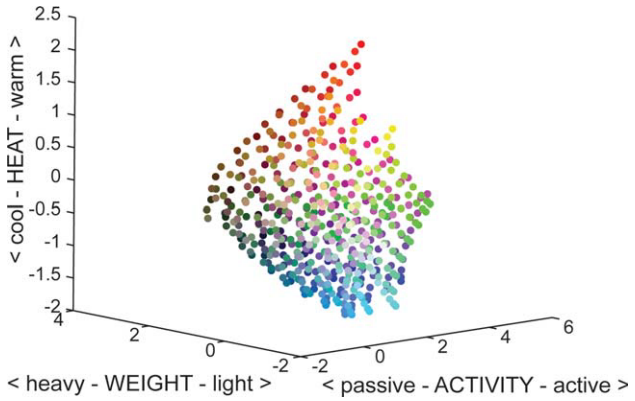

FIG. 1. Locations of the 512 bins from the RGB-histogram in the three-dimensional emotion space spanned by activity, weight, and heat. Bins (dots) are colored with the corresponding RGB colors.

Color Emotions

In a series of articles Ou *et al.*[2–4] investigated the relationship between color emotion and color preference. Color emotion models for single colors and two-color combinations are derived from psychophysical experiments. Observers were asked to assess single colors on 10 color emotion scales. Then they show that factor analysis can reduce the number of color emotions scales to only three categories, or color emotion factors: activity, weight, and heat

$$\text{activity} = -2.1 + 0.06 \times \left[(L^* - 50)^2 + (a^* - 3)^2 + \left(\frac{b^* - 17}{1.4}\right)^2\right]^{\frac{1}{2}} \quad (1)$$

$$\text{weight} = -1.8 + 0.04(100 - L^*) + 0.45\cos(h - 100^\circ) \quad (2)$$

$$\text{heat} = -0.5 + 0.02(C^*)^{1.07}\cos(h - 50^\circ) \quad (3)$$

$$h = \arctan\left(\frac{b^*}{a^*}\right) \quad (4)$$

$$C^* = \sqrt{a^{*2} + b^{*2}} \quad (5)$$

where L^*, a^* and b^* are CIELAB coordinates. In the next section, we describe how to use these emotion factors for multicolored images.

Color Emotions for Images

Using RGB-histograms for measuring statistical properties of color images is very common in Content Based Image Retrieval. We make use of this and utilize a method that will transform ordinary RGB-histograms of images to emotion descriptors. Typically, the histograms consist of 512 entries with eight quantization levels (bins) per color channel. RGB-histograms for all images in the database are collected and saved in a matrix H of size $N \times 512$ (rows $\times$ columns), where N is the number of images in our database. For each bin in the histogram we calculate the corresponding color emotion vector, using Eqs. (1) to (5). The result is saved in the matrix E of size 512×3 (three emotion values for each bin). The 512 RGB-bins plotted in the emotion space spanned by activity, weight, and heat can be seen in Fig. 1. In all calculations, we assume images are saved in the commonly used sRGB color space, and we use the standard illumination D50 when transforming sRGB values to CIELAB values. Next, the histogram matrix H is multiplied with the emotion matrix E to obtain a matrix $C = H \cdot E$, of size $N \times 3$. The n-th row in H, denoted by h_n describes the probability distribution of the RGB vectors in image n. The n-th row in C, denoted by C_n, obtained by the scalar product $c_n = h_n \cdot E$, is thus an expectation vector describing the expected value of the color emotion vectors of the pixels in image number n in our database. In other words, this vector contains the predicted mean score for each of the emotion factors: activity, weight, and heat. In Fig. 2, thirty images are plotted according to their predicted emotion coordinates in the three-dimensional emotion space.

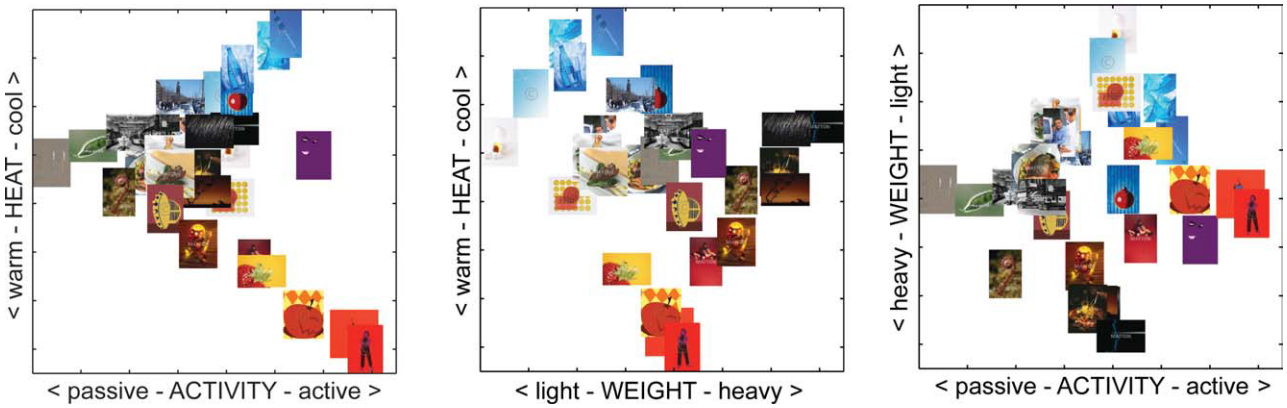

FIG. 2. Three different views of 30 images plotted according to their emotion coordinates in the three-dimensional emotion space spanned by activity, weight, and heat.

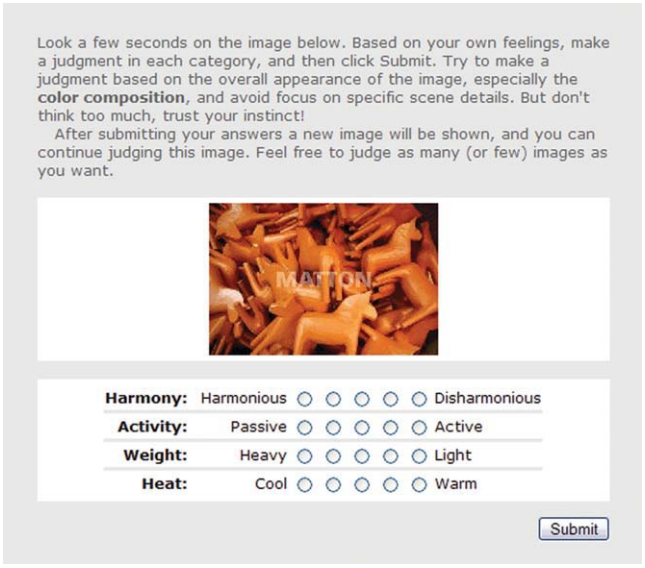

FIG. 3. The Graphical User Interface for the pilot study (Results from the Harmony factor are presented elsewhere.).

PSYCHOPHYSICAL EXPERIMENTS

When designing the user study, we followed the methodology presented by Engeldrum[20] to a large extent. The study consists of two tests, following each other in time. First a pilot study, that will indicate if we are aiming in the right direction. The study will also help us to select images, from now on called samples, for the second test. The pilot study is designed to be as simple as possible for the observer. The second test, here called the main study, is slightly more demanding for the observer. The advantage though is that the results are more reliable, and easier to evaluate. Our final conclusions are mainly based on the main study. Both tests are using web pages for communicating with the observer. The reason for describing both the pilot study and the main study is that the main study is to some extent based on results and experience gained in the pilot study.

The Pilot Study

In this study, observers are judging one sample at a time, using a category scaling method. Samples are random and independently selected from a database of 5000 images, containing both photos and graphics (the same database was used in Solli and Lenz[1]). Observers are asked to judge the selected samples on different emotion-related scales. For simplicity, each scale is divided into five equally sized intervals, or categories, and the observer can only pick one of them. The entire graphical user interface, including user instructions, can be seen in Fig. 3. The maximum sample size, height or width, is 200 pixels. Advantages with this method are that observer instructions are rather easy to understand, and observers can quit judging samples whenever they want. However, the simplicity will bring some negative aspects as well. Foremost, the observer criterion may drift (move along the emotion scale), especially for the first few images that are judged. The reason is that observers are not familiar with the full range of the scale until they have seen relatively many samples. Another drawback is that we only obtain a rough estimation of visual distance between samples since the interval spanned by each category is rather large. And finally, when judging samples, observers tend to use all categories an equal number of times (according to Engeldrum[20]), which may not resemble the reality since samples are displayed randomly.

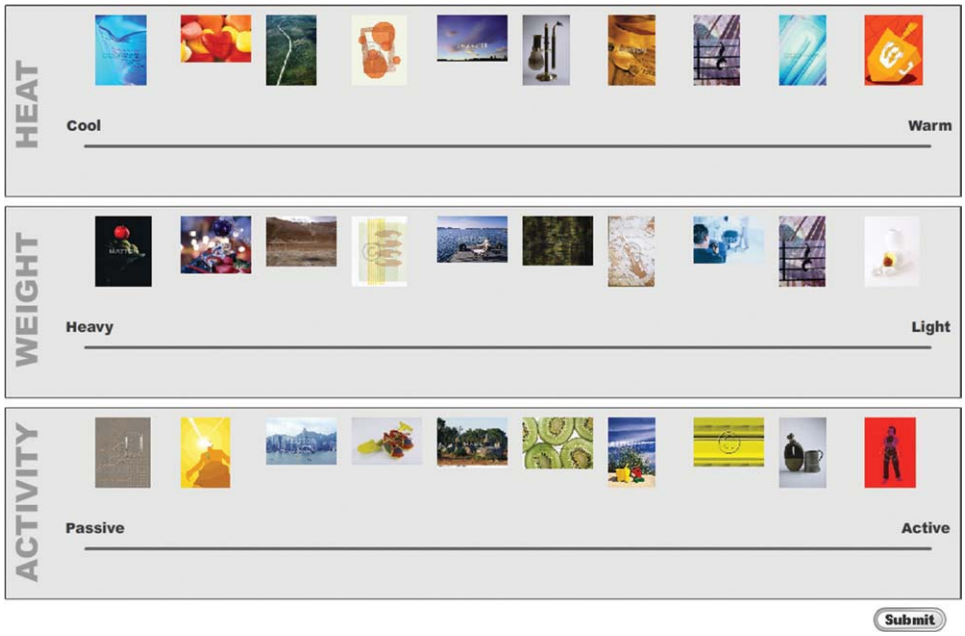

FIG. 4. The Graphical User Interface for the main study. Observers are asked to place each set of samples on the emotion ruler below. They are instructed that the distance between samples should correspond to the emotional difference.

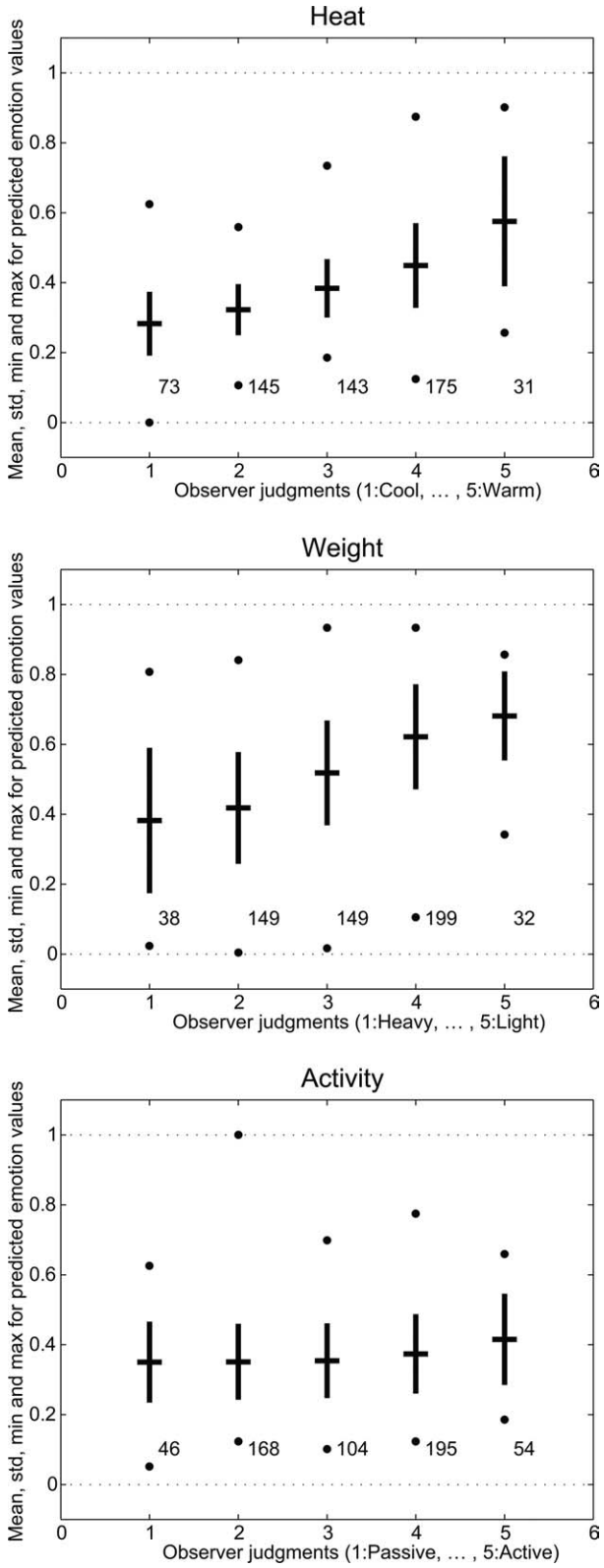

FIG. 5. Observer judgments for heat, weight, and activity from the pilot study are plotted against predicted emotion values. The *x*-axis represents judgment categories, and the *y*-axis represents predicted values. For each category, the short horizontal line corresponds to the mean of the derived values, the vertical line corresponds to the standard deviation, and dots correspond to the minimum and maximum among predicted values for samples in that category. The numbers along the dotted baseline shows how many samples that were assigned to each category.

The Main Study

An interval scaling method is utilized in this study, which, compared to the pilot study, will increase the accuracy of the emotional distance. Observers are shown a set of samples, and then asked to place all samples on a ruler ranging from one emotion attribute to another. Observers are also instructed to place the samples so that the distance between samples corresponds to the emotional difference. Samples with the same emotion response can be positioned above, or on top of each other. Totally 10 samples are used for each emotion scale. The selection of samples is based on judgments obtained in the pilot study. In the pilot study, each emotion scale was divided into five intervals or categories, and observers were assigning samples to those categories. From the set of samples assigned to each category, two samples are randomly drawn, giving us totally 10 samples for each emotion scale. Using this selection should provide a more homogeneous coverage of the emotion space than a random selection. Even if a full coverage is not necessary (observer judgments can still be evaluated for parts of the emotion space), it is of course desirable to be able to evaluate the entire emotion space within the current study. The selection procedure is an important and challenging task that is further discussed in "Conclusions" section.

The user interface for the main study, including selected samples, can be seen in Fig. 4. By coincidence, one sample was drawn for both heat and weight. Examining the samples visually, it looks like none of them contains a strong emotional content not related to color (the author's opinion), as discussed in previous section. Using the mouse pointer, observers can click on images and drag them to desired positions. Here the maximum sample size, height, or width, is 70 pixels. The reason for using smaller samples than in the pilot study is simply to make sure that all samples and the entire scale fits within the observer's browser window. This method demands from the observer that a complete set of samples should be judged, and the observer also needs to consider the emotional distance between samples.

An important consideration in the design of psychophysical experiments is the "rubber band effect." This means that most interval scales involve two arbitrary constants, a multiplier and some additive constant. To overcome this problem and enable easier comparisons between judgments from different observers, we follow Engeldrum[20] and calibrate observers to a common scale by adjusting judgments until they have the same mean and variance. For each judgment j, containing i samples, the mean is subtracted, and the result is divided by the standard deviation, as follows

$$a_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j} \tag{6}$$

where x_{ij} is the position of sample i in judgment j, $\bar{x}_j$ is the mean position for all samples in judgment j, s_j is the standard deviation for all samples in judgment j, and a_{ij} is

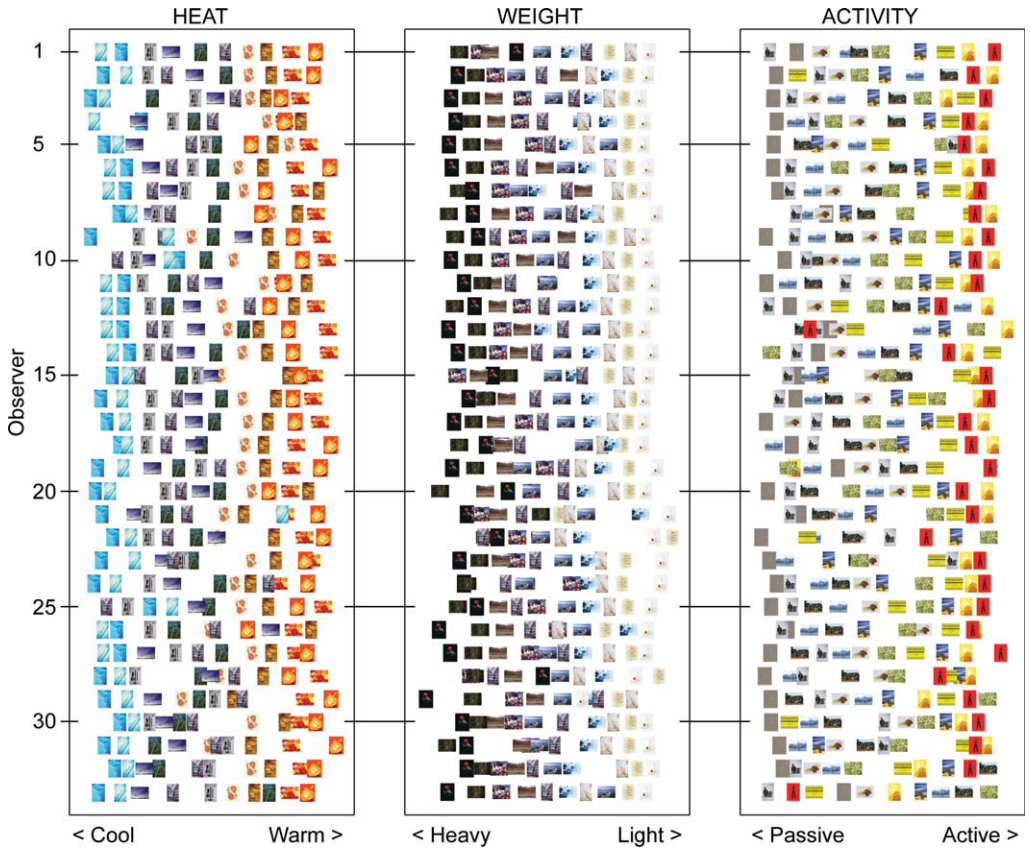

FIG. 6. Observer judgments for heat, weight, and activity. Judgments are scaled to have the same mean and variance.

the resulting position for sample i in judgment j. An alternative approach is to select a sample from each end point of an emotion factor (e.g., most cool and most warm) and use them as anchors. Then each judgment is normalized based on anchor positions. However, since two samples were randomly selected from each end point category, using only one of them as anchor will probably influence the result depending on which one we choose. But more important, at this stage we cannot know if the selected sample is a representative observer judgment suitable to be used as anchor. Moreover, judgments made in the pilot study cannot be fully trusted since the study was conducted on a public web page. Therefore the normalization by mean and variance is adopted.

RESULTS AND DISCUSSIONS

The Pilot Study

The web address of the page containing the pilot study was sent to a number of recipients, including colleagues, students and friends, both males and females ranging from ~20–65-years old. The address was also displayed on a public web page, probably generating a few more unknown participants. Totally 52 observers participated in the study, judging in total 567 samples (as described earlier, these are random and independently selected from a larger database containing 5000 images). The number of observers is based on a measurement of the number of unique computers used when judging images. Several observers may have used the same computer, or a single observer may have used more than one computer. However, we assume that the consequences on the overall result are of minor importance and can be ignored. In Fig. 5, observer judgments for heat, weight, and activity are plotted against emotion values derived using the predictive model described earlier and in Solli and Lenz.[1] For both heat and weight a clear relationship between judgments and derived values can be observed. For activity the relationship is much weaker. Only samples that are judged to have maximum activity are somewhat distinguishable from remaining ones. The figure also shows how many samples were assigned to each category. Apparently, "neutral" samples (close to the center of each scale) are more common than "extreme" samples (end points).

The Main Study

The address to the web page containing the main study was also sent to a number of recipients, but not displayed on a public web page. The group of recipients was comparable to the group in the pilot study: colleagues, students, and friends, both males and females ranging from ~20- to 65-years old. However, most of the recipients did not participate in the pilot study (although we cannot

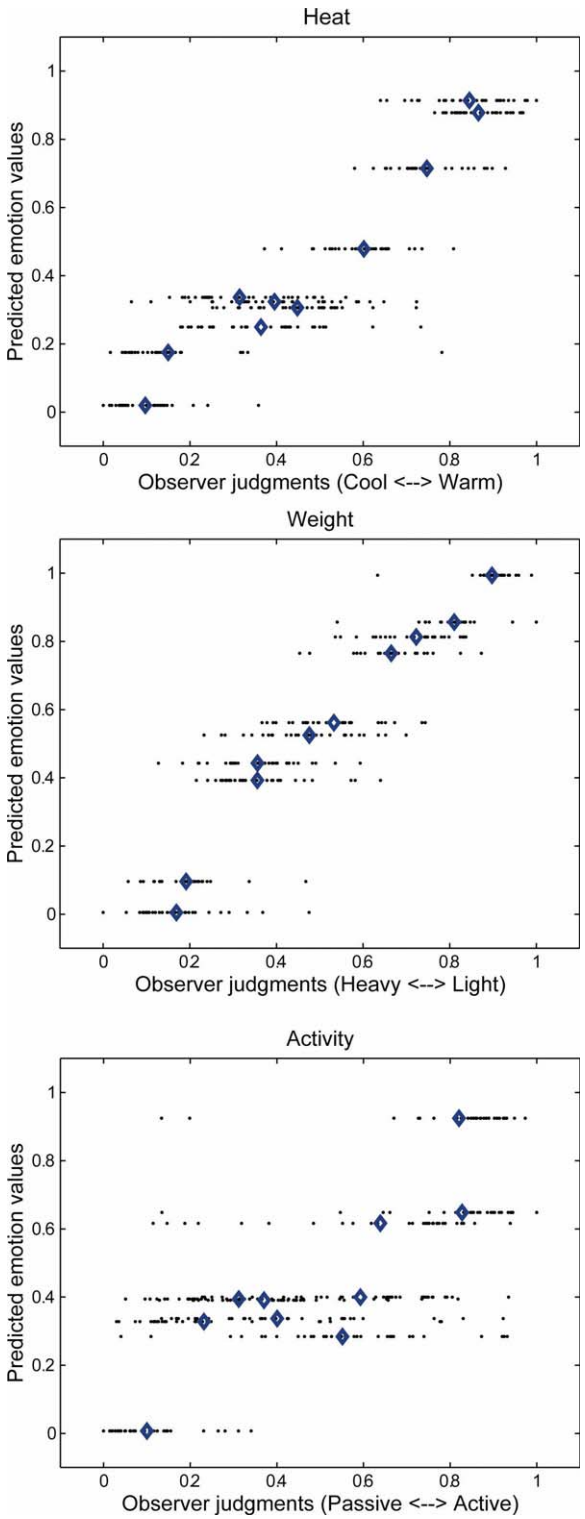

FIG. 7. Observer judgments for heat, weight, and activity from the main study are plotted against predicted emotion values. The *x*-axis represents positions on the emotion ruler, and the *y*-axis represents predicted values. Diamonds correspond to mean positions for each sample. For each diamond there is a distribution of dots (on the same horizontal level) showing individual judgments. For increased readability, the plot area is larger than the possible value interval ([0 1] for both the *x*- and *y*-axis). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

guarantee that they did not visit the public web page for the pilot study). As the presumed numbers of observers were lower than in the pilot study, participants were asked to judge samples only once to avoid single observers influencing the data too heavily. Totally 35 observers participated. The response from two observers were removed because they pushed the submit button before moving any of the images (we assume that was a mistake). This leaves 33 observer judgments for the evaluation process. In Fig. 6, all judgments are shown, for heat, weight, and activity, respectively. The mean and variance are used for scaling each judgment.

For measuring the interobserver agreement we adopt the method and terminology used by Ou and Luo.[12] The interobserver agreement measures the consistency of observer judgments, in other words how well observers agree with each other. The agreement can be derived by averaging the root mean square (RMS), also known as root mean square deviation (RMSD), or root mean square error (RMSE), between each observer's judgment and the overall mean. The RMS is derived by

$$\mathrm{RMS} = \sqrt{\frac{\sum_i (x_i - \bar{x}_i)^2}{N}} \tag{7}$$

where x_i represents the scale position given by an observer for sample i, $\bar{x}_i$ represents the mean scale position of all observers for sample i, and N represents the number of samples. However, what is not mentioned by Ou and Luo[12] is that if one wants to compare RMSs with different units (different scale, different amounts of categories, etc.) the RMS needs to be normalized, for instance by calculating the normalized root mean square

$$\mathrm{NRMS} = \frac{\mathrm{RMS}}{x_{\max} - x_{\min}} \tag{8}$$

where $x_{\max}$ and $x_{\min}$ are maximum and minimum of possible scale values. Interobserver agreements obtained are 0.099 for the heat factor, 0.079 for weight, and 0.160 for activity. As expected from earlier results, the agreement is better for heat and weight than for activity. Notice that a good agreement corresponds to a low value, and vice versa. The agreement values obtained are similar or even lower than results presented in related studies (for instance Ou and Luo[12]), indicating that the agreement is satisfactory. As each observer only judges one set of images we do not need to consider the intraobserver agreement.

TABLE I. Pearson r correlation coefficient and R^2 correlation between observer judgments and predicted emotion values.

	Pearson *r*		R^2	
Emotion factor	Mean positions	All positions	Mean positions	All positions
Heat	0.97	0.90	0.95	0.81
Weight	0.98	0.93	0.97	0.86
Activity	0.86	0.68	0.73	0.47

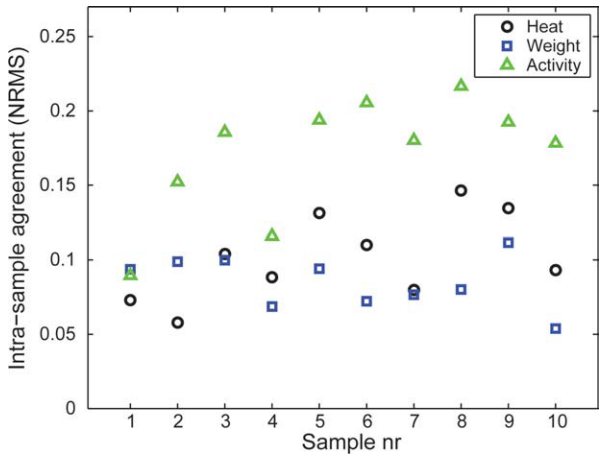

FIG. 8. Intrasample agreement for different samples and emotion factors. Notice that even if samples from different emotion factors have the same sample number, the image content is not the same. See all samples in Fig. 4 (sample number 1–10 from left to right).

Verification of the Predictive Model

The next step is to compare the observers' emotional response to the proposed predictive model (presented earlier and in Solli and Lenz[1]). In Fig. 7, judgments are plotted against emotion values derived using the predictive model. Diamonds corresponds to mean positions for each sample. For each diamond there is a distribution of dots (on the same horizontal level) showing individual judgments. The evaluation of the predictive model uses the Pearson product-moment correlation coefficient r between observer judgments and predicted values, for both mean positions (diamonds) and all positions (dots) in each emotion scale. The correlation coefficient is defined as

$$r = \frac{1}{N-1}\sum_{i=1}^{N}\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right) \tag{9}$$

where x_i and y_i are positions on the emotion scale, $\bar{x}_i$ and $\bar{y}_i$ are mean positions, and S_x and S_y are standard deviations of all positions for observer judgments and predicted values respectively. N represents the number of samples. The correlation coefficient measures the strength and direction of a linear relationship between sets of variables, this time judgments and predictions. For a perfect increasing (or decreasing) linear relationship the correlation is 1 (or -1). For independent variables the correlation is 0. Heat and weight resulted in correlation coefficients as high as 0.97 and 0.98. For activity the correlation is 0.86. The results obtained can also be found in Table I. The last column contains R^2 values (the square of the multiple correlation coefficient), which can easily be obtained from r since the correlation is attained from a simple linear regression ($R^2 = r \times r$). R^2 values indicates how much of the variation in the judgments that can be predicted by the model (e.g., $R^2 = 0.7$ means that the model explains 70% of the variance). Using the rather simple correlation coefficient r for measuring correlation can probably be questioned. However, since we want to measure the association between two variables, not the exact agreement, we search for an evaluation method invariant to scaling and translation of variables, and consequently we find the correlation coefficient suitable. In addition, considering previously published work within this research area we notice that using the correlation coefficient enables us to compare results with others. It is, however, not obvious how to draw conclusions from single correlation values. But by comparing our results with results reported in related studies (see Refs. 2,7,9,12 to mention a few) we conclude that the obtained correlation values lie within the range of what in general is considered to be a good correlation. Especially, the results for heat and weight show a very good correlation.

For comparison, in the pilot study the correlation coefficient was 0.58 for heat, 0.52 for weight, and 0.14 for activity. This confirms that the method of category scaling is much less reliable than interval scaling. In addition, we suspect that some of the data in the pilot study is unreliable since the user test was conducted on a public web page.

If we compare the interobserver agreement with the Pearson r correlation coefficient for heat, weight, and activity, we see that there is a relationship between poorer interobserver agreements and lower correlation coefficients. We may even suspect that some discrepancy in the correlation coefficient is due to observers varying opinions, and not due to limitations in the predictive model.

Intrasample Agreement (for Finding Problematic Samples)

Knowing samples with observer judgments that are highly inconsistent can be of importance if one wants to improve the predictive model. The normalized root mean square is used for calculating the intrasample agreement, a measure of how judgments concerning specific samples are spread along the emotion scale. Equations (7) and (8),

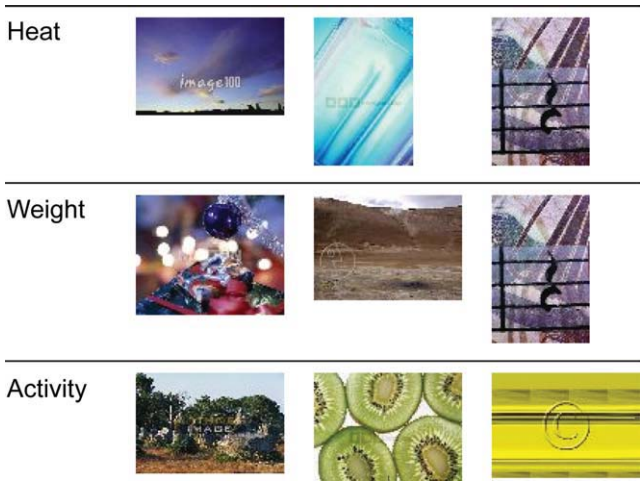

FIG. 9. The three samples from each emotion factor that obtained the worst intrasample agreements.

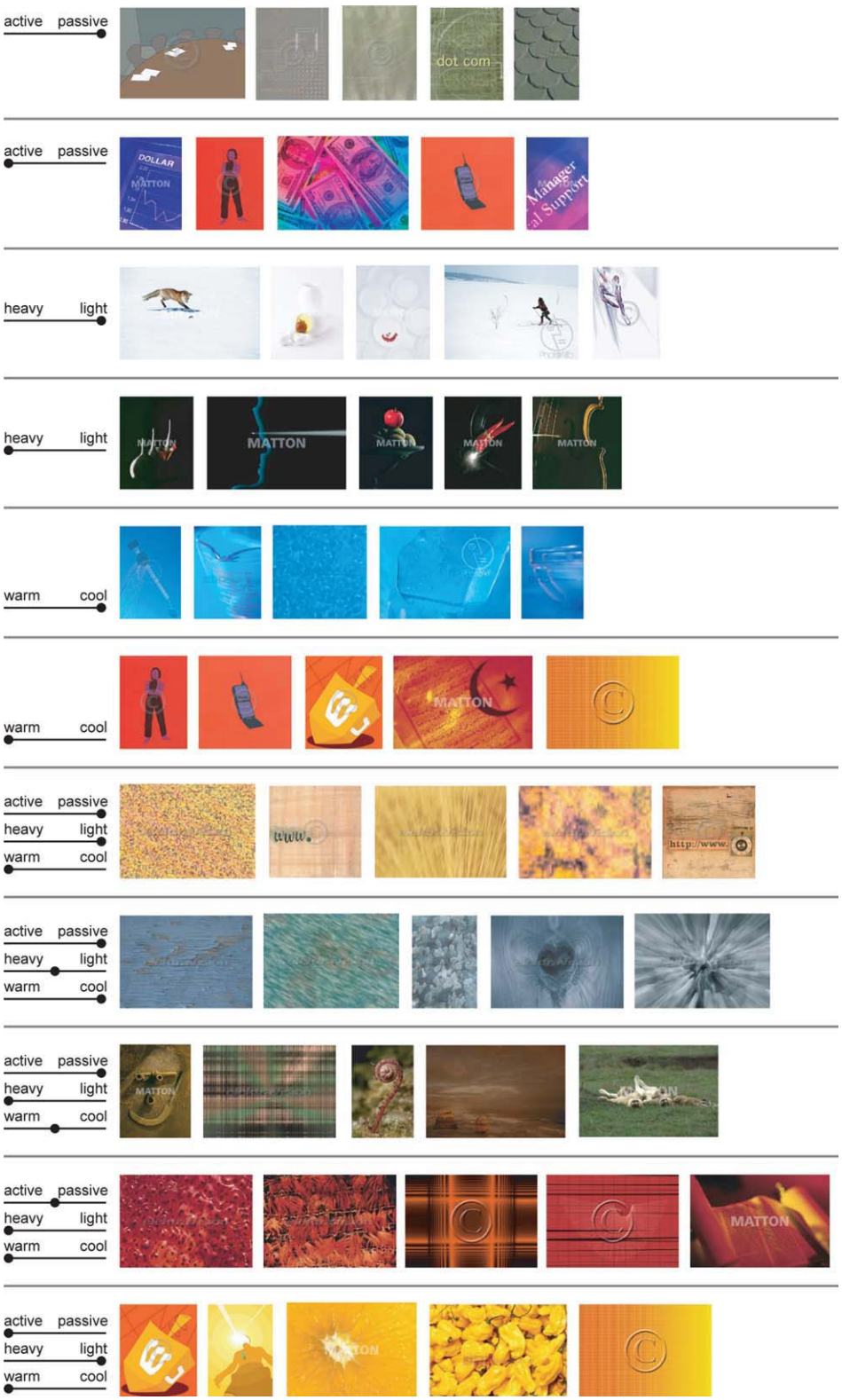

FIG. 10. Retrieval by emotion words. Sliders show the selection and values of the emotion words. The five best matches for each query are shown.

presented earlier, are used for calculating the intrasample agreement, but with i representing different judgments instead of samples. Consequently, N represents number of observers. Results obtained for samples 1–10 in all three emotion scales are plotted in Fig. 8. The mean intrasample agreements are 0.10 for heat, 0.08 for weight, and 0.17 for activity. The result corresponds well with values obtained for the interobserver agreement, reinforcing that

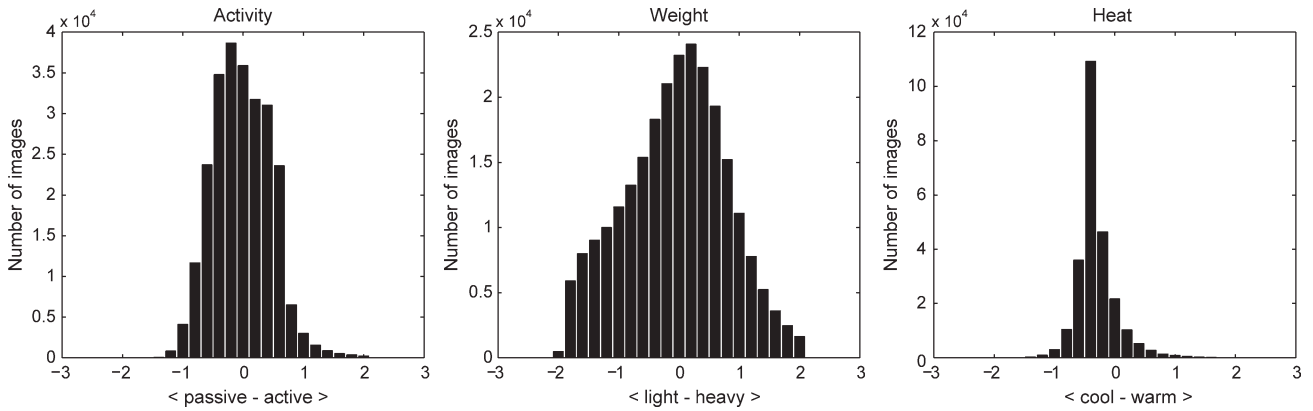

FIG. 11. The distribution of emotion values for a collection of 249,587 images downloaded from the Internet. Emotion values are not normalized, thereby the slightly different scaling of emotion factors compared to Figs. 1 and 7.

the activity factor is harder for observers to comprehend and apply on multicolored images than the heat and weight factors. On the other hand, the result may also show that observers have slightly varying opinions about activity for samples.

The three samples from each emotion factor that obtained the worst intrasample agreements are plotted in Fig. 9. We can observe a mixture of both highly abstract graphical images and ordinary photos. Samples with poor intrasample agreement cannot be visually distinguished from the remaining ones (all samples can be seen in Fig. 4). However, more research is needed to find out if those samples can be distinguished by some currently unknown statistical or computational properties.

APPLICATION IN IMAGE RETRIEVAL

We summarize selected parts from Solli and Lenz[1] to give the reader an overview of the intended application. Our first application is called retrieval by emotion words. When searching for an image, a search interface is used, where the query vector is specified with the help of 1–3 sliders, corresponding to different emotions. For a given query vector of emotion scores we retrieve the images whose emotion vectors are nearest neighbors of the query vector in the L2 norm. A few search examples, using a database containing 5000 images, both photos and graphics, are illustrated in Fig. 10. The image database is a randomly selected subset of a much larger database used in previous research.

An additional search mode called retrieval by query image is presented in Solli and Lenz.[1] Here the user can submit a query image to obtain images from the database with similar emotional appearance. Figure 1 demonstrates that the RGB space is unevenly spaced in emotion space. We use a kd-tree decomposition to obtain a more balanced decomposition of emotion space computed from the RGB histogram. For this, we split the emotion space perpendicular to one of the axes (heat, weight, or activity). Cycling through the axes, splitting each new region by the median, the result is a distribution with equal number of entries in each region, generally called a balanced tree with equal number of entries in each leaf. We split until we get 64 leaves, each containing eight entries. This is the color emotion histogram with 64 bins. Knowing which color emotion value belongs to which leaf, we can create a matrix that will transform RGB-histograms to color emotion histograms. The L2 norm is used for measuring distances between color emotion histograms. Compared to retrieval by emotion words, the use of emotion histograms enables a richer description of each image. However, the method is only applicable when a query image is available. A more extensive description of this search mode together with retrieval results can be found in Solli and Lenz.[1] Both retrieval methods are implemented in a public demo search engine (http://media-vibrance.itn.liu.se/imemo/). For an easier understanding of the search modes, we encourage readers to interact with the search engine.

CONCLUSIONS

The findings of this study show that people in general perceive color emotions for multicolored images in similar ways, and that observer judgments highly correlate with the predictive model recently proposed in image retrieval. Images were judged by observers on three emotion factors: heat, weight, and activity. Interobserver agreement obtained for the heat factor are 0.099, 0.079 for weight, and 0.160 for activity. Comparing the results with other studies involving single colors or two-color combinations shows that the agreement is good, especially for heat and weight. For measuring the correlation between observer judgments and predicted values the Pearson product-moment correlation coefficient r was used. Heat and weight obtained correlation coefficients of 0.97 and 0.98. For activity, the correlation is down to 0.86, but still within the range of what in general is considered a good correlation. A poorer interobserver agreement for the activity factor has a negative influence on the correlation. Also the intrasample agreement for heat and weight is better than for activity, indicating both that

the activity factor is harder for observers to comprehend, and that observers may have slightly more varying opinions about the activity factor than the other two factors. No visual difference could be established between images with high and low intrasample agreement.

The selection of samples in the main study is an important and challenging question. As this is the first study within this subject, it is difficult to decide how to select an ideal set of samples. One attempt could be to create a set that is representative for the database used in the target application. But the subsequent question is then: Representative in what way? The selection can be based on many different measurements, like emotional properties, the color content, etc. As shown in Fig. 5, the results obtained in the pilot study indicate that emotionally "neutral" samples (located in the middle of the emotion scale) are more common than "extreme" samples (located at end points). That this is a realistic assumption is confirmed in the following preliminary experiment where we used 249,587 sample images from the image search engine Picsearch (www.picsearch.com). For these images we computed emotion values, and histograms describing the distributions of emotion values. The histograms are plotted in Fig. 11. Also for this database, collected from the Internet, it is obvious that "neutral" samples are more common than "extreme" samples. A set of samples emotionally representative for the entire target database would thus contain more "neutral" samples than "extreme" samples. However, if the findings are implemented in an image retrieval system, we expect that users utilizing the emotion search are mainly interested in samples with strong emotional content. Consequently, it might be beneficial to develop and evaluate the method with a sample set containing equal number of "neutral" and "extreme" samples, or even more "extreme" samples than "neutral" ones. With this in mind we conclude that the sample selection method used in this initial study was appropriate, and that the obtained knowledge can be used for refining the selection method in upcoming research.

FUTURE WORK

In present research concerning image retrieval, focus often lies on scene and object recognition. However, we predict that one of the upcoming great challenges in image indexing and Content Based Image Retrieval will be the understanding of emotion-related attributes, like "warm," "cold," "happy," "harmony," etc. The present study is the first step toward a broader use of emotion related properties in Content Based Image Retrieval. Upcoming research will focus on the inclusion of additional emotion properties, and how to efficiently communicate those properties in a user interface for image indexing and retrieval. For the present study in particular, one thing to consider is whether the activity factor should be altered to some factor that users more easily comprehend, resulting in better interobserver agreement.

1. Solli M, Lenz R. Color Emotions for image classification and retrieval. Proceedings IS&Ts 4th European Conference on Colour in Graphics, Imaging, and Vision. Springfield, VA: IS&T; 2008. p 367–371.
2. Ou LC, Luo MR, Woodcock A, Wright A. A study of colour emotion and colour preference. Part I: Colour emotions for single colours. Color Res Appl 2004;29:232–240.
3. Ou LC, Luo MR, Woodcock A, Wright A. A study of colour emotion and colour preference. Part II: Colour emotions for two-colour combinations. Color Res Appl 2004;29:292–298.
4. Ou LC, Luo MR, Woodcock A, Wright A. A study of colour emotion and colour preference. Part III: Colour preference modeling. Color Res Appl 2004;29:381–389.
5. Kobayashi S. The aim and method of the color image scale. Color Res Appl 1981;6:93–107.
6. Sato T, Kajiwara K, Hoshino H, Nakamura T. Quantitative evaluation and categorising of human emotion induced by colour. Adv Col Sci Technol 2000;3:53–59.
7. Gao XP, Xin JH. Investigation of human's emotional responses on colors. Color Res Appl 2006;31:411–417.
8. Gao XP, Xin JH, Sato T, Hansuebsai A, Scalzo M, Kajiwara K, Guan S, Valldeperas J, Lis José M, Billger M. Analysis of cross-cultural color emotion. Color Res Appl 2007;32:223–229.
9. Xin JH, Cheng KM, Taylor G, Sato T, Hansuebsai A. Cross-regional comparison of colour emotions Part I: Quantitative analysis. Color Res Appl 2004;29:451–457.
10. Xin JH, Cheng KM, Taylor G, Sato T, Hansuebsai A. Cross-regional comparison of colour emotions. Part II: Qualitative analysis. Color Res Appl 2004;29:458–466.
11. Beke L, Kutas G, Kwak Y, Sung GY, Park DS, Bodrogi P. Color preference of aged observers compared to young observers. Color Res Appl 2008;33:381–394.
12. Ou LC, Luo MR. A colour harmony model for two-colour combinations. Color Res Appl 2006;31:191–204.
13. Wang WN, Yu YL. Image emotional semantic query based on color semantic description. Proceedings of International Conference on Machine Learning and Cybernetics. Guangzhou, China, 2005. p 4571–4576.
14. Corridoni JM, Del Bimbo A, Pala P. Image retrieval by color semantics. Multimedia Syst 1999;7:175–183.
15. Hong SY, Choi HY. Color image semantic information retrieval system using human sensation and emotion. Issues Inf Syst IACIS 2006;VII:140–145.
16. Cho SB, Lee JY. A human-oriented image retrieval system using interactive genetic algorithm. IEEE Trans Syst Man Cybernet 2002; 32:452–458.
17. Yoo HW. Visual-based emotional descriptor and feedback mechanism for image retrieval. J Inf Sci Eng 2006;22:1205–1227.
18. Wang WN, Yu YL, Jiang SM. Image retrieval by emotional semantics: A study of emotional space and feature extraction. Proc IEEE Int Conf Syst Man Cybernet 2006;4:3534–3539.
19. Lee J, Cheon YM, Kim SY, Park EJ. Emotional evaluation of color patterns based on rough sets. Proceedings 3rd International Conference on Natural Computation, ICNC, 2007, Vol. 1. Haikou, Hainan, China. p 140–144.
20. Engeldrum PG. Psychometric Scaling: A Toolkit for Imaging Systems Development. Winchester, Massachusetts, USA: Imcotek Press; 2000.