Evaluation of local and global descriptors for emotional impact recognition

Syntyche Gbehounou, François Lecellier, Christine Fernandez-Maloigne

Evaluation of local and global descriptors for emotional impact recognition

Syntyche Gbèhounou, François Lecellier, Christine Fernandez-Maloigne

XLIM Laboratory, UMR CNRS 7252, University of Poitiers, F-86962 Futuroscope Chasseneuil cedex, France

Abstract

In order to model the concept of emotion and to extract the emotional impact from images, one may search suitable image processing features. However, in the literature, there is no consensus on the ones to consider since they are often linked to the application. Obviously, the perception of emotion is not only influenced by the content of the images, it is also modified by some personal experiences like cultural aspects and semantic associated to some colours or objects. In this paper, we choose low level features frequently used in CBIR especially those based on SIFT descriptors. To take into account the complex process of emotion perception, we also consider colour and texture features and one global scene descriptor: GIST. We supposed the chosen features could implicitly encode high-level information about emotions due to their accuracy in the different CBIR applications of the literature.
We test our methodology on two databases: SENSE and IAPS.

Keywords: images, local descriptors, subjective evaluations, emotions, classification

1. Introduction

In past decades, many achievements have been made in computer vision in order to replicate the most amazing capabilities of the human brain, for example image classification according semantic content or people tracking

Email addresses: syntyche.gbehounou@univ-poitiers.fr (Syntyche Gbèhounou), francois.lecellier@univ-poitiers.fr (François Lecellier), christine.fernandez@univ-poitiers.fr (Christine Fernandez-Maloigne)

in video surveillance. However, there are some aspects of our behaviour or perception which remain difficult to apprehend, for example emotion prediction from an image or a video. This has several application such as: film classification, road safety education, advertising or e-commerce, by selecting appropriate images depending on the situation.

In order to predict the emotional impact of an image or a video, one first need is to describe what an emotion is and how to categorize them. There are two emotion classifications used in the literature [1]:

1. **Discrete approach**: emotional process can be explained with a set of basic or fundamental emotions, innate and common to all human (sadness, anger, happiness, disgust, fear, ...). There is no consensus about the nature and the number of these fundamental emotions. This modeling is usually preferred in emotion extraction based on facial expressions.
2. **Dimensional approach**: on the opposite, the emotions are considered in this model as the result of fixed number of concepts such as pleasure, arousal or power, represented in a dimensional space. The choosen dimensions vary depending to the needs of the model. Russel's model is the most considered, Fig. 1, with the dimensions valence and arousal:
   - ***The valence*** corresponds to the way a person feels when looking at a picture. This dimension varies from negative to positive and allows to distinguish between negative emotions and pleasant ones.
   - ***The arousal*** represents the activation level of the human body.

   The advantage of the dimensional approach is to define a large number of emotions whithout the limitation of a fixed number of concept as the discrete ones. In spite of this advantage, some emotions can be confused (such as fear and anger in the circumplex of Russel) or unrepresented (among others surprise in Russel's model).

In the literature, a lot of works are based on the discrete modeling of the emotions; for example those of Paleari and Huet [2], Kaya and Epps [3], Wei et al. [4] or Ou et al. [5, 6, 7]. In this paper, our goal is to obtain a classification into three different classes "Unpleasant", "Neutral" and "Pleasant". To tackle this objective, we choose a dimensional approach, since in discrete one the number and nature of emotions remain uncertain. Moreover, there are concepts which cannot be assigned to a specific class (surprise for example can be pleasant or unpleasant).

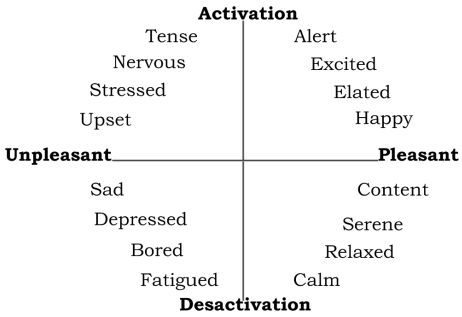

Fig. 1: Russel's emotions modeling. The axe Unpleasant/Pleasant corresponds to the arousal and the second one to the valence.

The extraction of emotional impact is an ambitious task since the emotions are not only content related (textures, colours, shapes, objects, ...), but also depend on cultural and personal experiences.

In the past decade, lots of papers have been devoted to the links between emotions and colours [4, 5, 6, 7, 8, 9, 10, 11]. Several of them consider emotions associated with particular colours through culture, age, gender or social status influences. There is a consensus among the authors to conclude that a link exists between colours and particular emotions. As stated by Ou et al. [5], colours play an important role in decision-making, evoking different emotional feelings. The research on colour emotion or colour pair emotion is now a well-established area of research. Indeed, in a series of publications, Ou et al. [5, 6, 7] studied the relationship between emotions, preferences and colours and have established a model of emotions associated with colours from psychophysical experiments.
Another part of the literature is dedicated to facial expression interpretation [2]. In this work, emotions are associated with facial features (such as eyebrows, lips). Since facial expressions are common among humans to express basic emotions (happy, fear, sadness, surprise, ...), it seems to be the easiest way to predict them. Nevertheless, in this case, the authors extract emotions carried by the images and not really those felt by someone looking at these pictures.
More recently some authors looked at the emotion recognition as a Content Based Image Retrieval (CBIR) task [12, 13, 14]. Their underlying idea consists in considering the traditional image retrieval techniques to extract the emotional impact of images. To achieve this goal, the authors used a

multistage method, at first by extracting traditional image features (colours, textures, shapes) and then combined those features into a classification system after a learning step. For example, Wang and Yu [15] used the semantic description of colours to associate an emotional semantic to an image. Concerning textures, the orientation of the different lines contained in the images is sometimes considered. According to Liu et al. [1], oblique lines could be associated with dynamism and action; horizontal and vertical ones with calm and relaxation.

Our work is part of this last family of approaches. We evaluated some low level features well adapted for object recognition and image retrieval [16, 17, 18, 19, 20, 21, 22] and conducted our work on two databases:

- A set of natural images that was assessed during subjective evaluations: Study of Emotion on Natural image databaSE (SENSE) [23];
- A database considered as a reference on psychological studies of emotions: International Affective Picture System (IAPS) [24].

This paper is organized as follow. We describe the image databases in Section 2 and the features used for emotion recognition in Section 3. The classification process is explained in Section 4. In Section 5 we summarize our results. We conclude about our study and provide some perspectives in Section 6.

2. Image databases

In the domain of emotion extraction, the choice of the database is not trivial since there is no reference for all emotion studies and applications, some authors even built their own dataset without spreading it. We choose in this study to considered two databases: the first one is composed of low semantic images and the second of more semantic ones. In this paper, "low-semantic" means, that the images do not shock and do not force a strong emotional response. We think that even low semantic images, including abstract representation, may produce emotions according to the viewer sensitivity and the viewing time. We define this "low-semantic" criteria in response to some high semantic images on IAPS [24], a reference in psychological studies on emotions. In this database, which will be described in subsection 2.1, there

are images with blood, dirt, high semantic photomanipulations or naked people which might induce a bias in the assessment of the images. There is a risk of overreaction to neutral images viewed right after strong pleasant or unpleasant ones. Our aim in this paper is to evaluate the behaviour of our strategy developed for a low semantic database on a more semantic one.

2.1. The International Affective Picture System (IAPS)

This dataset is developed since the late 1980s at NIMH Center for Emotion and Attention (CSEA) at the University of Florida [24], which is composed of photographs used in emotion research. It is considered as a reference in psychological studies and many papers present results on this base [1, 13, 14].
The images of IAPS are scored according to the affective ratings: pleasure, arousal and dominance. It corresponds to a dimensional representation of emotions. The affective norms for the pictures in the IAPS were obtained in 18 separate studies involving approximately 60 pictures for each session. Each of the 1182 images from their dataset was evaluated by about 100 participants.[1] The emotion of the images we have chosen is based of those defined in different papers [25, 26, 27] using three class classification: "Pleasant", "Neutral" and "Unpleasant".

2.2. SENSE

Study of Emotion on Natural image databaSE (SENSE) is the database used in [23, 28]. It is a low semantic, natural and diversified database containing 350 images free to use for research and publication. It is composed of animals, food and drink, landscapes, historic and tourist monuments as shown on Fig.2 with some examples.
Some transformations were applied on some images of the database: geometric modifications and changes on colour balance, so some images are repeated twice or more. In this database, only 17 images (4.86%) contain human faces to ensure that the facial expression does not induce bias. It is composed of low semantic images since they minimize the potential interactions between emotions on following images during subjective evaluations. This aspect is important to ensure that the emotion indicated for an image is really related to its content and not to the emotional impact of the previous one. In

[1] It is the size of the database when we received it.

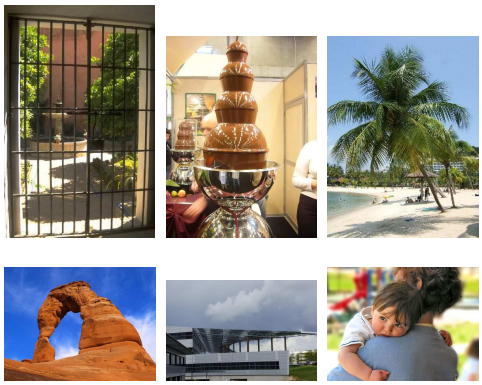

Fig. 2: Some images from SENSE.

fact, in case of strong emotions, one can imagine that the emotion felt for an image could modify the decision for the following image(s). This low-semantic aspect also allow different applications, for example, an application of re ranking images according to their emotional impact or an application of cook recipe image retrieval that eventually contains low semantic images. The short viewing duration is another constraint in the evaluation to reduce the semantic interpretation.
During the tests the observers assessed nature and power of emotional impact of the images. For the nature, the different choices are "Negative", "Neutral" or "Positive" and the power varies from "Low" to "High". These information where chosen as the easiest way to evaluate a "primary" emotion for low semantic images. Discrete modeling is not adapted in this case. In fact, as stated in introduction, in a discrete representation of emotions, emotional process is explained with a set of basic or fundamental emotions, innate and common to all human, often based on facial expressions.
In this database, each image was assessed by an average of 104.81 observers and the emotion is given after an average observation time of 6.5 seconds.

3. Images features

As stated is the introduction, many works find a link between colours or textures and emotions, we choose to extract those information using local features (interest point descriptors) and global ones (textures, colours). Since SIFT and other local descriptors are well adapted for CBIR, we will sum-

marize some important information about them and their colour extensions. Then, we present global descriptors considered in this work.

3.1. Features based on local information

In order to extract local features, one will always need a two step algorithm:

1. Find a set of interest points;
2. Compute the descriptor associated to each point.

In this work, to detect the interest points, we choose Harris-Laplace [29] which has shown good performance for category recognition according to Zhang et al. [30].
As far as descriptors are concerned, we opt for those evaluated by van de Sande et al [22]:

- SIFT and some colours extensions (CSIFT and OpponentSIFT),
- Colour Moments,
- Colour Moment Invariants.

Those choices were lead by the efficiency of SIFT and its extensions, in the domain of image retrieval. SIFT, introduced by Lowe [19, 20] was originally computed on greyscale images so, for this descriptor we converted the images to greyscale with the formula of NTSC standard (Equation (3)). The original version, proposed by Lowe, is invariant to translation, image scaling, rotation, change in illumination and affine or 3D projection.
We also compute two colour extensions of SIFT. The first one, OpponentSIFT, was proposed by van de Sande et al. [22]. They describe all the channels in the opponent colour space using SIFT descriptors as shown in equation (1):

$$\begin{pmatrix} 0_1 \\ 0_2 \\ 0_3 \end{pmatrix} = \begin{pmatrix} \frac{R-G}{\sqrt{2}} \\ \frac{R+G-2B}{\sqrt{6}} \\ \frac{R+G+B}{\sqrt{3}} \end{pmatrix}. \tag{1}$$

The information in the O_3 channel is equal to the intensity information, while the two others describe the colour information in the image.
We consider also a second colour extension of the greyscale SIFT: C-SIFT. This descriptor was firstly suggested by Abdel-Hakim and Farag [16] using

the colour invariants developed by Geusebroek et al. [31]. This approach builds the SIFT descriptors in a colour invariant space. Its performance was proved in the evaluations made by van de Sande et al. [22] on PASCAL VOC 2007 [32].

Besides SIFT descriptors (and extensions), we also tested colour moments and colour moment invariants in order to deal with fully colour descriptors. Colour moments are measures used to differentiate images based on their colour features. Once calculated, these moments provide a measurement for colour similarity between images.
For a colour image corresponding to a function I with RGB triplets, for image position (x, y), the generalized colour moments are defined by the equation (2):

$$M_{pq}^{abc} = \iint_{\Omega} x^p y^q \left[I_R(x,y)\right]^a \left[I_G(x,y)\right]^b \left[I_B(x,y)\right]^c \, \mathrm{d}x\mathrm{d}y, \tag{2}$$

where Ω is a region around the local feature.
M_{pq}^{abc} is known as generalized colour moment of order $p+q$ and degree $a+b+c$. Only generalized colour moments up to the first order and the second degree are considered to avoid numeric instability, thus the resulting invariants are functions of the generalized colour moments M_{00}^{abc}, M_{10}^{abc} and M_{01}^{abc}:

$$with\,(a,b,c) \in \left\{ \begin{array}{c} (1,0,0)\,,(0,1,0)\,,(0,0,1) \\ (2,0,0)\,,(0,2,0)\,,(0,0,2) \\ (1,1,0)\,,(1,0,1)\,,(0,1,1) \end{array} \right\}.$$

Those moments are only invariant to light intensity shift [22]. In order to add some colour invariants, there is also an invariant version of colours moments: colour moment invariants proposed by Mindru et al. [33]. The authors use generalized colour moments for the definition of combined invariants to the affine transform of coordinates and contrast changes. Compared to colour moments these descriptors are invariant to light intensity shift and change and light colour change and shift [22].

3.2. Features based on global information

We also tested some global information features (colour, texture and global scene description).

To identify the different colours of each image, we used colour segmentation by region growing [34] which is a two step algorithm: initialize the

seeds and then grow region from seeds in order to obtain a segmented image. For the initialization of the seeds, we performed an analysis of greyscale histogram. This analysis was made in greyscale to save time in homogeneous areas. To convert colour images to greyscale we use again the equation (3) according to the NTSC standard.

$$grey = 0.299R + 0.587G + 0.114B \tag{3}$$

The seeds are then the maxima of the greyscale histogram. The region growing was performed in CIE Lab 1976 colour space in order to use a perceptual distance. In our case, we consider the distance computed with ΔE and obtained with equation (4) (for two colours C_1 and C_2). Then on the segmented version, we retained only the average colour of different regions.

$$\Delta E = \sqrt{(L_1^* - L_2^*)^2 + (a_1^* - a_2^*)^2 + (b_1^* - b_2^*)^2} \tag{4}$$

To extract the textures of images, we use Wave Atom transform introduced by Ying and Demanet [35] on greyscale images. These features can be considered as a variant of 2D wavelet packets with a parabolic wavelength scale and prove their efficiency on local oscillatory textures.
Wave Atom transform is a multi-scale transform and yields information from different levels. It is an oriented decomposition where the number of coefficients for each orientation depends on the decomposition level. Before applying Wave Atom transform we resized all image to $256 * 256$ with zero padding if it is needed. With this new size, we obtain 5 levels of decomposition and only retained the scales 4 and 5 due to the small size of the three other scales. On scale 4 we obtain a set of 91 orientations, each of them containing $2^4 * 2^4$ (256) coefficients. Scale 5 contains 32 orientations and 1024 coefficients per orientation.

Our global scene description of the image is obtained using GIST introduced by Oliva and Torralba in 2001 [36]. This descriptor leads to a low dimensional representation of image. It is obtained with a set of perceptual dimensions (naturalness, openness, roughness, expansion, ruggedness) that represent the dominant spatial structure of a scene. These dimensions are estimated using spectral and coarsely localized information. For our study we compute GIST on images resized to $256*256$ with zero padding to respect the recommendations provided by the authors.

4. Classification process

Traditionally, for CBIR task, an image is represented by a visual signature, which concentrates the usefull information of the content of the image into a smaller descriptor. For a recognition or retrieval task, two images are considered as visually close if their visual signatures are close too. There are different ways to obtain this visual signature [17, 37, 38], but the large majority of them depends on a visual codebook and visual words.
We evaluate two different visual signatures methods: Bag of Visual Words (BoVW) [37] and VLAD [17]. Here, the visual codebook is a set a visual words (a set of features) obtained by a *K-Means* algorithm. The idea is to resume an image database with K visual words (features). Then, after obtaining the signature, the decision process on emotion recognition uses a classification algorithm such as SVM for example.

4.1. Compact representation of features vectors with visual signature

The visual signature named Bag of Visual Words (BoVW) was initially proposed by Sivic and Zisserman [37] with applications on images and videos and is inspired by the method "bag of words" used in text categorization. An image is represented here by the histogram of occurrences of the various visual words in the codebook.

VLAD is another visual signature which describes the distance between each feature and its nearest visual word. VLAD can be seen as a simplification of the Fisher kernel [38].
Let $\mathcal{C} = \mathcal{C}_1, \ldots, \mathcal{C}_K$ a visual codebook composed of K visual words obtained with *K-means* algorithm, each local descriptor $\mathcal{D}$ is associated to its nearest visual word $\mathcal{C}_i$ as shows the equation (5).

$$\mathcal{V}_k = \sum_{\mathcal{D}_n : NN(\mathcal{D}_n) = \mathcal{W}_k} (\mathcal{D}_n - \mathcal{W}_k). \tag{5}$$

The idea of the VLAD descriptor is to accumulate, for each visual word $\mathcal{C}_i$, the differences $\mathcal{D} - \mathcal{C}_i$ of the vectors $\mathcal{D}$ assigned to $\mathcal{C}_i$. This characterizes the distribution of the vectors with respect to the center. Assuming the local descriptor to be d-dimensional, the dimension D of VLAD representation is $D = K * d$.

Except for GIST, on all our descriptors the visual codebooks are obtained with a *K-Means* algorithm. The size of the *K-Means* codebook is obtained with the equation (6):

$$K = \sqrt[4]{N * d}, \tag{6}$$

where K is the number of visual words and N is the number of descriptors and d the size of vector of characteristics.
For VLAD we only compute K=64 visual words based on the results obtained by Jégou et al [17].
In the particular case of GIST, we use PCA as advised by Oliva and Torralba [36]. To perform PCA we compute the principal components on the GIST descriptors from the full database considered. We select the K axes that preserve 99% of the initial information.
After computing the visual signature, we apply a L2-normalization to allow the comparison of the signatures coming from different images.

4.2. Classification with SVM

For emotion recognition, Liu et al. [1] among others, use an SVM classifier with a linear kernel in its multiclass extension "One against one" for classification. They showed that this method gives the best results using their descriptors and databases, so we naturally choose to test our descriptors and databases with this classifier.
The inputs of the classifier are the visual signatures according to the considered descriptor. We performed a two class classification: positive and negative emotions. We do not consider the "Neutal" class since, during our database assessment on SENSE, this class was given by observers which whether cannot decide about the nature of the emotion or really feel neutral when looking at the images. Moreover, the neutral images are rarely considerd in the literature [4, 13] due to the same restriction. On IAPS, some authors [25, 26, 27] list in their articles the images according to their hedonic valence that induces the three groups: Pleasant, Neutral and Unpleasant images, which respectively correspond to positive, neutral and negative. However, to perform a fair comparison between our results on the two databases in the rest of this paper we only keep positive and negative images.
After classification for each descriptor we combined the results for the different descriptors using "Majority Voting" algorithm as shown on the Fig. 3.

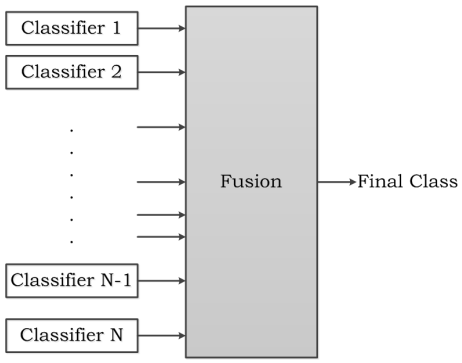

Fig. 3: Algorithm of classification fusion. Concerning "Majority Voting" the final class is the class predicted by a maximum of classifiers.

5. Classification results

In this Section we present our results for local and global feature evaluation for emotional impact recognition. We firstly discuss our results on SENSE and IAPS, then we compare those from IAPS to some baselines from the literature. The configuration of the different image databases (learning and test sets) is given in the appendix.

5.1. Results on SENSE and IAPS and discussions

In Table 1, we summarize the results obtained after classification for each descriptor. In this table:

- WA4 and WA5 respectively means Wave Atoms Scale 4 and Wave Atoms Scale 5;
- CM denotes Colour Moments and CMI Colour Moment Invariants;
- OpSIFT means OpponentSIFT;
- We use the notation Dataset_Visual codebook to resume the different configurations we have tested. Then in SENSE_I configuration, the visual signatures of the images of SENSE are computed using the visual vocabulary from IAPS. The different configurations allow us to determine whether or not the results are dependent on the image database used to create the visual dictionary.

In [39], we have studied the impact of the visual signature according to the nature of the features. We noticed that VLAD representation matches better for the local descriptors and BoVW for global ones. In general, on local features, VLAD outperformed BoVW by 10% and on global features underperformed them by 20%, interested reader may refer to [39] for more details. Then for the following results we use BOVW for global feature descriptors and VLAD for local feature descriptors.
One can easily see on Table 1 that the selected descriptors have a different behaviour whether the nature of emotion is positive or negative. However the average of correct classification for each nature of emotion remains higher than the random selection (50%). In general, when considering IAPS as test database or as visual codebook, the negative images are more easily classified. It comes from the nature of images in IAPS since many of them are voluntarily modified to ensure a strong emotional response. For SENSE database, the average results are more consistent for negative and positive emotions with 55.55% and 54.16%. We can also conclude that the visual dictionary has little impact on the behaviour of descriptors for SENSE and IAPS. Those global average results prove that our choice to consider classic CBIR features for emotion evaluation gives interesting results.

When differentiating the descriptors, the results are less consistent. For example, SIFT have approximately the same results for negative and positive emotions with yet a higher classification rate for negative ones (63.28%). It is also the only descriptor which allows a classification significantly higher than random one for both positive and negative emotions whether the dataset and the visual dictionary. CSIFT and OpSIFT compete with SIFT for global recognition of emotions, but OpSIFT is less efficient for negative images with SENSE or using SENSE visual codebook. In spite of this less efficiency, it is the best descriptor for IAPS using IAPS codebook with 60.66% for negative images and 63.79% for positive ones. CSIFT gives the best negative estimation of all global and local descriptors with 77.75% in average and even 90% for SENSE using IAPS codebook. Colour descriptors tend to extract more negative images except for SENSE using SENSE dictionary. In this case, for CM and Colours, the classification rates for positive images are higher than 80%. The other descriptors give less reliable results on both positive and negative images with a very low score for WA4 on positive images on IAPS with the codebook obtained with SENSE. However WA4 and WA5 are complementary since WA4 tends to extract negative images and WA5 positive

Table 1: Classification rates after classification for each descriptor. SENSE_I configuration corresponds to the results obtained where the visual signatures of the images of SENSE are computed using the visual vocabulary from IAPS. For SENSE_S configuration the visual signatures are obtained using SENSE visual codebook. For IAPS_I and IAPS_S the visual signatures of the images of IAPS are respectively computed using the visual vocabularies from IAPS and SENSE.

Descriptors		**Nature of emotions**	**Configuration Test database_Visual codebook**				**Average**
			SENSE_S	***SENSE_I***	***IAPS_S***	***IAPS_I***	
Global descriptors	***Colours***	Negative	40%	70%	85.25%	78.69%	68.49%
		Positive	80.21%	43.75%	27.59%	29.31%	45.22%
	WA4	Negative	50%	50%	77.05%	68.85%	61.48%
		Positive	30.21%	52.08%	20.69%	32.76%	33.94%
	WA5	Negative	30%	60%	57.38%	44.26%	47.91%
		Positive	50%	65.62%	41.38%	58.62%	53.91%
	GIST	Negative	90%	40%	42.62%	62.3%	58.73%
		Positive	27.08%	61.46%	56.90%	37.93%	45.84%
Local descriptors	***CM***	Negative	10%	80%	40.98%	60.66%	47.91%
		Positive	88.54%	54.17%	68.97%	51.72%	**65.85%**
	CMI	Negative	70%	60%	60.66%	86.89%	69.39%
		Positive	57.29%	58.33%	55.17%	27.59%	49.60%
	SIFT	Negative	70%	70%	52.46%	60.66%	63.28%
		Positive	56.25%	52.08%	51.72%	53.45%	53.38%
	CSIFT	Negative	80%	90%	73.77%	67.21%	**77.75%**
		Positive	50%	54.17%	53.45%	50%	51.91%
	OpSIFT	Negative	60%	60%	65.57%	60.66%	61.56%
		Positive	47.92%	52.08%	48.28%	63.79%	53.02%
	Average	Negative	**55.55%**	**64.44%**	**61.75%**	**65.58%**	**61.83%**
		Positive	**54.16%**	**54.86%**	**47.13%**	**45.02%**	**50.29%**

ones.

Regarding the obtained results, all the descriptors may enhance the classification rate since the misclassified images are different. However, it appears rather clearly that negative images will be more easy to extract than positive ones. To verify this hypothesis we performed a fusion based on Majority Voting. The results are presented in Table 2.

Table 2: Comparison of correct average classification rates on SENSE and IAPS before and after fusion with Majority Voting.

		Before fusion	After fusion
SENSE_S	Negative	55.56%	60%
	Positive	54.17%	57.29%
	Average	54.87%	58.65%
SENSE_I	Negative	64.44%	90%
	Positive	54.86%	64.58%
	Average	59.65%	77.29%
IAPS_S	Negative	61.75%	75.41%
	Positive	47.13%	41.38%
	Average	54.44%	58.40%
IAPS_I	Negative	65.58%	77.05%
	Positive	45.02%	46.55%
	Average	55.30%	61.80%

In Table 2 we compare the classification rates before and after fusion with Majoriting Voting. As we think considering the previous results, there is a significant improvement after the fusion. For example, regardless the codebook used, the recognition of negative images is increased by 15% on average for the two datasets. Moreover, the best classification rates are obtained after merging using the dictionary built from IAPS. Before the fusion, 54.86% and 45.02% of positive images were respectively recognized on SENSE and IAPS against 64.58% and 46.55% after. If we generally consider these results after fusion, we see that they have especially been improved on our image database (SENSE), independently of visual dictionaries and

emotions:

- $\sim +15\%$ for negative images and $\sim +6\%$ for positive ones;
- $\sim +17\%$ with the codebook from IAPS and $\sim +3.7\%$ with the codebook from SENSE.

Note that for IAPS, positive image average results (47.13% and 45.02%) are lower than a simple random selection (50%). There are two hypotheses for this: the construction of the database itself or the fact that negative images are easier to recognize.
The proposed algorithm highlights the complementarity of the chosen features. In fact, we have also tested different fusion combinations and the better configuration is the fusion of all of the features to obtain the best classification of positive and negative images. We also prove that CBIR methods are effective for the emotional impact recognition.

5.2. Comparison with literature on IAPS

As stated in the introduction, we want to point out the difficulty of comparison of results on emotion recognition since the datasets vary and, even on IAPS, there is no consensus on the training set. Regarding the divergences induced by this disparity, the goal of this comparison is only to present the interval of the classification rates.
We chose three results on IAPS to make the comparison:

- Wei et al. [4] using a semantic description of the images for emotional classification of images. The authors chose a discrete modeling of emotions in 8 classes: "Anger", "Despair", "Interest", "Irritation", "Joy", "Fun", "Pride" and "Sadness". The classification rates they get are between 33.25% for the class "Pleasure" and 50.25% for "Joy."
- Liu et al. [1] proposing a system based on colour, texture, shape features and a set of semantic descriptors based on colours. Their results on IAPS are 54.70% in average after a fusion with the Theory of evidence and 52.05% with MV fusion. For their classification, they held four classes by subdividing the dimensional model Valence/Arousal into four quadrants; those defined by the intersection of the axes.
- Machajdik et al. [13] using colour, texture, composition and content descriptors. They chose a discrete categorization in 8 classes: "Amusement", "Anger", "Awe", "Contentment", "Disgust", "Excitement",

"Fear" and "Sad". The average rates of classification are between 55% and 65 %. The lowest rate is obtained for the class "Contentement" and the highest for the class "Awe". The results are from the best feature selections implemented during their work.

If we compare our results with those three, we can conclude that our method is really relevant. Our classification rates are in the high average on IAPS: 54.44% and 55.30% before fusion; 58.82% and 62.18% after. Moreover, our approach allows us to compete with the best scores from the literature even outperform them in terms of classification rate.

6. Conclusions and future works

In this paper, we propose an evaluation of different standard CBIR features for emotion prediction on low and high semantic images. The emotional impact of images is the result of many parameters both intrinsic and extrinsic. On the one hand, colours, textures or shapes are intrinsic parameters. On the other hand, culture, state of mind and experience of the viewer can be considered as extrinsic. Our goal in this work is to model the intrinsic ones using local and global features, and to deal with the large variability of images. So we have tested our algorithm on two databases: one low semantic (SENSE) and the other more semantic (IAPS). Whether they are local or global, the different descriptors have proven their efficiency for emotional impact recognition on both databases. Moreover, the compact representation of local features using VLAD seems to focus their emotional response and give very promising results for SIFT based descriptors.

One of the important perspective in our study will be evaluations of saliency for different emotions. For this purpose, we plan to hold new psychovisual tests with an eye tracker. We can then assess visual attention of observers based on the emotional impact of images. Another interesting perspective is to use another algorithm of combining, for example Evidence Theory that gives best results compared to Majority Voting on the emotion recognition solution proposed by Liu et al. [1].

[1] E. Liu, N.and Dellandréa, L. Chen, Evaluation of features and combination approaches for the classification of emotional semantics in images, in: International Conference on Computer Vision Theory and Applications, 2011.

[2] M. Paleari, B. Huet, Toward emotion indexing of multimedia excerpts, Proceedings on Content-Based Multimedia Indexing, International Workshop (2008) 425–432.

[3] N. Kaya, H. H. Epps, Color-emotion associations: Past experience and personal preference, in: Proceedings of AIC Colors and Paints, Interim Meeting of the International Color Association, 2004.

[4] K. Wei, B. He, T. Zhang, W. He, Image Emotional Classification Based on Color Semantic Description, Vol. 5139 of Lecture Notes in Computer Science, Springer Berlin / Heidelberg, 2008, pp. 485–491.

[5] L. C. Ou, M. R. Luo, A. Woodcock, A. Wright, A study of colour emotion and colour preference. part i: Colour emotions for single colours, Color Research & Application 29 (3) (2004) 232–240.

[6] L. C. Ou, M. R. Luo, A. Woodcock, A. Wright, A study of colour emotion and colour preference. part ii: Colour emotions for two-colour combinations, Color Research & Application 29 (4) (2004) 292–298.

[7] L. C. Ou, M. R. Luo, A. Woodcock, A. Wright, A study of colour emotion and colour preference. part iii: Colour preference modeling, Color Research & Application 29 (5) (2004) 381–389.

[8] C. Boyatziz, R. Varghese, Children's emotional associations with colors, The Journal of Genetic Psychology 155 (1993) 77–85.

[9] M. P. Lucassen, T. Gevers, A. Gijsenij, Adding texture to color: quantitative analysis of color emotions, in: Proceedings of CGIV, 2010.

[10] M. M. Bradley, M. Codispoti, D. Sabatinelli, P. J. Lang, Emotion and motivation ii: Sex differences in picture processing, Emotion 1 (3) (2001) 300–319.

[11] L. Beke, G. Kutas, Y. Kwak, G. Y. Sung, D. Park, P. Bodrogi, Color preference of aged observers compared to young observers, Color Research & Application 33 (5) (2008) 381–394.

[12] M. Solli, R. Lenz, Emotion related structures in large image databases, in: Proceedings of the ACM International Conference on Image and Video Retrieval, ACM, 2010, pp. 398–405.

[13] J. Machajdik, A. Hanbury, Affective image classification using features inspired by psychology and art theory, in: Proceedings of the international conference on Multimedia, 2010, pp. 83–92.

[14] V. Yanulevskaya, J. C. Van Gemert, K. Roth, A. K. Herbold, N. Sebe, J. M. Geusebroek, Emotional valence categorization using holistic image features, in: Proceedings of the 15th IEEE International Conference on Image Processing, 2008, pp. 101–104.

[15] W. Wang, Y. Yu, Image emotional semantic query based on color semantic description, in: Proceedings of the The 4th International Conference on Machine Leraning and Cybernectics, Vol. 7, 2005, pp. 4571–4576.

[16] A. E. Abdel-Hakim, A. A. Farag, CSIFT: A SIFT Descriptor with color invariant characteristics, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition), 2006.

[17] H. Jégou, M. Douze, C. Schmid, P. Pérez, Aggregating local descriptors into a compact image representation, in: Proceedings of the 23rd IEEE Conference on Computer Vision & Pattern Recognition, IEEE Computer Society, 2010, pp. 3304–3311.

[18] Y. Ke, R. Sukthankar, PCA-SIFT: a more distinctive representation for local image descriptors, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 2, 2004, pp. 506–513.

[19] D. G. Lowe, Object recognition from local scale-invariant features, International Conference on Computer Vision 2 (1999) 1150–1157.

[20] D. G. Lowe, Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision 60 (2004) 91–110.

[21] D. Nistér, H. Stewénius, Scalable recognition with a vocabulary tree, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Vol. 2, 2006, pp. 2161–2168.

[22] K. E. A. van de Sande, T. Gevers, C. G. M. Snoek, Evaluating color descriptors for object and scene recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 32 (9) (2010) 1582–1596.

[23] P. Denis, V. Courboulay, A. Revel, S. Gbehounou, F. Lecellier, C. Fernandez-Maloigne, Improvement of natural image search engines results by emotional filtering, EAI Endorsed Transactions on Creative Technologies.
URL https://hal.archives-ouvertes.fr/hal-01261237

[24] P. J. Lang, M. M. Bradley, B. N. Cuthbert, International affective picture system (IAPS): Affective ratings of pictures and instruction manual. technical report A-8, Tech. rep., University of Florida (2008).

[25] M. M. Bradley, S. Hamby, A. Lw, P. J. Lang, Brain potentials in perception: Picture complexity and emotional arousal, Psychophysiology 44 (3) (2007) 364–373.

[26] M. M. Bradley, M. Codispoti, P. J. Lang, A multi-process account of startle modulation during affective perception, Psychophysiology 43 (5) (2006) 486–497.

[27] A. Keil, M. M. Bradley, O. Hauk, B. Rockstroh, T. Elbert, P. J. Lang, Large-scale neural correlates of affective picture processing, Psychophysiology 39 (5) (2002) 641–649.

[28] S. Gbèhounou, F. Lecellier, C. Fernandez-Maloigne, V. Courboulay, Can salient interest regions resume emotional impact of an image?, in: Computer Analysis of Images and Patterns, Vol. 8047 of Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2013, pp. 515–522.

[29] K. Mikolajczyk, C. Schmid, Indexing based on scale invariant interest points, in: Proceedings of the 8th IEEE International Conference on Computer Vision, Vol. 1, 2001, pp. 525–531.

[30] J. Zhang, M. Marszalek, S. Lazebnik, C. Schmid, Local features and kernels for classification of texture and object categories: A comprehensive study, International Journal of Computer Vision 73 (2) (2007) 213–238.

[31] J. Geusebroek, R. van den Boomgaard, A. W. M. Smeulders, H. Geerts, Color invariance, IEEE Transactions on Pattern Analysis and Machine Intelligence 23 (12).

[32] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, A. Zisserman, The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results (2007).

[33] F. Mindru, T. Tuytelaars, L. Van Gool, T. Moons, Moment invariants for recognition under changing viewpoint and illumination, Computer Vision and Image Understanding 94 (13) (2004) 3–27.

[34] C. Fernandez-Maloigne, Advanced Color Image Processing and Analysis, Springer, July 2012.

[35] L. Demanet, L.and Ying, Wave atoms and time upscaling of wave equations, Numerische Mathematik 113 (2009) 1–71.

[36] A. Oliva, A. Torralba, Modeling the shape of the scene: A holistic representation of the spatial envelope, International Journal of Computer Vision 42 (2001) 145–175.

[37] J. Sivic, A. Zisserman, Video Google: A text retrieval approach to object matching in videos, in: Proceedings of the International Conference on Computer Vision, 2003, pp. 1470–1477.

[38] F. Perronnin, C. R. Dance, Fisher kernels on visual vocabularies for image categorization, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE Computer Society, 2007.

[39] S. Gbehounou, Image database indexing: Emotional impact evaluation, Theses, Université de Poitiers (Nov. 2014).

IAPS

Learning set

Negative images: *248 images.* 1019, 1050, 1051, 1052, 1080, 1090, 1110, 1111, 1114, 1120, 1201, 1202, 1271, 1274, 1275, 1280, 1303, 1304, 1525, 1930, 2095, 2120, 2141, 2205, 2278, 2301, 2455, 2456, 2490, 2520, 2590, 2683, 2691, 2692, 2700, 2703, 2715, 2717, 2722, 2730, 2751, 2753, 2799, 2800, 2900, 2981, 3000, 3001, 3015, 3016, 3017, 3019, 3051, 3053, 3059, 3060, 3062, 3063, 3064, 3068, 3071, 3080, 3100, 3101, 3103, 3110, 3120, 3130, 3140, 3150, 3160, 3168, 3180, 3181, 3185, 3191, 3212, 3213, 3215, 3216, 3225, 3230, 3261, 3266, 3301, 3350, 3400, 3500, 3550, 4621, 5970, 5971, 6021, 6022, 6190, 6200, 6211, 6212, 6213, 6220, 6231, 6241,

6242, 6243, 6250, 6260, 6263, 6300, 6312, 6313, 6315, 6350, 6370, 6410, 6415, 6510, 6530, 6540, 6550, 6555, 6561, 6562, 6563, 6570, 6821, 6825, 6830, 6831, 6836, 6838, 7135, 7136, 7361, 7380, 8230, 8485, 9001, 9002, 9006, 9007, 9010, 9031, 9040, 9041, 9043, 9046, 9050, 9075, 9102, 9120, 9140, 9145, 9163, 9180, 9181, 9183, 9185, 9186, 9187, 9220, 9252, 9253, 9254, 9265, 9290, 9291, 9295, 9300, 9302, 9320, 9321, 9322, 9326, 9330, 9331, 9332, 9341, 9342, 9373, 9395, 9405, 9409, 9410, 9412, 9414, 9415, 9417, 9419, 9421, 9423, 9424, 9425, 9427, 9428, 9429, 9430, 9433, 9435, 9440, 9452, 9471, 9480, 9490, 9491, 9500, 9520, 9530, 9560, 9570, 9571, 9584, 9590, 9599, 9600, 9610, 9611, 9621, 9622, 9623, 9630, 9810, 9830, 9831, 9832, 9901, 9902, 9903, 9904, 9908, 9909, 9910, 9911, 9920, 9921, 9922, 9925, 9930, 9940, 9941, 2055.1, 2352.2, 2375.1, 2900.1, 3005.1, 4664.2, 6250.1, 6570.1, 9635.1

Positive images: 228 images. 1340, 1410, 1440, 1441, 1463, 1500, 1510, 1540, 1600, 1601, 1603, 1604, 1620, 1630, 1670, 1710, 1721, 1722, 1731, 1740, 1811, 1812, 1850, 1910, 1999, 2030, 2035, 2040, 2050, 2057, 2058, 2060, 2071, 2075, 2080, 2091, 2151, 2152, 2153, 2154, 2156, 2158, 2160, 2165, 2208, 2209, 2216, 2222, 2250, 2260, 2274, 2299, 2303, 2304, 2306, 2310, 2314, 2331, 2332, 2339, 2341, 2344, 2345, 2346, 2352, 2360, 2362, 2370, 2387, 2388, 2391, 2395, 2501, 2510, 2530, 2540, 2560, 2598, 2650, 2655, 2791, 4002, 4003, 4180, 4220, 4250, 4290, 4310, 4490, 4500, 4520, 4550, 4599, 4601, 4603, 4607, 4609, 4610, 4611, 4612, 4616, 4617, 4622, 4623, 4626, 4628, 4640, 4641, 4645, 4650, 4651, 4652, 4656, 4658, 4659, 4660, 4666, 4670, 4676, 4677, 4681, 4687, 4689, 4690, 4700, 5000, 5001, 5010, 5199, 5200, 5201, 5202, 5215, 5220, 5260, 5270, 5450, 5460, 5470, 5480, 5594, 5600, 5611, 5621, 5626, 5629, 5631, 5660, 5725, 5760, 5764, 5779, 5781, 5811, 5814, 5820, 5829, 5830, 5831, 5833, 5870, 5890, 5891, 5910, 5994, 7200, 7220, 7230, 7270, 7280, 7282, 7284, 7289, 7325, 7330, 7350, 7390, 7400, 7405, 7410, 7460, 7470, 7480, 7481, 7501, 7502, 7508, 7530, 7570, 7580, 8021, 8030, 8034, 8041, 8080, 8090, 8120, 8161, 8162, 8163, 8180, 8185, 8186, 8190, 8200, 8208, 8210, 8260, 8300, 8320, 8330, 8340, 8370, 8371, 8380, 8400, 8461, 8465, 8470, 8490, 8496, 8497, 8499, 8500, 8502, 8503, 8510, 8531

Test set

Negative images: 61 images. 1070, 1113, 1220, 1300, 2053, 2276, 2457, 2688, 2710, 2750, 2811, 3010, 3030, 3061, 3069, 3102, 3131, 3170, 3195, 3220, 3300, 3530, 6020, 6210, 6230, 6244, 6311, 6360, 6520, 6560, 6571, 6834, 7359, 9000, 9008, 9042, 9090, 9160, 9184, 9250, 9280, 9301, 9325, 9340, 9400, 9413, 9420, 9426, 9432, 9470, 9495, 9561, 9592, 9620, 9800, 9900, 9905, 9912, 9927, 2345.1, 3550.1

Positive images: 58 images. 1460, 1590, 1610, 1720, 1750, 1920, 2045, 2070, 2150, 2155, 2170, 2224, 2300, 2311, 2340, 2347, 2373, 2398, 2550, 2660, 4210, 4470, 4597, 4608, 4614, 4624, 4643, 4653, 4664, 4680, 4695, 5030, 5210, 5300, 5551, 5623, 5700, 5780, 5825, 5836, 5982, 7260, 7286, 7352, 7430, 7492, 7545, 8031, 8116, 8170, 8193, 8280, 8350, 8420, 8492, 8501, 8540, 2352.1

SENSE

Learning set

Negative images: 53 images. 320, 305, 338, 341, 319, 92, 171, 340, 313, 87, 211, 335, 93, 332, 330, 314,

307, 225, 334, 210, 141, 155, 327, 322, 333, 226, 216, 24, 172, 151, 90, 318, 154, 149, 107, 339, 323, 204, 170, 329, 303, 18, 100, 150, 308, 182, 189, 302, 106, 301, 57, 300, 348

Positive images: 53 images. 260, 343, 266, 76, 47, 164, 45, 2, 267, 196, 82, 337, 286, 264, 117, 29, 41, 342, 283, 113, 80, 125, 298, 78, 278, 272, 115, 10, 34, 345, 205, 131, 56, 279, 265, 159, 79, 287, 240, 165, 122, 346, 288, 281, 273, 53, 277, 129, 95, 81, 187, 297, 193

Test set

Negative images: 10 images. 35, 52, 68, 96, 103, 116, 137, 220, 311, 328

Positive images: 96 images. 4, 15, 20, 23, 25, 26, 28, 31, 32, 37, 43, 44, 46, 48, 50, 54, 55, 58, 60, 61, 65, 69, 70, 72, 73, 74, 83, 85, 86, 94, 97, 101, 102, 110, 123, 124, 126, 127, 128, 132, 133, 134, 136, 138, 139, 142, 146, 156, 160, 162, 163, 166, 179, 181, 186, 188, 190, 197, 199, 206, 207, 208, 213, 230, 238, 241, 244, 245, 246, 248, 249, 251, 253, 254, 255, 256, 258, 261, 262, 270, 271, 274, 276, 280, 282, 284, 285, 289, 292, 293, 295, 296, 299, 344, 349, 350