Content-based Mood Classification for Photos and Music

A generic multi-modal classification framework and evaluation approach

Peter Dunker
dkr@idmt.fraunhofer.de

Stefanie Nowak
nwk@idmt.fraunhofer.de

André Begau
bga@idmt.fraunhofer.de

Cornelia Lanz
cornelia.lanz@stud.tu-ilmenau.de

Fraunhofer Institute for Digital Media Technology
Ehrenbergstraße 31, Ilmenau, Germany

ABSTRACT

Mood or emotion information are often used search terms or navigation properties within multimedia archives, retrieval systems or multimedia players. Most of these applications engage end-users or experts to tag multimedia objects with mood annotations. Within the scientific community different approaches for content-based music, photo or multi-modal mood classification can be found with a wide range of used mood definitions or models and completely different test suites. The purpose of this paper is to review common mood models in order to assess their flexibility, to present a generic multi-modal mood classification framework which uses various audio-visual features and multiple classifiers and to present a novel music and photo mood classification reference set for evaluation. The classification framework is the basis for different applications e.g. automatic media tagging or music slideshow players. The novel reference set can be used for comparison of different algorithms from various research groups. Finally, the results of the introduced framework are presented, discussed and conclusions for future steps are drawn.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Context Analysis and Indexing, Abstracting methods, Indexing methods; I.5.2 [**Pattern Recognition**]: Design Methodology-Classifier design and evaluation

General Terms

Algorithm, Design, Experimentation, Measurement

Keywords

multi-modal mood classification, image content analysis, music information retrieval, evaluation

MIR'08, October 30–31, 2008, Vancouver, British Columbia, Canada.

1. INTRODUCTION

Digital music and photos get more and more involved in everyones life based on the growing number of digital cameras and the broad use of mobile digital music players. Within the scientific community and in first commercial products, content-based image and audio classification has become a wide-spread technique for a wide range of applications e.g. in music genre detection, speech/music analysis, spoken language analysis [1] and in the visual domain major in the determination of scene classes like landscape/city/party or daytimes day/sunset/night [2]. Especially mood tagging as search and exploration paradigm for the access of music databases like moodlogic, lastfm, magnatunes as well as for the access of image databases like flickr or gettyimages is quite common.

Typically music oriented publications concentrate on the term *mood*, while publications in the image domain are using the term *emotion*. Usually mood describes a longer human feeling e.g. hearing a complete song. The term emotion is often used in image oriented publications and describes in principle the personal affectedness based on spontaneous perception e.g. appearing images [3]. In this publication we use the term mood as a general term for mood and emotion. In principle this publication concentrates on different main use cases: automatic multi-modal media tagging for photo and music search and retrieval, navigation and visualization of media archives and finally multimedia slideshow players to enrich photo presentations by accompanying music as well as to enrich music playlists by accompanying photos.

Following the defined use cases the purpose of this paper is to find a mood model that fits all use cases, to develop a generic framework for photo and music classification and to set up a novel database for a comparable evaluation.

Therefore a review of common mood models as well as related work in this field is given in section 2. Based on a selected two dimensional mood model, a generic classification framework is presented in section 3 that considers an universal classification of images and music within the same framework. A reference database with publicly available images and music is created. In section 4.2 the selection process and details on the selected media are presented. In section 4.3 the results of the presented classification framework using the novel reference database are presented and discussed. Finally, in section 5 conclusions are drawn and possible future work is depicted.

2. RELATED WORK

Within this section we give a review and discussion of the related work on single-modal and multi-modal mood classifications as well as the used mood models and test suites. A short summary describes the drawbacks of the referred approaches and leads to the advantages of the presented work.

2.1 Mood Concepts

In order to get a well working multi-modal mood classification that meets the demands of its users, it is important to find a suitable mood model. Looking at the research that has been reported in this field, it becomes obvious that there are two main groups of mood models. The first group contains models that consist of *listings of adjectives or nouns* and the second group contains *dimensional models*.
A common example for the first group is Kate Hevner's Adjective Circle [4] depicted in Fig. 1. It consists of 66 single

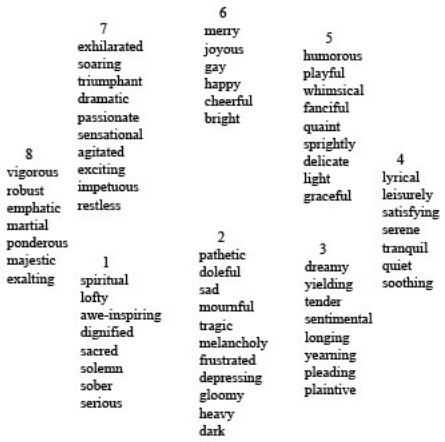

Figure 1: Kate Hevners Adjective Circle [4]

adjectives that are subdivided into eight groups. Chen et al. [5] chose emotional states based on Hevner's circle for their emotion-based music visualization using photos. The eight classes in the order of the numbers in Fig. 1 are called: sublime, sad, touching, easy, light, happy, exciting and grand. Farnsworth modified Hevner's concept and arranged the adjectives in ten groups [6].
Dimensional mood models consist of one or more dimensions where each represents a special mood characteristic. A very early approach has been presented by Wundt in 1896[7]. To some extent newer models are based on this concept. Wundt's mood space consists of the following three axes: pleasure/ displeasure, arousal/ nonarousal and stress/ relaxation as depicted in Fig. 2(A). Another three-dimensional approach is known as Albert Mehrhabians PAD. The dimensions are: pleasure/ displeasure, arousal/ nonarousal and dominance/ submissiveness [8].
Famous two-dimensional models are:
Thayer's model: One axis visualizes the amount of stress, the other the amount of energy[9], see Fig. 2(B).
Russell's model: It uses pleasant/ unpleasant for one dimension and a composition of stress and alertness for the second dimension. As a result the second dimension is spanned from sleepy to aroused.
Tellegen-Watson-Clark-Model (TWC): The TWC model is quite similar to the other models but contains a 45° turn. Its dimensions are called high positive affect/ low positive affect and high negative affect/ low negative affect [10].
Reisenzeins model: Reisenzein possesses the pleasant/ unpleasant and arousal/ non-arousal dimensions of Wundt's model [11].
While summarizing the review of the different mood models we found the following advantages and disadvantages for the suggested approaches. The category-based models can be easily used for tagging especially with a list of different adjectives for the same mood which generalizes the subjective perceptions of multiple users and gives a heavyset dictionary for search and retrieval applications. The dimension-based models also allow a percentage based assignment to a special mood dimension which improves the applicability for navigation and visualization use cases. A combination of both approaches points to an appropriate consensus.

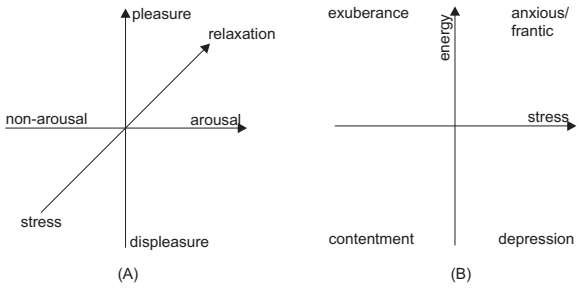

Figure 2: Mood model of Wundt(A) and Thayer(B)

2.2 Classification Approaches

Within the scientific community different approaches for single-modal and multi-modal mood classification were published. Here a rough overview of selected publications is given.

Music

Automatic mood classification for music is a comparatively common technique. The used musical attributes are typically divided into two groups, timbre-based attributes and rhythmic or tempo-based attributes. The tempo-based attributes can be represented by e.g. an Average Silence Ratio [12] or a Beats Per Minute value [13]. Lu [14] uses amongst others Rhythm Strength, Average Correlation Peak, Average Tempo and Average Onset Frequency to represent rhythmic attributes. Frequency spectrum based features like Mel-Frequency Cepstral Coefficients (MFCC), Spectral Centroid, Spectral Flux or Spectral Rolloff are also often used e.g. in [15, 16]. Wu and Jeng [17] use a complex mixture of various features: Rhythmic Content, Pitch Content, Power Spectrum Centroid, Inter-channel Cross Correlation, Tonality, Spectral Contrast and Daubechies Wavelet Coefficient Histograms. For the classification step in the music domain Support Vector Machines (SVM) [15, 18, 19, 13] and Gaussian Mixture Models (GMM) [14, 20] are typically applied. Liu et al. [21] utilize a nearest-mean classifier.
The comparison of classification results of different algorithms is difficult because every publication uses an individual test set or ground-truth. E.g. the algorithm of Wu and Jeng [17] reaches an average classification rate of 74,35% for 8 different moods with the additional difficulty that the results of the system and the ground-truth contain mood histograms which are compared by a quadratic-cross-similarity.

Image

Automatic mood classification for images is not that popular as for music. All approaches use a variant of color and gradient information for the image mood classification. Wang et al. [22] use image brightness, color temperature (warm-cool), saturation and contrast descriptions as features. Chen et al. [5] use a non-uniform quantized histogram of the HSV color space which is more suitable for human perception. Yoo [23] uses the non typical color code space which is based on Seons psychological evaluation of color patterns [24]. The gradient information is estimated by different features e.g. a Haar Wavelet Transformation [25], the Sum Of Gradients for the sharpness [22], a Hough Transformation [26] or Canny Edge Detectors together with Wavelet Coefficients [5]. Additionally, Yoo uses a granularity of homogeneous regions as Texture description.

The classification within the visual domain differs from the audio domain. Usually, audio features are extracted from a short audio frame, so a set of 100 training samples can be generated e.g. by 3 seconds of music. Image features are most often extracted as a single feature per image whereby 100 images result in 100 training samples. Therefore the classifiers are optimized for small sample size problem or using a preprocessing for a feature reduction or a sample number increasing. Guo [26] uses a combination of a fuzzy neural network to estimate a sequence of emotion semantemes and afterwards a double hidden Markov model to classify the mood. Wang [22] uses a Support Vector Machine of Regression which has the capability to generalize a small sample set. Chen et al. [5] use a Bayesian classifier and an Affinity Propagation Algorithm to label additional unknown images to generate more trainings samples for a following SVM. Wang [22] reports an average recognition rate of 86%, with the speciality that 12 mood pairs are classified individually. Therefore a random test would reach 50%.

Multi-modal

Automatic multi-modal mood classification describes a combined or parallel analysis of music and images e.g.[25, 5]. In addition to the described techniques the specialty of Cho's approach [25] is the interactive Genetic Algorithm that allows user interaction and continuous learning and optimization of classification parameters. This approach was designed as a human computer interface for music and image retrieval with user interaction. This procedure is not applicable for the use cases of this publication.

Chen et al. [5] present an approach for music and image mood classification with the application to generate an accompanying visualization for music. They adapt the 12 mood categories of Wu and create a ground-truth image set for training with a varying number of images per mood e.g. sad (11), easy (125), happy (62) or exciting (25). By using a Bayesian classifier and unlabeled images the number of training samples was increased with the compromise that this classifier has an accuracy of only 47% so that every second additional trainings item is wrong. Supposably reasoned by that restriction the final evaluation was performed as a user test to measure the user perception of coordination (connection music/image), interestingness (presence style), colorfulness (enrichment of audio by images) in comparison to a Microsoft Media Player visualization and a random photo slideshow. This user test reveals an increased user experience of the mood based combination of music and photos which is an important finding. Nevertheless it lacks an automatic and reproducable quantitative evaluation for comparison of ongoing development in other research groups.

2.3 Ground-Truth

One key challenge in the evaluation of mood classification is to have a ground-truth that can be used by different research groups to compare their results. The difficulty of annotating mood lies in its subjectiveness in contrast to technologies like face detection. Within the related work described in section 2.2 different media sources are mentioned e.g. private CD collections. Besides [27] who uses the USPOP CD collection no reproducible music data is referred to. For the image data usually downloads from the Internet are taken but without a traceable reference. Besides the media sources, also the media types or genre are quite different e.g. Classic[20, 19], Jazz [15] or film sound tracks without singing voice [17]. Used types of images are paintings [28] or photos of the daily life [5].

For generating the ground-truth by labeling the media data with mood tags usually a small number of persons is involved. This can be students [29] or experts [20, 30] and sometimes people with different nationalities [16]. Chen et al. [5] are using an online tagging system where 496 persons labeled 398 images.

The commonality of all reviewed publications is the lack of a golden standard for the ground-truth that allows a comparison of individual algorithms. Furthermore, a joint mood concept for photos and music is missing which can be used for tagging both, music and photos, and which allows all kinds of the above described use cases. Especially a mood model and an associated multi-modal classification approach for visualizing music and image items in the same user interface is missing.

3. MOOD CLASSIFICATION

Within this section the generic multi-modal mood classification framework is presented beginning with the selection of an appropriate mood model.

3.1 Mood Model Selection

The mood concepts introduced in section 2.1 provide a basis for the decision about which model is the best for our multi-modal mood classification approach. Due to the fact that there are numerous list-models with different numbers of items and diverse moods it is difficult to determine which concept refers to a complete and correct mood model. A dimensional approach is not automatically complete but it is complete and logical inside its area or space. Based on the mood area which is spanned by the mood dimensions the number of individual mood categories can be decreased or increased e.g. by choosing one or four moods per quadrant. A one-dimensional model would not be sufficient because e.g. the character of a high aroused state can vary from positive arousal (joy) to negative arousal (anger). Finally, the two-dimensional Reisenzein model has been chosen, because valence and arousal "'account for most of the independent variance in affective responses" [31]. Three-dimensional models increase the complexity without need and would expand a possible navigation user interface from a common 2D to a non-intuitive 3D interface.

Also, numerous studies have shown that emotions selected by audio or visual media can be mapped onto an emotion space with the dimensions arousal and valence [31]. Another

point that confirms the choice of this model is the fact that attributes of activation/ deactivation and positive/ negative quality are expressed by features in pictures or music.

Mood expression in Images and Music

The aspect of arousal/ non-arousal in a picture can be visualized by color, saturation, lightness and the orientation and character of lines. So called warm colors like red and orange are more active than passive colors like blue and turquoise [32]. Activity emerges from diagonal lines, whereas horizontal lines do express calmness [33]. Pleasure/ displeasure is expressed by the lightness of the color. Bright colors create a positive and friendly mood whereas dark colors create a gloomy impression [32].

High activation in music can be generated by fast tempo, frequently variations of tempo, high dynamic changes, high pitch range and bright timbre. Pleasure in music is created by fast tempo [34].

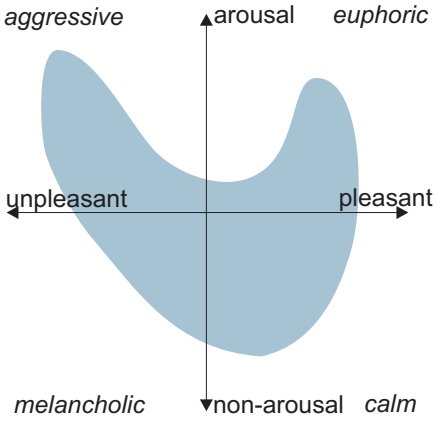

Figure 3: Reisenzein or valence/arousal model with labeled affect area (derived from [31])

Mood Dimensions and Categories

Physiological experiments have shown that only a part of the area of the two-dimensional valence-arousal model equates a human emotion [31] as depicted in Fig. 3. This raises the following questions: How many moods per quadrant should be chosen? Should the number of moods per quadrant depend on the size of the "affect space"? From which point of the area should the moods be picked? Because the answers to these questions are quite complex a simplified approach is to pick one mood per quadrant. Then the moods are named by the extreme values of one quadrant. This should assure that the single items are clearly distinguishable from each other. The following four discrete moods (one per quadrant) have been chosen: aggressive, melancholic, euphoric and calm (see Fig. 3).

The selection of the two dimensional mood model and its four categories can easily be mapped into each other and therefore keeps the feasibility for all defined use cases. Within this publication we concentrate on the four categories of the mood model.

3.2 Audio and Visual Features

In the field of multimedia information retrieval, features are measurable properties representing the media objects they are computed from. These descriptors have to meet the requirements to be relevant with respect to human perception while being robust and computable with reasonable effort.

Audio

Audio low-level features like the MPEG-7 Audio Spectrum Flatness descriptor [35], being directly extracted on frequency bins, are the basis for state of the art robust audio identification applications [36]. Since these low-level approaches lack semantic information, mid-level features are able to provide higher cognitive properties such as tempo, tonality, syncopation or melody grouping. Combining both kinds of features is successfully adopted in the fields of music similarity computation [37]. The presented strategy is pursued in this work for classifying the mood of a piece of music as a high-level semantic description. To represent the sound based musical attributes, the low-level features Normalized Loudness, Audio Spectrum Flatness, Spectral Centroid, Spectral Crest Factor and MFCC are combined with 3 MFCC based timbral mid-level features. Four rhythmic mid-level features based on Audio Spectrum Envelope represent the tempo based musical attributes. Altogether a set of five low- and seven mid-level features is used resulting in a total feature vector dimensionality of 219. A feature vector is extracted every 2.56 seconds representing an audio snippet of 5.12 seconds length to ensure a high temporal resolution.

Visual

The visual descriptors mainly consist of low-level features that cover the color and structure domain similar to algorithms reviewed in 2.2. Due to peculiarities of the classification approach the overall feature vector dimension for images should be less then the number of samples available for training. Therefore, we developed a Color Histogram feature which is calculated in the Hue Saturation Value (HSV) color space with an individual quantization of each channel to H-8, S-4 and H-4 bins, similar to the MPEG-7 Scalable Color Descriptor[35]. This feature covers the properties: brightness/darkness, saturation/pastel/pallid and the color tone/hue. To describe the structure or horizontal/vertical frequencies we developed a Haar Wavelet feature which describes the mean and variance of the energy for each band. Applying three wavelet decompositions for three orientations the feature vector dimension is 18. As an additional color oriented feature we developed a Color Temperature Histogram which is based on a first k-means clustering of all image pixels in the LUV color space. Afterwards, the color temperature of each centroid is calculated and a histogram with 8 color temperature bins in the range from 1.500 to 20.000 Kelvin is setup. This feature describes the warm/cool impact of images. The concatenated image feature vector has a dimensionality of 42.

Dimension Reduction

A widely used method to improve discriminability among classes while reducing the feature dimension is the Linear Discriminant Analysis (LDA) [38]. This linear transformation maximizes the ratio of between-class variance to the within-class variance thereby guaranteeing maximal separability. The resultant $N \times N$ matrix $\mathbf{T}$ is used to map a N-dimensional feature row vector $\mathbf{x}$ into the subspace $\mathbf{y}$ by a multiplication. Reducing the dimension of the transformed feature vector $\mathbf{y}$ from N to D is achieved by considering only the first D column vectors of $\mathbf{T}$ (now $N \times D$) for multiplication. Within our classification framework the number of D will be optimized for the special classification tasks.

3.3 Classification Models

There are two general classification approaches, a generative and a discriminative one. Both allow to classify unknown multimedia objects into different classes with a certain probability depending on the training of the model and the extracted features from the data.
Generative probabilistic models describe how likely a multimedia object belongs to a certain class of multimedia objects. These models form a probability distribution over the object's features, in this case over the audio and image features presented in section 3.2, for each class. In contrast, discriminative models try to predict the most likely class directly instead of modeling the class conditional probability densities. Therefore the model learns class boundaries between different classes during the training process and uses the distance to the boundaries as indicator which class is the most probable for the given data.
In the research community, there are different opinions about what approach to pursue. Supporters of the discriminative approach argue that "one should solve the problem directly and never solve a more general problem as an intermediate step."[39]. In favour for the generative approach is the fact that discriminative models do not take prior information into account and that all classes have to be considered simultaneously.[40]
For the proposed multi-modal mood classification, we investigate and evaluate both approaches. As a generative probabilistic model, a Gaussian Mixture Model is used. The discriminative classifier is a Support Vector Machine.

Gaussian Mixture Model (GMM)
For classifying a multimedia object by a GMM, one assumes that the single objects are generated by a mixture of Gaussian sources. By estimating the model parameters of the GMM, information about which mixture models which mood class can be obtained and can be used to separate the multimedia objects. The Gaussian distribution represents a set of independent data samples by its mean μ and variance σ^2 using the Gaussian probability density function. It is very useful to describe datasets with unimodal densities, but fitting a Gaussian to multi-modal datasets gives a mean value in an area with low probability and an overestimated variance. The idea of mixture models is to use a mixture of Gaussians, realized by linear superposition of Gaussian distributions. Data samples are thought of as originated from various sources and each source is modeled by a single Gaussian. Therefore, mixing coefficients $P(c)$ are introduced, that could be understood as prior probabilities, in which every source is present. Regarding a mixture of M Gaussians, the finite mixture density $p(x)$ is described as:

$$p(x) = \sum_{i=1}^{M} P(c_i) \frac{1}{\sigma_i \sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x-\mu_i)^2}{(\sigma_i)^2}} \quad (1)$$

The number of Gaussians M can be optimized within the proposed framework.

Support Vector Machine (SVM)
A SVM attempts to generate an optimal decision boundary (margin) between classes based on a set of labeled training feature vectors. The hyperplane that separates the classes is optimized in the sense that a maximum margin between the plane and the classes is achieved. In this work the publicly available library *LIBSVM* [41] is used which allows nonlinear hyperplanes for class separation. We utilize the Radial Basis Functions (RBF) kernel:

$$K(\mathbf{x}, \mathbf{y}) = e^{-\gamma \|\mathbf{x}-\mathbf{y}\|^2} \quad (2)$$

The LIBSVM allows an optimization of the kernel parameter γ and the penalty parameter C_{SVM} during training.

3.4 System Architecture

This section gives an overview of the overall system architecture used for the classification of audio and image media objects (Fig. 4). Due to the multi-modal approach, the main design requirements are universality and flexibility concerning the independence of media types and the selection of features and classification models. Both, the training

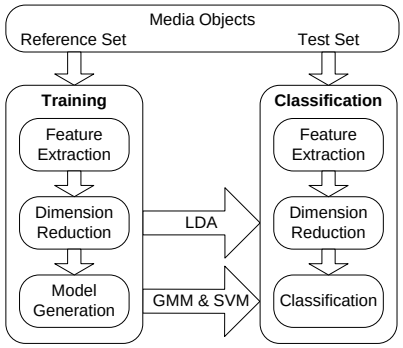

Figure 4: System block diagram

and the classification step, are divisible into 3 independent parts: feature extraction, dimension reduction and training/classification. The feature extraction is sufficiently described in the sections above. Additionally, an optimization step extends the training process.

Training
The feature extraction is done by a modular signal processing engine that is configurable for any type of media. The output is a binary file containing all extracted features for each media object. During the training step, the feature files for each class/mood are read by the system to perform the LDA. The resulting transformation matrix is saved for the use in the classification step while the transformed features are processed by the different training modules to generate the GMM and SVM models that are also saved for the classification step.

Classification
When defining the mood for an unclassified multimedia object of interest, the appropriate features are extracted first. After transforming the features into the LDA space, the classification with the GMM and SVM models is performed with resulting probability values for each mood. The classifier model probability values are normalized to a sum of 100%. The mood with the highest probability is voted as classified concerning its normalized probability as confidence C. In our experiments we use a confidence threshold value C_{thr} for rejecting classifications.

Optimization
Since the LDA and the different classification models come along with a set of various parameters, the system is designed to find an optimum parameter set in a 5-fold cross-validation routine. Within this step, the main training set is additionally divided into reference and test set according

to a Monte Carlo method while knowing to which mood a certain media item belongs. The test set is then classified against the models generated from the reference set. In an iteration loop, the LDA and classifier model parameters are now altered within predefined limits to repeat the Monte Carlo method in order find the optimum configuration.

approach	D	M	γ	C_{SVM}
Image GMM	5	6	-	-
Image SVM	5	-	0,00276	90,5
Audio GMM	4	4	-	-
Audio SVM	5	-	2,8284	90,5

Table 1: Optimized parameters for all classifiers

Table 1 shows all optimized parameter values: feature dimension D of the LDA transformation (see section 3.2), the SVM parameters γ and C_{SVM} (Eq. 2) as well as the number of Gaussians M within a GMM (Eq. 1).

4. EVALUATION

Both classifiers that were introduced in section 3.3 were tested in our framework for their use in the application of separating moods in images and music. For every mood (aggressive, melancholic, euphoric and calm) a training and a test run were processed while optimization was performed within the training step.

4.1 Evaluation Measures

In order to measure the classification performance we use three kinds of result preparations. Since we have no true negative ground-truth we incorporate the number of correct-classifications CC, the number of mis-classifications MC and the number of rejected or unclassified test items UC. The first measure is the overall accuracy for all moods without rejecting: $accuracy = \frac{CC}{CC+MC}$. The second preparation is the confusion diagram which shows the CC and MC for each mood class and each approach without rejecting, see Fig. 5. The third measure is the precision-recall (PR) curve which covers the rejecting capabilities by utilizing the varying threshold value C_{thr} for the normalized classification confidence C to reject classifications. The parameter C_{thr} varies between 25% and 95% with respect to a lowest normalized confidence maximum of 25% using four classes and only a small number of remaining classifications with above 95% confidence. Precision and recall are calculated as: $precision = \frac{CC}{CC+MC}$, $recall = \frac{CC}{CC+MC+UC}$.

4.2 Reference Set

Generating a reference set for multi-modal mood classification was one of the key challenges and goals of this work. We decided to utilize a two step mechanism for generating the reference set, first collecting already tagged data and second a personal review step to validate the tags. While collecting data we concentrated on music and images files that are publicly available on the Internet and most often under creative commons license.
The images are completely taken from flickr (Photo community: flickr.com) using several search keywords per mood e.g. *aggressive*: tension, anger, crude, evil, hell; *euphoric*: color, active, great, euphoria; *calm*: mellow, calm, joyful; *melancholic*: sad, contemplation. Based on that process we generated a list of about 300 images per mood. For the music collection we searched various online music platforms (e.g. last.fm, tonspion.de ...) for freely available music, partly direct from different band websites. We considered music from a broad range of genres e.g. Electronic, Metal, Indie, Pop, Reggae, Rock or Techno to get a mixture of popular music with a high variance of characteristics.
For the personal review step three persons reviewed all photos and songs. If one person gave a veto the media item was rejected. As a result of this procedure a collection of 100 music files and 100 photos for each mood are available. The reviewing process underlied two requirements: only one song per artist is chosen and images with a mood impact caused by semantic aspects are avoided. These requirements were not completely feasible. We got seven artists that are present with two songs each and some images have partly semantic impacts. Finally, we divided each mood set into a defined training and a test set each with 50 media items.

4.3 Results and Discussion

In this section we present the results of the classification experiments by discussing the three result preparations: overall accuracy, confusion diagrams and PR curves.
The mean recognition rates of our framework are: GMM/ Audio - 48,5%, SVM/Audio - 48,5%, GMM/Image - 44% and SVM/Image - 53,5%. These results seem to be worse compared with examples of the reviewed approaches especially when keeping in mind that a random approach achieves 25%. On the other hand it suggests that our reference set is very heterogenous. But this heterogeneity is needed to cover the different interpretations of mood of various users. Summarizing the overall results, each approach returns a correct classification for every second query.
A more detailed examination of the heterogeneity allows the confusions illustration depicted in Fig. 5. Especially for music we can see worse results and a high confusion of euphoric and calm moods which indicates a particular heterogeneity of these mood data sets The best results could be achieved for aggressive and melancholic music with an accuracy of about 68% and 58% which let assume that these moods are the most clear ones. We have to point out that the moods euphoric and calm were three of four times more often misclassifed than correctly classified. During a random subjective verification of the false classifications we noticed that many songs consist of different parts e.g. chorus and verse with quite different mood impacts. This problem should be concerned in future activies.
The image classification confusion is less than in music classification and the correct mood achieves the best results on each set. Nevertheless, calm seems also to be more problematic and confusable with melancholic images. Note that there is a low confusion of euphoric and melancholic images which could be explained by their opposite impression.
The final examination of the results is considered in the PR curve (Fig. 6), which depicts the recognition as well as the rejecting performance of the classification approaches. The diagram shows quite non-straight-line curves for each setup. This indicates that the number of test data, 200 items per setup, is not sufficient for a statistically reliable analysis and also substantiates the thesis of the high heterogeneity of the test data. Nevertheless, the typical curve with increasing recall while decreasing precision and vice versa is identifiable. The overall best classification results seems be the

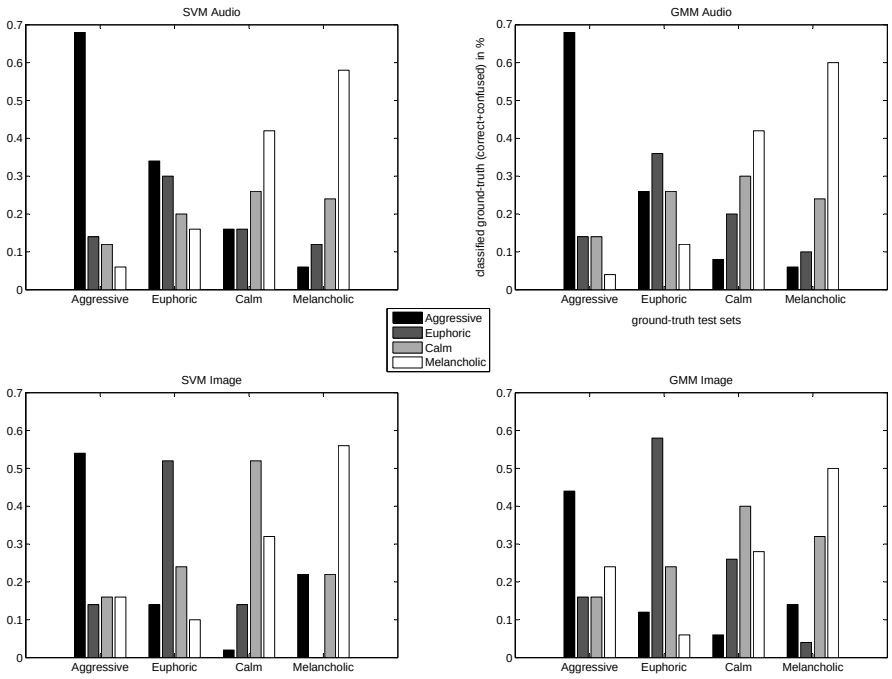

Figure 5: Confusion of the classification results of GMM and SVM classifiers. The x-axis contains the 4 mood ground-truth test sets, the y-axis depicts the percentage of the classified classes (correct and confused).

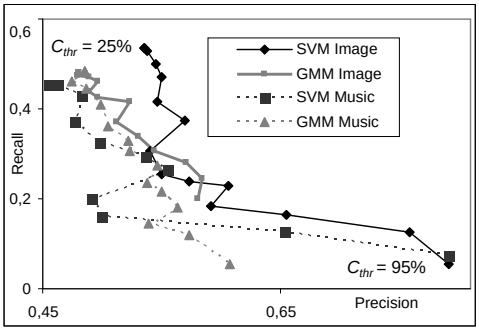

Figure 6: PR curve depending on the confidence threshold C_{thr} between 25% and 95%.

SVM/Image setting with the most often highest curve. A final statement for a better classifier scheme, generative or discriminative, cannot be given. But the PR curves and the knowledge about the reference set implies better results by a GMM in more heterogeneous data and by a SVM in more homogeneous data.

5. CONCLUSION AND FUTURE WORK

A perfect mood classification is not yet possible and will presumably never be possible due to the subjective impact of mood to individuals. Nevertheless, this paper contributed a review of various mood models and the selection of a universal dimension and category-based model. The mood model allows an easy mapping of dimension and category model type and a multi-modal use for all given use cases in this publication. A generic classification framework was presented that has the capability to process images and music in a parallel manner and allows an optimization of different system parameters. A novel reference set for music and photo mood classification was presented which can be used by different research groups for comparing their results. While dividing the reference set into fixed training and test set and applying the classification approaches in a well defined way a baseline evaluation for multi-modal mood classification for further comparisons could be achieved. Even if the results show that the reference data are not enough or not detailed enough annotated they can be used in combination with the results at least as a baseline e.g. for a new task within a multimedia retrieval benchmarking contest.
Summarizing the results, different topics in the field of multi-modal mood classification need to be addressed in the future. First, more mood specific features for music and photos are needed. Additional classifiers should be integrated in the classification framework. The evaluation workflow should be extended to a Monte Carlo approach. Then multiple reference data combinations as training and test sets are setup and more reliable results are generated through the usage of the average recognition rate. Furthermore, the reference set needs to be extended and more detailed annotated e.g. with multiple moods per item or a segment-wise annotation for music. Finally, investigation of user specifiable mood categories could lead to valuable applications by taking care of individual mood perception of the users.

6. ACKNOWLEDGMENTS

This work has been partly supported by the PHAROS Integrated Project (IST-2005-2.6.3), funded under the EC IST 6th Framework Program and by grant No. 01MQ07017 of the German THESEUS program.

7. REFERENCES

[1] T. Li, M. Ogihara, and Q. Li. A comparative study on content-based music genre classification. *Proc. ACM SIGIR '03*, pages 282–289, 2003.

[2] A. Bosch, A. Zisserman, and X. Munoz. Scene classification via pLSA. *Proc. ECCV*, 4:517–530, 2006.

[3] A. Laengle. Zur Begrifflichkeit der Emotionslehre in der Existenzanalyse. *Laengle A (Hg) Emotion und Existenz. Wien: WUV-Facultas*, pages 185–200, 2003.

[4] K. Hevner. Experimental studies of the elements of expression in music. *American Journal of Psychology*, 48(2):246–268, 1936.

[5] C.H. Chen, M.F. Weng, S.K. Jeng, and Y.Y. Chuang. Emotion-Based Music Visualization Using Photos. *LNCS*, 4903:358–368, 2008.

[6] E. Schubert. *Measurement and Time Series Analysis of Emotion in Music*. PhD thesis, University of New South Wales, 1999.

[7] A. Feist and E. Stephan. Entwicklung eines Verfahrens zur Erfassung des Gefühlszustandes (VGZ). `http://psydok.sulb.uni-saarland.de/volltexte/2007/952/`.

[8] A. Mehrabian. Framework for a comprehensive description and measurement of emotional states. *Genet Soc Gen Psychol Monogr*, 121(3):339–61, 1995.

[9] R.E. Thayer. *The Biopsychology of Mood and Arousal*. Oxford University Press, USA, 1989.

[10] B. van de Laar. Emotion detection in music, a survey. *Twente Student Conference on IT*, 1:700, 2006.

[11] R. Reisenzein. Worum geht es in der Debatte um die Basisemotionen? *Försterling, Stiensmeier-Pelster & Silny (Hg.), Kognitive und emotionale Aspekte der Motivation*, pages 205–237, 2000.

[12] Y. Feng, Y. Zhuang, and Y. Pan. Music information retrieval by detecting mood via computational media aesthetics. In *IEEE/WIC*, pages 235–241, 2003.

[13] D. Yang and W. Lee. Disambiguating music emotion using software agents. In *Proc. of the Int. Conf. on Music Information Retrieval*, 2004.

[14] L. Lu, D. Liu, and H.J. Zhang. Automatic mood detection and tracking of music audio signals. *IEEE Trans. Audio, Speech & Language Process*, 14(1), 2006.

[15] T. Li and M. Ogihara. Content-based music similarity search and emotion detection. *IEEE Int. Conf. on Acoustics, Speech & Signal Processing*, 5, 2004.

[16] M. Tolos, R. Tato, and T. Kemp. Mood-based navigation through large collections of musical data. *2nd IEEE Consumer Communications and Networking Conference*, pages 71–75, 2005.

[17] T.L. Wu and S.K. Jeng. Probabilistic estimation of a novel music emotion model. In *14th International Multimedia Modeling Conference*. Springer, 2008.

[18] M. Wang, N. Zhang, and H. Zhu. User-adaptive music emotion recognition. In *7th Int. Conf. on Signal Processing*, volume 2, pages 1352–1355, 2004.

[19] T. Li and M. Ogihara. Detecting emotion in music. *Proc. Int. Symp. Music Information Retrieval*, 2003.

[20] D. Liu, L. Lu, and H.J. Zhang. Automatic mood detection from acoustic music data. *Proc. Int. Symp. Music Information Retrieval*, pages 81–87, 2003.

[21] C.C. Liu, Y.H. Yang, P.H. Wu, and H.H. Chen. Detecting and classifying emotion in popular music. In *JCIS*, 2006.

[22] 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. Systems, Man and Cybernetics*, 4, 2006.

[23] H.W. Yoo. Visual-based emotional descriptor and feedback mechanism for image retrieval. *Journal of Information Science and Engineering*, 22(5):1205–1227, 2006.

[24] T. Soen, T. Shimada, and M. Akita. Objective evaluation of color design. *Color Res Appl*, 12(4):184–194, 1987.

[25] S.B. Cho. Emotional image and musical information retrieval with interactive genetic algorithm. In *Proc. IEEE*, volume 92, pages 702–711, 2004.

[26] Y. Guo and H. Gao. Emotion recognition system in images based on fuzzy neural network and hmm. *5th IEEE Int. Conf. on Cognitive Informatics*, 1, 2006.

[27] X. Hu, M. Bay, and J. S. Downie. Creating a simplified music mood classification ground-truth set. *Int. Symp. on Music Information Retrieval*, pages 309–310, 2007.

[28] W.N. Wang and Y.L. Yu. Image emotional semantic query based on color semantic description. *Proc of the 4th Int. Conf. on Machine Learning and Cybernetics*, 7:4571–4576, Aug 2005.

[29] W.N. Wang, Y.L. Yu, and J.C. Zhang. Image emotional classification: static vs. dynamic. *IEEE Int. Conf. on Systems, Man and Cybernetics*, 7, 2004.

[30] S. Kim, S. Kim, S. Kwon, and H. Kim. A music summarization scheme using tempo tracking and two stage clustering. *IEEE Workshop on Multimedia Signal Processing*, pages 225–228, 2006.

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

[32] I. Crüger. `http://www.ipsi.fraunhofer.de/~crueger/farbe/farb-wirk1.html`.

[33] J. Monaco. Film verstehen. Kunst, Technik, Sprache, Geschichte und Theorie des Films und der Medien. *Aufl., Hamburg/London/New York*, 2000.

[34] H. Gembris. Wie Musik auf den Menschen wirkt. *Korczak, D. & Hecker, J.(Hg.): Praktische Psychologie, Bd*, 23:236–274, 2000.

[35] B.S. Manjunath, P. Salembier, and T. Sikora, editors. *Introduction to MPEG-7*. John Wiley & Sons, 2002.

[36] P. Dunker and M. Gruhne. Audio-visual fingerprinting and cross-modal aggregation: Components and applications. *12th IEEE Int. Symp. on Consumer Electronics*, 2008.

[37] C. Dittmar, C. Bastuck, and M. Gruhne. Novel mid-level audio features for music similarity. In *Proc. Int. Conf. on Music Communication Science*, 2007.

[38] A.R. Webb. *Statistical Pattern Recognition*. John Wiley and Sons Ltd., 2nd edition, 2002.

[39] V.N. Vapnik. *Statistical learning theory*. Wiley New York, 1998.

[40] T. Westerveld. *Using generative probabilistic models for multimedia retrieval*. PhD thesis, University of Twente, 2004.

[41] C.W. Hsu, C.C. Chang, C.J. Lin, et al. A practical guide to support vector classification. *National Taiwan University, Tech. Rep., July*, 2003.

APPENDIX

A. REFERENCE DATA SETS

The links to all reference items and the constellation of trainings and test sets are available from the authors for research and comparison. The media items cannot be provided.