THÈSE

pour l'obtention du Grade de
DOCTEUR DE L'UNIVERSITE DE POITIERS
(Faculté des Sciences Fondamentales et Appliquées)
(Diplôme National - Arrêté du 7 août 2006)

École Doctorale: Sciences et Ingénierie pour l'Information, Mathématiques (S2IM)

Secteur de recherche : Traitement du Signal et des images

Defended by:

Syntyche GBEHOUNOU

****

Image databases indexing: Emotional impact assessing

****

Supervisor: Christine FERNANDEZ-MALOIGNE

Co-supervisor: François LECELLIER

****

Jury

Pr.	Ludovic MACAIRE, LAGIS, Université de Lille,	Reviewer
Pr.	Denis PELLERIN, GIPSA-lab, Polytech'Grenoble,	Reviewer
Pr.	Theo GEVERS, Université d'Amsterdam, Pays-Bas,	Examiner
MCF.	Emmanuel DELLANDRÉA, LIRIS, École Centrale de Lyon,	Examiner
Pr.	Christine FERNANDEZ-MALOIGNE, XLIM-SIC, Université de Poitiers,	Supervisor
MCF.	François LECELLIER, XLIM-SIC, Université de Poitiers,	Co-supervior

Contents

List of Figures

List of Tables

List of Tables

Notations and acronyms

Table 1: List of used notations.

Notation	Meaning
I	Greyscale image
(x,y)	Pixel coordinates in 2D
det(M)	Determinant of matrix M
trace(M)	Trace of matrix M
Ω	An image region
$\mathcal{D}$	A set of local descriptors
K	Size of the visual codebook
$\mathcal{W}$	Visual codebook
$d_{A,B}^{L2}$	Euclidian distance between two vectors A and B
$d_{A,B}^{Hamming}$	Hamming distance between two binary sequences A and B
$d_{A,B}^{\chi^2}$	χ^2 distance between two vectors A and B

Table 2: List of used acronyms.

Acronym	Meaning
AI	Artificial Intelligence
BoVW	Bag of Visuals Words
CBIR	Content Based Image Retrival
CM	Colour Moments
CMI	Colour Moment Invariants
DoG	Difference of Gaussians
EEG	Electroencephalography
EM	Expectation-Maximisation
FAST	Features From Accelerated Segment Test
FV	Fisher Vector (Vecteur de Fisher)
GLOH	Gradient Location and Orientation Histogram
GMM	Gaussian Mixture Model
IAPS	International Affective Picture System
IG	Information Gain
HVS	Human Visual System
KNN	K Near Neighbours
KP	Keypoint
LoG	Laplacian of Gaussian
MSER	Maximally Stable Extremal Regions
MV	Majoriting Voting
OpSIFT	Opponent-SIFT
PCA	Principal Component Analysis
SENSE	Studies of Emotions on Natural image DatabaSE
SIFT	Scale-Invariant Feature Transform
SSVEP	Steady-State Visually Evoked Potential
SURF	Speeded Up Robust Feature
SUSAN	Smallest Univalue Segment Assimiliating Nucleus
SVM	Support Vector Machine
tf-idf	term frequency-inverse document frequency
UKB	University of Kentucky Benchmark (Base d'images)
VLAD	Vector of Loccally Aggregated Descriptors
WA	Wave Atoms
WA4	Scale 4 of Wave Atoms
WA5	Scale 5 of Wave Atoms

Introduction

Turing [Turing 50] was the first in 1950 to introduce the idea of creating intelligent machines. Since lots of researches were done and the results are really promising. For example, Artificial Intelligence (AI) was used during the Gulf war to improve the decision support systems and the autonomous systems like the drones [His]. But one event in particular made a strong impression: in 1996 the "super computer" Deep Blue won face to Garry Kasparov, chess game world champion. So we discovered that AI can be more powerful in certain domain. From that time several works have been conducted and the more intelligent solutions are proposed in different fields: pattern recognition, computer vision, ... Despite the different improvements, some aspects resist: in particular the different aspects of our vision system which are combined to cognition. In this thesis we focus ourselves on one of them: emotional impact recognition.

Emotion recognition is an ambitious task. In fact, emotions are complex reactions and the literature is diversified from the emotion definition to the proposed approaches through the image sets. A consensual definition was proposed in 2003 by David Sander [Sander 13]. He defines emotion as a rapid phenomenon triggered by an event. The challenge of the emotional impact recognition solution is to find the discriminative image features for this task. We can divided the papers on emotion recognition into two categories. On the one hand those related to relations between emotions and colours. In these works emotions related to one colour [Beresniak 90, Boyatziz 93, Kaya 04, Ou 04a, Ou 04b, Ou 04c] and a set of colours (two and more) [Ou 06, Solli 09, Ou 11] were studied. On the other hand the researches on emotion recognition based on:

- Face detection [Tomkims 62, Scherer 84, Ekman 92, De Silva 97, Busso 04] associating an emotion to facial features (eyebrows, lips among other);
- Semantic description of colours [Wang 05, Hong 06];
- Image low level features (colours, texture, shapes, ...) [Yanulevskaya 08, Solli 10, Machajdik 10, Liu 11].

On this document we consider the approaches based on feature extraction. Note that the discriminative characteristics can be linked to the database. For example, Machajdik and Hanbury [Machajdik 10] show that colours are useful for emotion recognition on a database composed of abstract images. We chose colour, texture,

shape and object features often used in Content Based Image Retrieval (CBIR). Our research on emotional impact retrieval was inspired by this computer vision task and explains the first part of this rapport.

Content Based Image Retrieval

The idea is to retrieve images based on a set of criteria that can be overall colour, texture or objects. The approaches must be robust to geometric transforms, the change of point of view and lighting conditions. Images are often described with a set of features which can be local [Harris 88, Lowe 99, Mikolajczyk 01, Matas 02] or global [Swain 91, Oliva 01, Hays 07, Li 08, Douze 09]. Then these features can be matched or used to build visual signature [Sivic 03, Perronnin 07, Jégou 10] during retrieval step. In this latter case, two images are visually neighbours if their signatures are. The literature is diversified and new solutions are frequently proposed about descriptors, retrieval method to fit large database requirements.
Our first contribution concerns one of the most used visual signature: "Bag og Visual Word" first introduced by Sivic and Zisserman [Sivic 03]. The main idea is to characterize an image with a vector of visual word frequencies. Traditionnally the visual dictionary is obtained with a *K-Means* algorithm; sensitive to feature dimension [Parsons 04]. So we propose to build a visual vocabulary based on a random selection combined to an iterative process. This approach is independent of the descriptor dimensionality. For these works we used University of Kentucky Benchmark [Nistér 06] and Pascal VOC2012 [Everingham 12]. We chose the last one is used to build the visual codebook and the first one to test our approach.

Image emotional impact recognition

Our approach is based on CBIR techniques. We supposed that the existing descriptors can be useful and powerful for emotion recognition. In this thesis, we have assessed their relevance for the task by comparing our results to those of the literature. We also proposed a new image database for emotional impact study based on the weakness evoked by Machajdik and Hanbury [Machajdik 10]. One of the recurrent problems is the database evaluation. The majority of the authors does not publish their database or neither gives information about their evaluation conditions. However one image set appear as a consensus for different proposed approach comparison: International Affective Picture System (IAPS) [Lang 08]. It is widely assessed database which presents some restrictive terms of use. They are not compatible with our aim to study visual saliency impact for emotional impact evaluation and recognition. We think that this selective process used by our visual system can be useful to reduce the semantic interpretation during evaluation. It can also be interesting to define region of interest for feature extraction.

Outline

This document is divided into two parts: the first one focused on CBIR and the second one is related to the image emotional impact.
In Chapter 1, we present some solutions of the literature for CBIR and we finish with

a brief presentation of visual saliency. This state of the art is focused on the solutions which have inspired our research. Our new approach for codebook computation is explained and discussed in Chapter 2. We also evaluated the local features according to their visual saliency in this chapter.
In the second part, Chapter 3 is dedicated to the state of the art of the emotion recognition. In the last chapter, we propose a set of criteria to describe and compare the databases for emotion study. Our database is also presented together with the results of our computational solution based on CBIR techniques.

Part I

Content Based Image Retrieval

Chapter 1

Proposed solutions for CBIR

Contents

A CBIR task is generally composed of two steps:

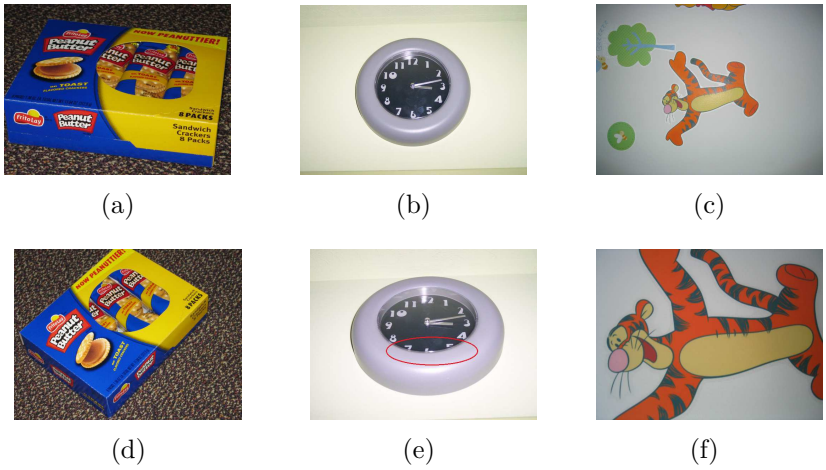
1. Transform the image to a matrix: this matrix corresponds to a set of values supposed to be robust to geometric transforms and changes of point of view or lighting conditions;
2. Compare the matrix representations of the images.

In this chapter we will present some solutions of the literature about these two steps.

1.1 Image descriptors

Ideally, the descriptors must be robust to different changes such as:

- Geometric transforms (rotation, translation, etc.);
- Point of view changes;
- Scale changes;
- Lighting condition changes.



(a) (b) (c)

(d) (e) (f)

Figure 1.1: Some images presenting different geometric, point of view and lighting condition changes. They are from the database proposed by Nistér et al. [Nistér 06]. On image 1.1(e), an occlusion due to the point of view change is shown.

Some of these changes are represented on Figure 1.1. The robustness to point of view or lighting condition changes is complex to obtain. In fact, point of view change can induce an occlusion (an incomplete scene/object) as shown by image 1.1(e). The cropping illustrated by the image 1.1(f) can be a result of scale or point of view change.
The complexity of these variations must be included in the definition of the different image descriptors used for retrieval.

Two solutions exist in the literature for image description for CBIR:

- **Global descriptors** are related to global information often colours and textures;
- **Local descriptors** which describe the local variation of a pixel or a region. They can be related to colour, texture, geometry or a combination of these low level informations.

1.1.1 Global descriptors

Different solutions have been proposed. The most known is the usage of colour histograms introduced by Swain and Ballard in 1991 [Swain 91]. Their results are really interesting mainly when colour are discriminative for the considered images.
Another global descriptor with good results in the literature [Hays 07, Li 08, Douze 09] is GIST introduced by Oliva and Torralba in 2001 [Oliva 01]. These descriptors are obtained with a set of perceptual dimensions (naturalness, openness, roughness, expansion, ruggedness) that represent the dominant spatial structure of a scene estimated using spectral and coarsely localized information.
Spite of their interesting results, local descriptors are prefered to global ones. In fact these latter do not allow to distinguish the background of the object and have some trouble face to cropping and occlusion.

1.1.2 Local descriptors

Local description includes a first step of feature detection. The goal is to capture interesting details in order to improve the robustness to the different changes. We will not evoke dense detection [Perronnin 08, Gordoa 12] which consist in select local features ignoring every geometric aspects, considering a regular grid.

Local feature detection

Local feature detectors are widely used in literature as the first step of many systems in image processing (image retrieval, image recognition ...) [Mikolajczyk 05b, Schmid 00, Bay 06, Abdel-Hakim 06, Mikolajczyk 01, van de Sande 10]. They can be divided in three groups [Tuytelaars 08]:

- Corner detectors which define a corner as a point in 2D image with high curvature;
- Blob detectors producing coherent sets of pixels having constant properties. All pixels of a blob can be considered similar to each other;
- Region detectors which directly or indirectly are concerned with images regions extraction.

In this document we prefer the term "local features" to englobe the outputs of the different detectors (points, regions, blobs).

We describe here four corner and blob detectors:

- **Harris detector** which is a corner detector proposed by Harris and Stephen in 1988 [Harris 88]. It is based on the auto-correlation matrix used by Moravec in 1977 [Moravec 77]. It measures the intensity differences between a main window and windows shifted in different directions. Harris and Stephen in their improved version proposed to use the matrice M_{Harris} defined by the equation (1.1).

$$M_{Harris}(x,y) = \begin{bmatrix} \sum\limits_{W} I_x(x_k,y_k)^2 & \sum\limits_{W} I_x(x_k,y_k)I_y(x_k,y_k) \\ \sum\limits_{W} I_x(x_k,y_k)I_y(x_k,y_k) & \sum\limits_{W} I_y(x_k,y_k)^2 \end{bmatrix}, \tag{1.1}$$

  where I_x and I_y are the partial derivatives.

  Corners are the points with a high value C defined with the equation (1.2). Harris detector is invariant to the rotation and not to the scale change.

$$C = det(M_{Harris}) - k * trace(M_{Harris})^2 \tag{1.2}$$

- **Harris-Laplace detector** which resolves the scale invariance problem of the Harris detector. It was introduced by Mikolajczyk and Schmid [Mikolajczyk 01]. They introduce the scale invariance by detecting the points firstly with a Harris function in multiple scales. Then the points are filtered according to the local measure. They use the laplacian and only points with a maximal response are considered in the scale-space.

- **Difference of Gaussians (DOG)** was used by Lowe in the Scale-Invariant Feature Transform (SIFT) algorithm [Lowe 99]. This detector approximates the Laplacian of Gaussian whose kernel is particularly stable in scale-space [Mikolajczyk 02]. The local maxima allow to detect a blob structures.

- **Features From Accelerated Segment Test (FAST)** introduced by Rosten and Drummond [Rosten 05, Rosten 06] for the real-time frame-rate applications. It is a high speed feature detector based on the SUSAN (Smallest Univalue Segment Assimiliating Nucleus) detector [Smith 97]. For each pixel, a circular neighbourhood with fixed radius is defined. The central pixel p is called "nucleus". All the pixels inside the disc whose intensity is close to the nucleus value with some threshold receive a high weighting. The pixels in the image whose value corresponds to a local minimum are considered as local features. In the case of FAST detector, only the 16 neighbours as shown in the Figure 1.2 defined on the circle are handled. p is a local feature if at least 12 contigus neighbours have a intensity inferior to its value and some threshold.

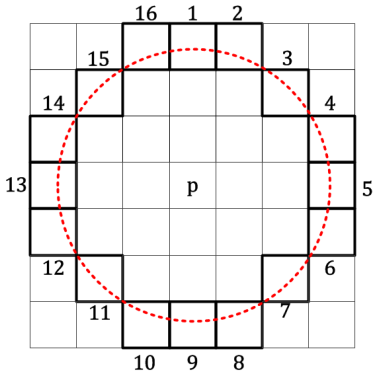


Figure 1.2: Neighbour definition for FAST detector.

Harris-Laplace detector has the best repeatability of the four mentioned aboved according to Mikolajczyk and Schmid [Mikolajczyk 01] for planar scenes. FAST confirms its high-speed detection because, compared to Harris detector, its computation is fifteen times faster [Rosten 05]. Despite this advantage for applications with time constraints, this detector is much less robust to the noise compare to DOG.

Some local feature descriptors

The most popular descriptor in object recognition is the SIFT descriptor proposed by Lowe in 1999 [Lowe 99]. The efficiency of SIFT and its different extensions have been demonstrated in numerous papers in object recognition and image retrieval [Jégou 10, Ke 04, Lowe 99, Lowe 04, Nistér 06, van de Sande 10].
The original version of SIFT [Lowe 99] is defined in greyscale and different colour variants of SIFT have been proposed. For example, OpponentSIFT, proposed by van de Sande et al. [van de Sande 10] which are recommended when no prior knowledge about the data set is available. OpponentSIFT describes all the channels in

the opponent colour space (equation (1.3)) using SIFT descriptors.

$$\begin{pmatrix} 0_1 \\ 0_2 \\ 0_3 \end{pmatrix} = \begin{pmatrix} \frac{R-G}{\sqrt{2}} \\ \frac{R+G-2B}{\sqrt{6}} \\ \frac{R+G+B}{\sqrt{3}} \end{pmatrix} \tag{1.3}$$

The information in the O_3 channel is the intensity information, while the other channels describe the colour information in the image.

The greyscale SIFT is 128-dimensional whereas the colour versions of SIFT, e.g. OpponentSIFT, are 384-dimensional. This high dimensionality induces slow implementations for matching the different feature vectors. The first solution to resolve this aspect proposed by Ke and Sukthankar is PCA-SIFT [Ke 04] with 36 dimensions. This variant is faster for matching, but less distinctive than SIFT according to the comparative study of Mikolajczyk and Schmid [Mikolajczyk 05a]. During this study, they proposed GLOH (Gradient Location and Orientation Histogram) a new variant of SIFT, which is more effective than SIFT with the same number of dimensions. However GLOH is more computationally expensive. Again with the aim to propose a solution less computationally expensive and more precise, Bay et al. [Bay 06] introduced SURF (Speeded Up Robust Features) which is a new detector-descriptor scheme. In fact, the detector used in SURF scheme is based on Hessian matrix and applied to integral images to make it fast. Their average rate of the recognition of objects of art in a museum shows that SURF is better than GLOH, SIFT and PCA-SIFT.
There are other interesting local point descriptors such as:

- Colour moments: they are measures that can be used to differentiate images based on their colour features. Once computed, these moments provide a measurement for colour similarity between images. They are based on generalized colour moments [Mindru 04] and are 30-dimensional. Given a colour image represented by a function I with RGB triplets, for image position (x, y), the generalized colour moments are defined by the equation (1.4).

$$M_{pq}^{abc} = \iint x^p y^q \left[I_R(x,y)\right]^a \left[I_G(x,y)\right]^b \left[I_B(x,y)\right]^c \, \mathrm{d}x\mathrm{d}y \tag{1.4}$$

  M_{pq}^{abc} is referred to as a generalized colour moments of order $p+q$ and degree $a+b+c$. Only generalized colour moments up to the first order and the second degree are considered, thus the resulting invariants are functions of the generalized colour moments M_{00}^{abc}, M_{10}^{abc} and M_{01}^{abc}, with:

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

- Colour moment invariants computed from the algorithm proposed by Mindru et al. [Mindru 04]. The authors use generalised colour moments for the construction of combined invariants to the affine transform of coordinates and contrast changes. There are 24 basis invariants involving generalized colour moments in all 3 colour bands.

1.2 Techniques for similar image retrieval

In the literature two techniques exist:

- Compare the vector of descriptors with a matching solution;
- Buid a visual signature per image and compare these signatures.

We are focus ourselves on the latter solution.

1.2.1 Image representation

As far as the image representation is concerned, plenty of solutions have been proposed. The most popular is commonly called "Bag of Visual Words (BoVW)" and was inspired by the bag of words used in text categorisation. Given a visual vocabulary, the idea is to characterize an image by a vector of visual word frequencies [Sivic 03]. The visual vocabulary construction is often done through low level feature vector clustering using for instance *K-Means* [Csurka 04, Sivic 03] or Gaussian Mixture Models (GMM) [Farquhar 05, Perronnin 06]. However, a weighting value can be applied to the components of this vector. The standard weighting scheme is known as "term frequency-inverse document frequency", *tf-idf* [Sivic 03], and is computed as described by the equation (1.5).
Suppose there is a vocabulary of K words, then each document is represented by a K-vector $V_d = (t_1, \ldots, t_i \ldots, t_k)^\top$ of weighted word frequencies with components:

$$t_i = \frac{n_{id}}{n_d} \log \frac{N}{n_i}, \tag{1.5}$$

where n_{id} is the number of occurrences of word i in document d, n_d is the total number of words in the document d, n_i is the number of occurrences of term i in the dataset and N is the number of documents in the dataset.

Besides BoVW approaches, many other efficient methods exist, for example Fisher Kernel or VLAD (Vector of Locally Aggregated Descriptors). The first one has been used by Perronnin and Dance [Perronnin 07] on visual vocabularies for image categorisation. They proposed to apply Fisher kernels on visual vocabularies represented by means of a GMM. In comparison to the BoVW representation, fewer visual words are required by this more sophisticated representation.
VLAD is introduced by Jégou et al. [Jégou 10] and can be seen as a simplification of the Fisher kernel. Considering a codebook C = $c_1, \ldots, c_K$ of K visual words generated with *k-Means*, each local descriptor x is associated to its nearest visual word $c_i = NN(x)$. The idea of the VLAD descriptor is to accumulate, for each visual word c_i, the differences $(x - c_i)$ of the vectors x assigned to c_i.

1.2.2 Visual signatures comparison

There are various methods to compare image representation and a majority of them are easily available[1].

[1] https://gforge.inria.fr/projects/yael/, http://www.cs.umd.edu/~mount/ANN/, http://www.cs.ubc.ca/research/flann/

The different methods to retrieve images can be classified into two groups:

- Learning methods for image classification and categorization;
- Distances based comparison of visual signatures for image retrieval.

In the first class of methods SVM (Support Vector Machine) offers good results [van de Sande 10]. Note that a linear kernel can be utilized with Fisher vectors and this configuration gives satisfying results [Perronnin 07].
In the second class of methods, similar image retrieval are traditionnally done with Near Neighbour research with equation (1.6):

$$NN(\mathcal{S}_1) = argmin\ dist\left(\mathcal{S}_1 - \mathcal{S}_2\right), \tag{1.6}$$

with $\mathcal{S}_1$ the visual signature which is considered as the request, $\mathcal{S}_2$ one visual signature of the database and *dist* the distance between the compared image representations. Often euclidian distance is used to compare the image representation but χ^2 distance can be considered. The equation (1.7) gives the computation for this distance between to visual signatures of size K:

$$d^{\chi^2}_{S_1,S_2} = \sum_{i=1}^{K} \frac{\left(S_1(i) - S_2(i)\right)^2}{S_1(i) + S_2(i)}. \tag{1.7}$$

We have presented some "classical" techniques for image retrieval. In fact, new solutions considering human visual system information emerge. They used for example visual saliency information. We are interested by this information in our researches so you decide to briefly present this concept.

1.3 Visual saliency

Visual attention models are used to identify the most salient locations in an image. They are widely applied to many image-related research domains.
In the last decades, many visual saliency frameworks have been published. Borji et al. [Borji 13] have proposed an interesting comparative study of 35 different models of the literature. They also mentionned the ambiguity between saliency and attention. According to them, visual attention is a broad concept covering many topics (e.g., bottom-up/top-down, overt/covert, spatial/spatio-temporal). On the other hand it has been mainly referring to bottom-up processes that render certain image regions more conspicuous; for instance, image regions with different features from their surroundings (e.g., a single red dot among several blue dots).
Many of visual saliency frameworks published are inspired from psycho-visual features [Itti 98, Le Meur 06a] while others make use of several low level features in different ways [Gao 08a, Zhang 08]. The works of Itti et al. [Itti 98] can be considered as a noticeable example of the bio-inspired models. An input image is processed by the extraction of three conspicuity maps based on low level characteristics computation. These three conspicuity maps are representative of the three main human perceptual channels: colour, intensity and orientation. These maps are combined to generate the final saliency map as described on Figure 4.6.

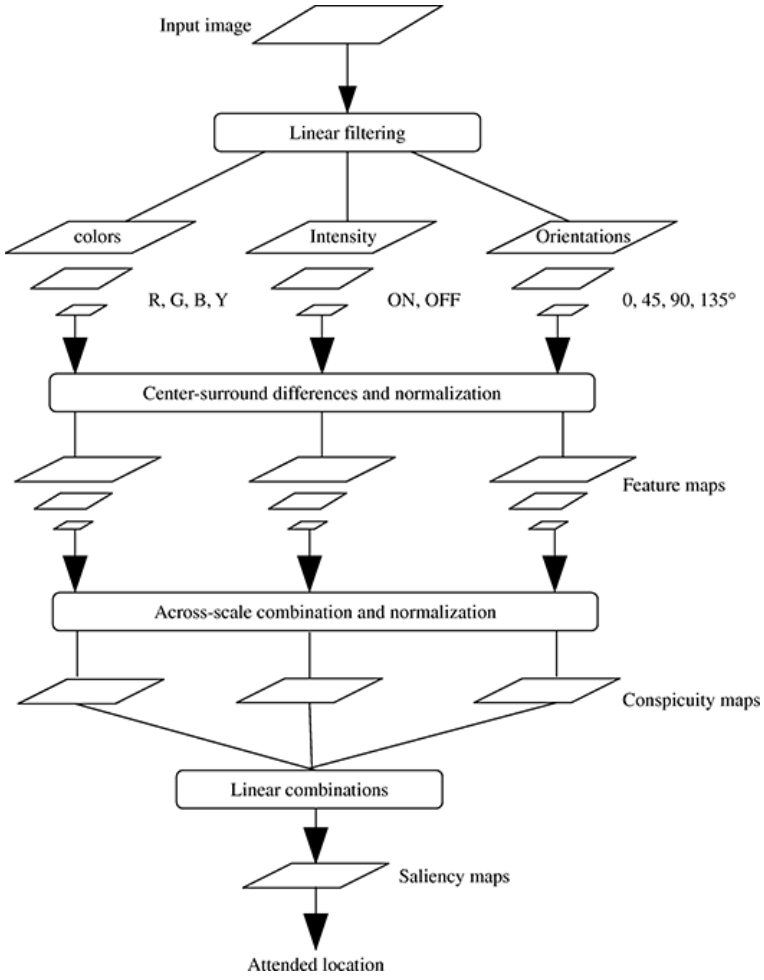


Figure 1.3: Architecture of the computational model of attention proposed by Itti et al. [Itti 98].

For image retrieval including saliency usage, the idea consists in taking advantages of visual saliency to decrease the amount of information to be processed [Gao 08a, Gao 08b, Liu 08, Zdziarski 12]. These methods usually take the information given by the visual attention model at an early stage; image information will be either discarded or picked as inputs for next stages based on its saliency value. For example Gao et al. [Gao 08a] propose to rank all the local features according to the saliency value and only the distinctive points are reserved for the matching stage. Zdziarski et al. [Zdziarski 12] share the same idea: SURF descriptor are computed only for pixels with saliency value above a fixed threshold. Their experiments show that the number of features can be reduced without affecting the performance of the classifier.

In this chapter, we have presented some solutions for image retrieval from description to visual signature comparison. Local descriptors are preferred because of their accuracy and their robustness to lots of variations. They are often high dimensional so a new solution is to filter local features according to human vision

system using e.g. visual saliency. Then we have briefly present one of the first the visual saliency models: those proposed by Itti et al. [Itti 98].
In this first chapter we introduce the different tools we use in our research (detectors, descriptors) described in Chapter 2. In this latter, we will present a new approach to build visual codebook. We also study the saliency of the features detected by the four detectors we present in Subsection 1.1.2.

Chapter 2

Our approach for image retrieval

Contents

In this chapter, we present our first contribution for image retrieval. Firstly, we describe a new algorithm to build a visual dictionary based on a random selection of visual words. With this approach we obtain satisfying results with a codebook composed of only 294 words for the descriptor with the best results, compared to a average size of 20 000 for the literature. Secondly, we include visual saliency in our retrieval system by two ways:

- Weight the descriptor vectors by the visual saliency of the local features before BoVW computation;
- Study the saliency of the detected local features and the importance of these features for the retrieval according to their saliency.

2.1 Image databases used

There are diversified database for image retrieval and classification [Nistér 06, Everingham 07, Huiskes 08, Huiskes 10, Everingham 12]. We chose two datasets during our experiments for image retrieval:

- **University of Kentucky Benchmark** introduced by Nistér et al [Nistér 06]. In the remainder, we will refer to this dataset as "UKB" to simplify the reading. UKB is really interesting for image retrieval and presents three main advantages:

  1. It is a large benchmark composed of 10 200 images grouped in sets of 4 images showing the same object. Figure 2.1 is an good illustration of the diversity of the set of 4 images (changes of point of view, illumination, rotation, etc.);
  2. It is easily accessible and a lot of results are available to make an effective comparison. In our case, we will compare our results to those obtained by Jégou et al. [Jégou 10] and Nistér et al. [Nistér 06];
  3. The evaluation score of the results on "UKB" is simple: it counts the average number of relevant images (including the query itself) that are ranked in the first four nearest neighbours when searching the 10 200 images. A score of 3 indicates that over the whole dataset, the system retrieves a mean of 3 images over the 4 existing for the same object.

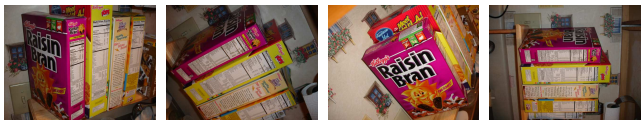


Figure 2.1: An example set of 4 images showing the same object in UKB.

- **PASCAL Visual Object Classes challenge 2012** [Everingham 12] called PASCAL VOC2012. This benchmark is composed of 17 215 images represent realistic scenes and they are categorised in 20 objects classes, e.g. person, bird, airplane, bottle, chair and dining table. We used the full dataset to construct the vocabulary.

2.2 The chosen descriptors

We chose:

- CM (Colour Moments);
- CMI (Colour Moment Invariants);
- SIFT (Scale-Invariant Feature Transform);

- SURF (Speeded Up Robust Feature);
- Opponent-SIFT referred as OpSIFT for the result presentation.

SURF excepted the descriptors were computed with ColorDescriptor Software developped by[van de Sande 10] using Harris-Laplace detector. For this latter we use k=0.6, the Harris threshold is set to 10^{-9} and harris threshold 0.03. SURF are computed with Opencv and the hessian threshold was fixed to 300 for local feature detection.

2.3 New method for visual codebook computation: Iterative Random visual words Selection (IteRaSel)

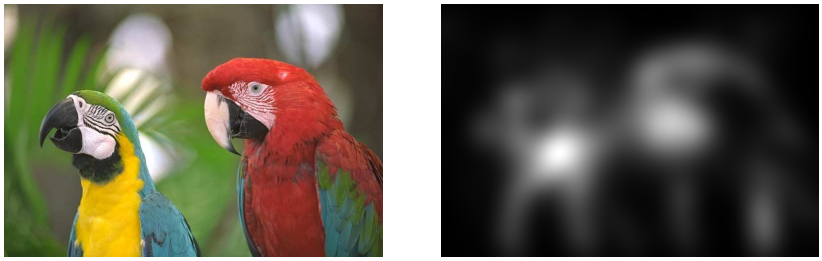
Most of the time, the BoVW uses the well-known *K-Means* algorithm [Csurka 04]. But the effectiveness of these clustering algorithm tends to drop drastically with respect to the high dimension of the image features. As Parsons et al. [Parsons 04] mentioned, the curse of dimensionality occurs and makes the result of the clustering close to random. This randomness aspect has been one of the starting point of our reflexion. A second aspect of our reflexion was to minimize the importance of feature selections. Without a priori knowledge of the image dataset, our approach is designed to simplify the indexing step and the parameter tuning according to selected features. Thus, one of our objectives is to simplify the construction of the visual vocabulary. We propose to replace the clustering algorithm step by selecting randomly a large set of visual words. This random selection can either be made by:

- Randomly picking keypoints in all images of a collection and using their computed descriptors as visual words;
- Creating synthetic visual words with knowledge of the feature space distribution.

We chose the first solution for visual words selection. After selecting the visual word the second step is to identify the visual words that have the best information gain IG in the set of random selected visual words. To do so, we defined a information gain IG_w based on *tf-idf* weighting scheme and a saliency score, given by the following formula:

$$IG_w = \underbrace{\frac{n_{wD}}{n_D} \log \frac{N}{n_w}}_{tf-idf} + \underbrace{\frac{\sum Sal_{wD}}{n_{wD}}}_{Visualsaliency} \tag{2.1}$$

where IG_w is the IG value of the visual word w, n_{wD} is the frequency of w over all the keypoints of the dataset D, n_D the total number of keypoints in the dataset, N the number of images in the dataset, n_w the number of images containing the word w, and Sal_{wD} is the saliency score for all the keypoints assigned to word w.
The part of our IG score comes from saliency maps obtained with Itti et al.' model



(a) Image (b) Saliency map

Figure 2.2: An example of saliency map.

[Itti 98] computed using the graph-based visual saliency software[1]. Note that the different saliency values obtained are normalized between 0 and 1: 0 for a non salient local feature and 1 for salient pixel.
Algorithm 1 describes our aproach for vocabulary computation.

Algorithm 1: Visual vocabulary construction

Data: $\mathcal{D}$ the set of descriptors on the whole database, K the size of the codebook
Result: $\mathcal{W}$ the final visual vocabulary
Initialize $\mathcal{W}$ with of randomly selected words;
while *Size of* $\mathcal{W} > K$ **do**
  Assign each keypoint to its closest word w;
  Compute IG_w for each word w with equation (2.1);
  Sort words according to their IG_w value;
end

The sorting step evoked in Algorithm 1 is important to deleted non informative features. In fact all words with an IG value equals to 0 are removed. The features with the highest information gain are also deleted according to a threshold α fixed to 10% after a lot of experimentations. This allows to remove the "sporadic" features which can decrease the retrieval results because there are in many images background. The *while* block code is repeated until the desired number of words in the visual vocabulary is reached, re-assigning only the keypoints that were left alone due to the previously deleted visual words. At the end of this step, the visual vocabulary is built.

The proposed algorithm is based on a random set of visual words. When using a random approach, it is natural to assume that several runs of the algorithm would create different visual vocabularies, and therefore would yield different experimental results. In order to avoid a hypothetical lack of stability that may appear for some descriptors, we propose a stability extension to our approach, inspired from strong clustering methods. For this purpose, we generate β visual vocabulary, the iterative process is executed for each of them (see Algorithm 1). We combine the obtained

[1] `http://www.klab.caltech.edu/~harel/share/gbvs.php`

words together and run the iterative process again, until the desired number of visual words is reached. Combining few sets of visual words will improve the overall informative gain of the vocabulary. However, since the already obtained results are almost stable, experiment results show that $\beta = 3$ is a good parameter value, and choosing $\beta > 3$ would give similar results for a longer construction time. We will present in the next section the results about the different β value test.

With this new codebook algorithm we used BoVW representation for image visual signature. The distance between the descriptor vectors and the visual words were evaluted with euclidian distance and those between visual signatures with χ^2 distance.

2.4 IteRaSel evaluations

We used Pascal VOC2012 to compute the visual codebook and the testing database is UKB. For easier reading we propose the following notations for the presentation of the results:

- BoVW corresponds to an image representation obtained with a *K-Means* codebook and a BoVW visual signature;
- IteRaSel corresponds to an image representation obtained with a *IteRaSel* codebook and a BoVW visual signature.

2.4.1 Random selection of visual words study

Firstly we compare a random selection of visual words to *K-Means* codebook. In the two configurations the visual signature used is BoVW. The results are presented in Figure 2.3. Only three descriptors are illustrated, the two others have the same behaviour. The overall tendency is that *K-Means* based BoVW approach outperforms simple random selection of words for small vocabularies. However, when the number of randomly selected words is high enough, the figure shows a stable score value for random selection, that is equivalent to *K-Means* scores. Thus, by simply taking a high number of random visual words, the scores are equivalent.

2.4.2 Random visual word selection combined to iterative process

The results presented here is for CMI descriptor. In fact, it offers the best retrieval rate as Figure 2.3 shows. From the results illustrated on this same figure we have concluded that random visual word selection gives similar results compared to *K-Means*. In this subsection we study the impact of the iterative process included in the Algorithm 1.
Figure 2.4 presents the UKB scores with respect to a sub-selection of the 2048 starting visual words. These exhaustive results, made from different runs, highlight important facts. First, the results are very stable. Indeed, starting from a random

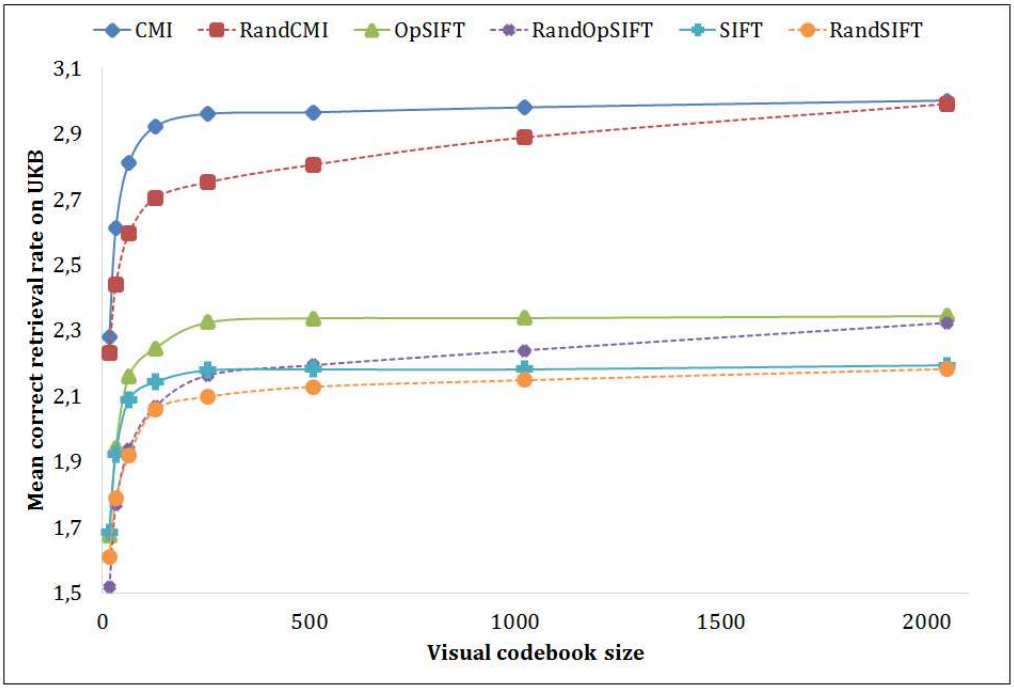


Figure 2.3: *K-Means* based BoVW vs Random selection of words based BoVW. CMI corresponds to the results obtained with *K-Means* and randCMI those obtained with a random selection. It is the same for other descriptors.

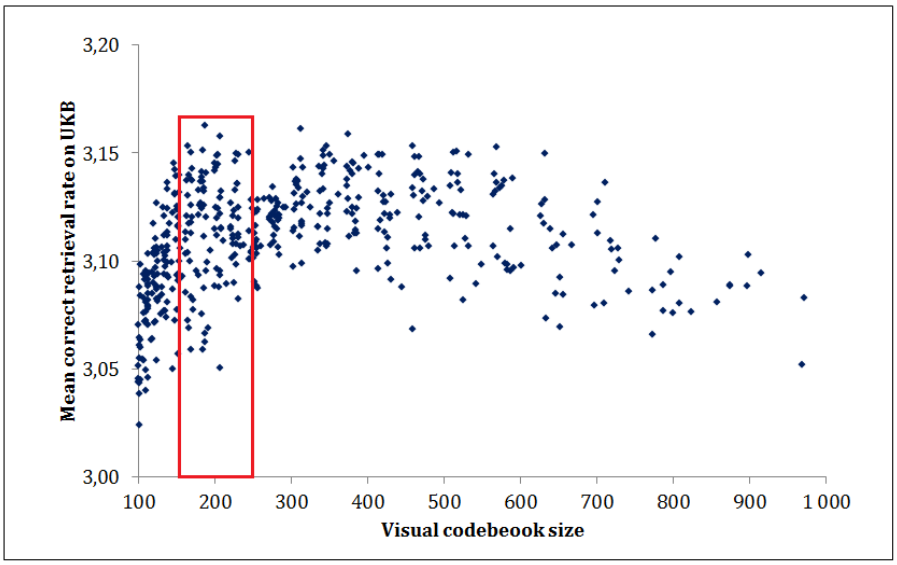


Figure 2.4: Recursive selection scores for 2048 random starting words. Between 150 and 200 the average retrieval rate are consistently between 3.05 and 3.15.

vocabulary, we see that for a given number of words, the score lies within a quite narrow window. For example, around a sub-selection of 150 words, the score value ranges from 3.05 and 3.15 approximately. The second important fact is the high

score values of these sub-selections. In the previous results, neither normal BoVW nor the random selection of vocabulary has given a score over 3.
We also studied the impact of the size of the initial randoml codebook. To do so we choose the following sizes of initial codebook: {1024, 2048, 4096, 8192, 16 384, 32 768, 65 536}.

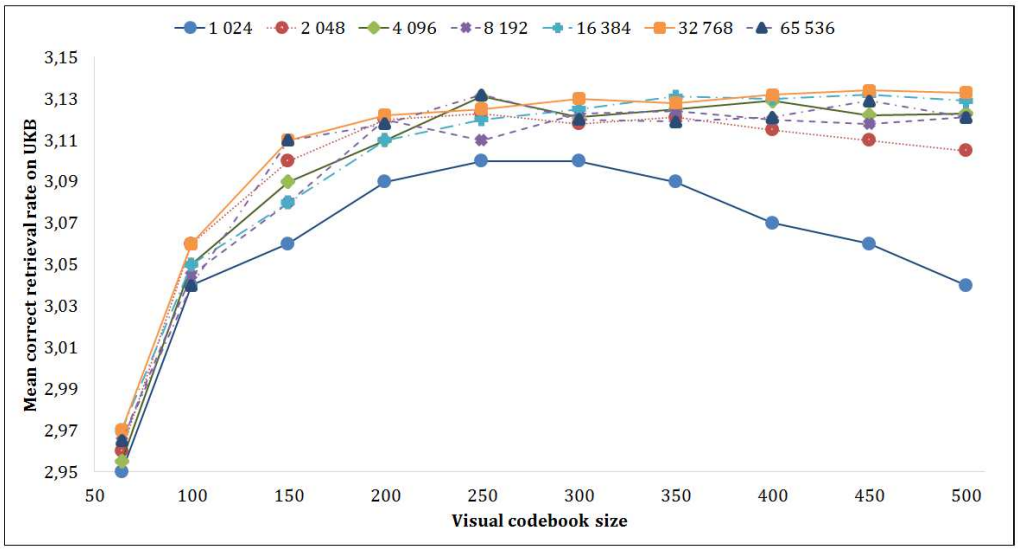


Figure 2.5: Recursive selection mean scores for 1024 to 65536 random starting words.

Figure 2.5 presents the mean UKB score of 10 runs with respect to number of visual words using different starting random sets, from 1024 to 65538 words. There is a clear evolution of the results. The starting number of visual words affects the overall results. Starting with 1024 words is not enough, however going higher than 4096 words has no more clear effect on results. For the next experiments, we select 4096 as the maximum value of initial random visual words.

2.4.3 Stabilization process by mixing β codebooks

Figure 2.6 shows the interest of mixing vocabularies together: regardless of the size of the initial visual vocabulary words, scores increase with β, including $\beta = 2$ to $\beta = 3$. For $\beta > 3$ the UKB score does not improve significantly so we chose $\beta = 3$ as a optimum value.

2.4.4 IteRaSel evaluation with the codebook mixing process

The visual codebooks used for the results presented here are obtained by this way:

1. Ramdom selection of a subset of 4096 visual words in the set $\mathcal{D}$ of the descriptors of the images of Pascal VOC2012;

2. Computation of a codebook from the Algorithm 1;

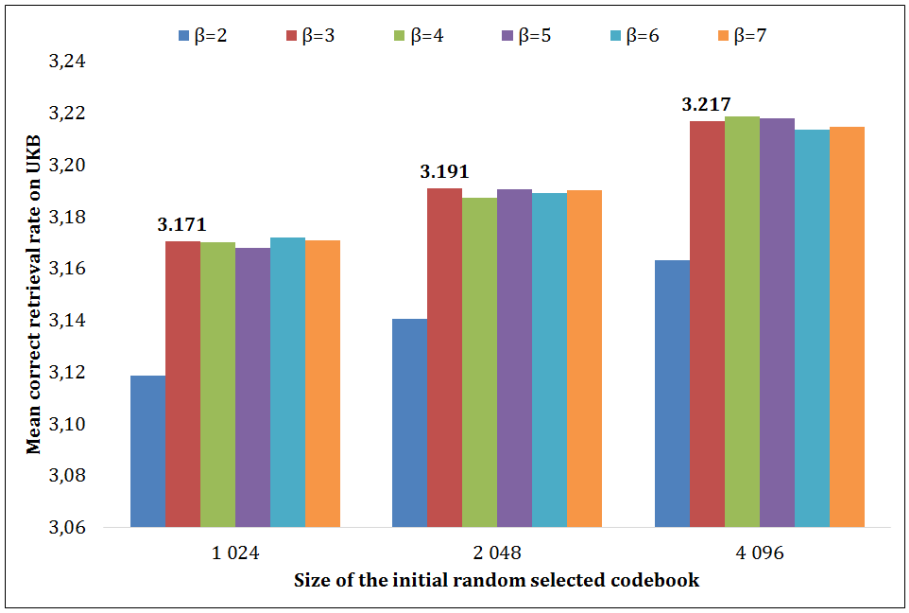


Figure 2.6: Mean score after mixing dictionaries in multiple configurations: $\beta = \{2, \ldots, 6, 7\}$. The codebooks were computed with 1024, 2048 et 4196 visual words randomly selected.

3. The steps 1 and 2 are repeted 2 times to obtain 3 codebooks;
4. The 3 codebooks are used to compute a unique codebook with the Algorithm 1; in this case the random initialization of $\mathcal{W}$ is skipped;
5. The bags of visual words are computed with the final dictionary;
6. The visual signatures are compared with χ^2 distance.

The average correct retrieval rates obtained with the five chosen descriptors are summarized in Table 2.1. The results show that for most of selected features, there

Table 2.1: Average correct retrieval rates for UKB. *K-Means* and IteRaSel correspond to the algorithm used to compute the visual dictionary.

Descriptors	*K-Means*	IteRaSel	%(IteRaSel/*K-Means*)
CMI	2.95 (K=2048)	**3.22** (K=294)	+7.4%
CM	2.62 (K=2048)	**2.81** (K=265)	+7%
SURF	2.69 (K=2048)	2.75 (K=253)	+2.75%
OpSIFT	2.30 (K=2048)	2.46 (K=159)	+6.9%
SIFT	2.19 (K=2048)	2.30 (K=187)	+6.5%

is a significative improvement, $\sim 7\%$ with IteRaSel in the UKB score compared to the BoVW approach. This demonstrates that without adding any prior knowledge regarding which feature to use, and how to use it, our iterative random selection of

visual words approach performs better. These results are even more interesting than the descriptors that obtain the highest average scores have a small dimensionality: 24 for CMI (3.22) and 30 for CM (2.81).

We subsequently focused our work on the descriptor CMI which has the advantage of providing good results for very small dimensionality compared with SIFT or its colour extensions.

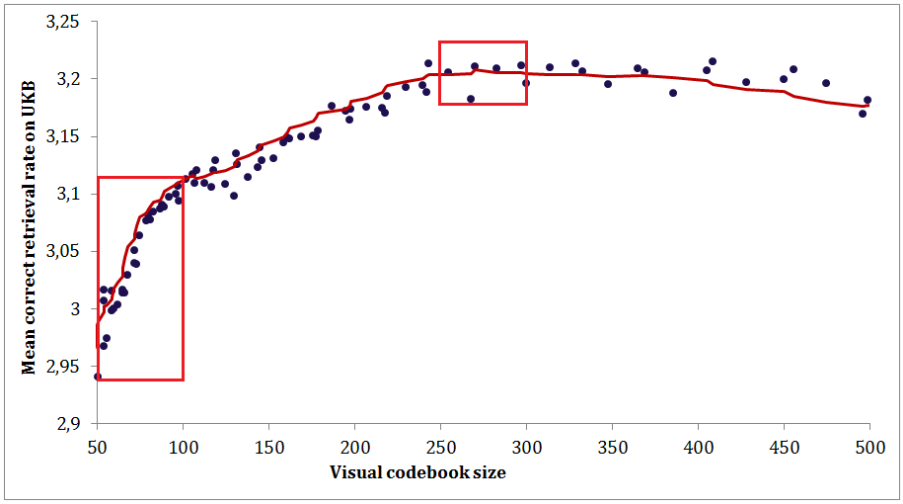


Figure 2.7: Mean correct scores obtained with different sizes of codebook from a initial random codebook of size 4096. The red curve is a trend curve of the scatterplot.

Figure 2.7 presents the results of the visual codebook size impact study. The final sizes of the dictionary are between 50 and 500 and the initial random visual codebook is 4096. At first we observe interesting results with few visual words: between 250 and 300. Secondly a mean score of 3 is noticed for codebook size between 50 and 100. They are better than those obtained with a *K-Means* vocabulary of 2 048 visual words (see Figure 2.3).

The first evaluations of the different steps of our algoritm show it validity. The method is simple but easily reproducible.

2.4.5 Comparison with the state of the art

In this comparison only CMI is used. We preferred to compare our best results with the best from the literature. For this comparison no normalization was applied on the visual signature. Traditionnally a L2 norm was used [Jégou 10, Jégou 12] but in our case we noticed a little decreasing of our mean score 3.07 with L2 norm against 3.27 without normalization. We will discuss about normalization impact later.
The mean scores presented in Table 2.2 are the best in the cited papers.
Table 2.2 analysis indicates that BoVW give the worst results spite of the size of the codebook. We can also notice that the results in this case depend on the size

Table 2.2: Comparison of our best score with some of the literature

Image representation	Best score
FV[Jégou 12] (SIFT K=256)	**3.47**
IteRaSel (CMI K=294)	**3.22**
VLAD [Jégou 10] (SIFT K=64)	3.17
BoVW* (CMI K=2048)	2.95
BoVW* [Jégou 12] (SIFT K=20 000)	2.87

* is a BoVW build with a codebook from *K-Means* algorithm.

of the visual vocabulary. FV and VLAD scores are the best but these image representation include more information than the simple frequency counting for "Bag of visual words".

The comparison with the literature results confirm the previous results about IteRaSel. This new visual codebook computation gives hopeful results on UKB especially for a small codebook. Different improvements are possible and are discussed in the next subsection.

2.4.6 Discussions

Through the different evaluations we show that IteRaSel clearly outperforms *K-Means* especially for small visual codebook. This algorithm is fully repoducible and easy to implement. Different improvements can be introduced. For example the signature normalization. As we introduced in Subsection 2.4.5, we notice a negative effect after applying L2 norm. So we decide to study the impact of this process for three distances: χ^2, L1 and L2 distances. To do this we define a new norm with the equation (2.2).

$$norm = (nb_{KP})^p, \quad p \in \{0, 0.1, 0.2, \ldots, 0.9, 1\}, \tag{2.2}$$

nb_{KP} is the number of local features for an image. Note that if p=1 it is L1 norm and no normalization for p=0.
As shows in Figure 2.8 the normalization has an important effect on the results and this regardless the distance. For the three distances when p$\in$ $[0.5, 1]$ the highest score is obtained. The normalisation is even more important on UKB because there are just 4 similar images. For example, for χ^2 distance between p=0.6 and p=0 the improvement is about 3.5%. The selected distance for image signature comparison is also significant: choose χ^2 distance enhances the results about 17.6% compared to euclidian distance.

The second part of this chapter is dedicated to visual saliency usage in image retrieval. We used for this study the visual attention model proposed Itti et al. [Itti 98]. For this second part only CMI descriptor is used.

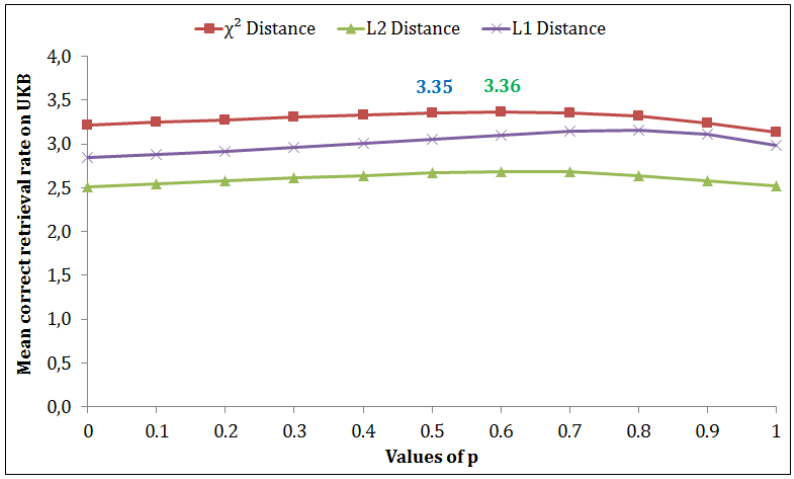


Figure 2.8: Normalization impact on image retrieval.

2.5 Weighting of the descriptor vectors by the local feature visual saliency

The first saliency usage we tested is descriptor vector weighting by the local feature visual saliency.
Considering average correct retrieval score we notice no change. The first reason can be that globally the local features detected has a similar visual saliency. So include saliency by this way has no effect. We extended this study by analysis the rank of the four similar images during the retrieval. In fact for each request all the images in the database have a rank $\in [0, 10\ 199]$.
Figure 2.9 shows the results of the study of the ranks of four similar images retrieved for UKB. For this illustration the four first image ranks were summed and we present the results for three values of the normalization parameter p defined with the equation (2.2). On this figure, when:

- The rank sommation is the same (configuration Saliency=NoSaliency); it informs that the four first images retrieved is the good ones;

- The summation is different in the two configuration then at least one of the four first retrieved is not good.

The conclusion to this study is that weighting descriptor vectors by the local feature saliency improve the rank of the four similar images.

The first usage of saliency information we have tested is not successfull regard the retrieval score for UKB. We verified the hypothetis that the local features detected have not significant saliency value by evaluating the saliency of keypoints detected by four algorithms.

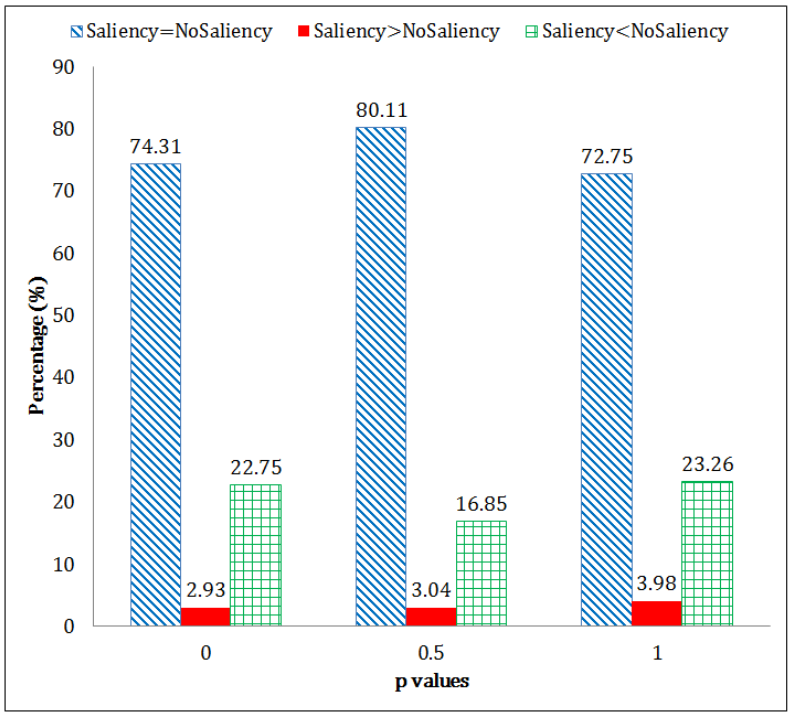


Figure 2.9: Study of the ranks of four similar images retrieved for UKB. On this figure "Saliency" means taht the descriptor vector have been weighted by the visual saliency of the lcoal features they descibe and "NoSaliency" means that no weighting has been applied.

2.6 Local features saliency study

This study was conducted for four databases:

- UKB [Nistér 06];
- Pascal VOC2012 [Everingham 12];
- The dataset proposed by Le Meur and Baccino [Le Meur 06b] for saliency study which contains 27 images. We will refer to this dataset as "LeMeur";
- The database introduced by Kootstra et al. [Kootstra 11] composed of 101 images refered as Kootstra in this paper. It is also used for saliency model evaluation.

The inclusion of the two image databases traditionally used for the study of visual saliency allows us to verify the dependence of our results to the nature of the base. We evaluate the four detectors presented in Chapter 1:

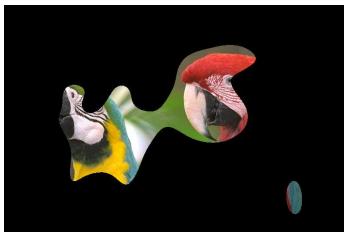
- Harris;
- Harris-Laplace;
- DoG;

- FAST.

In our experiments, we use k=0.4 for Harris detector. The Harris threshold was defined equal to 0.05 multiplied by the best corner quality C computed with the equation (1.2). The neighbourhood size is 3x3. For Harris-Laplace detector[2], we use k=0.6, the Harris threshold is set to 10^{-9} and harris threshold 0.03.
DOG detector settings are the default values for these detector proposed by Lowe[3]. The threshold needfull in the FAST algorithm to compare the intensity value of the nucleus and its neighbours is set to 30 in our experiments.

To evaluate the local features saliency we need to find a threshold t. The different saliency values obtained are normalized between 0 and 1 so an instinctive threshold might be 0.5. We preferred to define a threshold that conserves the easy recognition of the scenes/different objets with very few pixels. We evaluated different values. The different images in Figure 2.10 show the results with three values of threshold: 0.3, 0.4 and 0.5. We chose the threshold equal to 0.4. So we consider that a local feature is salient if the saliency on its position is greater than or equal to 0.4.

(a) Original image



(b) t=0.3

(c) t=0.4

(d) t=0.5

Figure 2.10: Image quantized with different threshold.

Before studying the local features saliency, we study the average saliency repartition of the pixels of the different databases. In all the figures presented in this section, the horizontal line corresponds to the median value. The values represented in red are the outliers. Their values are larger than $q3 + w(q3 - q1)$ or smaller than $q1 - w(q3 - q1)$, where $q1$ and $q3$ are the 25th and 75th percentiles, respectively. w is equal to 1.5 for our displaying and this value corresponds to approximately $\pm 2.7\sigma$ and 99.3 coverage if the data are normally distributed.
In Figure 2.11, the saliency values are uniformly quantized into 10 intervals. This quantization allows to observe the distribution of the pixels according to their

[2]We use colour descriptors software developped by van de Sande et al. [van de Sande 10].
[3]We use Opencv implementation of Harris detector and SIFT detector to compute DOG.

saliency. The "outliers" on this figure are due to the fact that some images contain more salient information than others within the same database. Their rates are low for the different databases: 1.48% for LeMeur, 1.58% for Kootstra, 2.28% for UKB and 1.99% for Pascal VOC2012. For the four datasets, the first interval $[0, 0.1]$ is the one with the highest median m: $m > 30\%$ for LeMeur, $m > 20\%$ for Kootstra, $m > 40\%$ for UKB and $m \sim 30\%$ for Pascal VOC2012. These preliminary results are consistent because a saliency model is supposed to mimic our visual attention system by selecting very little information but the most relevant. We can conclude from this first analysis that in general, the images on our databases are mostly not salient. This is confirmed with the Figure 2.12.
In this figure we can notice that LeMeur and Kootstra have the highest median values. This is understandable since they are designed for visual saliency study. UKB and Pascal VOC2012 may contain several visually attractive information of different sizes linked to the complexity of the scene or object items. Note also that the size of the attractive objects or scenes plays a significant role. Although the percentage of salient pixels is greater on LeMeur and Kootstra, we can conclude that on the four image databases very few pixels have a visual saliency value greater than or equal to 0.4. Then we can make the same assumption for the behaviour of detectors: very few salient points will be detected.
For the local feature detector saliency evaluation our aim is not to find the best configuration of the various parameters involved in the calculation of local features for better results. We took the default values provided by the authors assuming that they correspond to a certain average optimization.
We recall here that the fact that a detector produces more or less salient points is not necessarily related to its performance in image retrieval. This aspect is not considered in the any evaluatation of detectors proposed in the literature. In the Section 2.7, we studied for Harris-Laplace the importance of the salient local features for images retrieval on UKB.

The results of the study of visual saliency of the local features are shown in the Figure 2.13, page 35.
If we consider the average of the different medians m (one median per database), we obtain $m \sim 50\%$ for Harris, $m \sim 32\%$ for Harris-Laplace, $m \sim 35\%$ for DoG and $m \sim 37\%$ for FAST. Harris detector appears as the one that extracts the most salient features despite the nature of the images of these databases. It could be explained by the fact that it measures intensity differences in the image space, that can be interpret as a measure of contrast, interesting for visual saliency. The difference between the three other detectors is minimal. The results of Harris-Laplace and DoG could be explained by the scale change they incorporate.
Note that there is no correlation between the percentage of salient pixels in the images and those of salient local features detected.

Our study of local feature detector saliency confirms that they do not detect the most salient information[4]. These observations are comprehensible since the local detectors used and the visual saliency models are not based on the same concept.

[4]Those from the chosen detectors.

The fact that the Harris detector produces the most salient corners is interesting. It can advise to use Harris detector if any scale change invariant is need for local features filtration.

In the following, we focus on Harris-Laplace, and assess the importance of the local features according to their visual attention for image retrieval on UBK. The salient concept is not no more linked to the previous threshold $t = 0.4$. The local features are ranked according to their saliency value.

2.7 Impact of local feature filtering based on visual saliency

For this study the local features are detected with Harris-Laplace. We consider again only CMI. To achieve image retrieval task on UKB, we chose BoVW. The codebook used here is computed with IteRaSel algorithm and is the same we used to compare our results with literature in the first part of chapter.

As we previously mentionned we ranked the local features according to their saliency values. For our study we filtered local features in two configurations:

- "More salient": the more salient features are removed;
- "Less salient": the less salient features are removed.

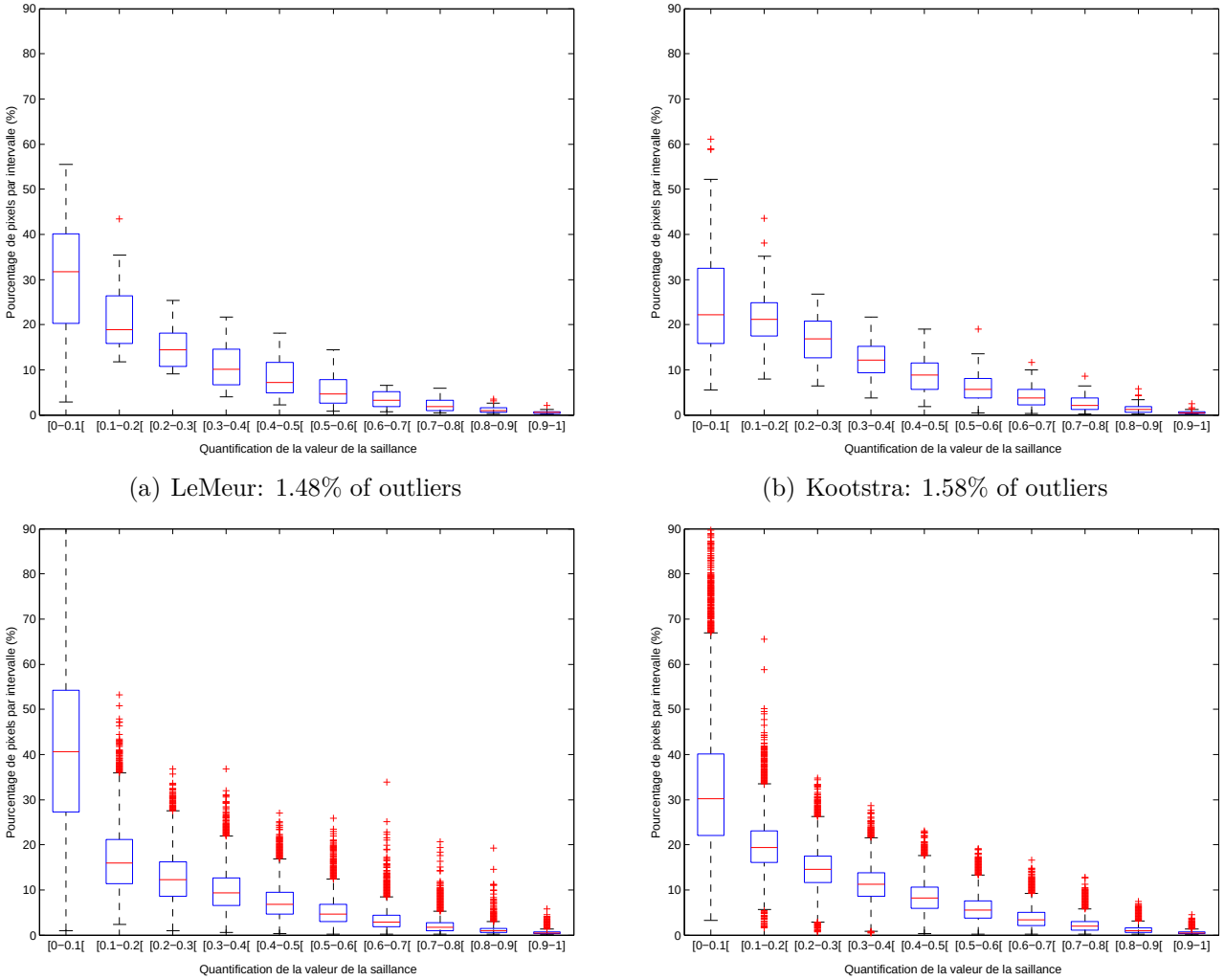
The image signature is then built with the residual local features after filtering. The results are presented in Figure 2.14, page 36 for CMI.
The results clearly highlight the importance of salient local features for the retrieval. For example removing 50% of the more salient features with CMI induces lost of correct retrieval of 20% against 3.55% for the 50% of the less salient ones.
Our findings go in the same direction as the previous: local features can be filtered according to their saliency without affecting significantly the retrieval results. The more salient local features description are very important to have an acceptable retrieval. These conclusions are valid for Harris-Laplace detector. We have tested these assumption in a different detection configuration: dense quantization. Indeed increasing works consider this feature detection approach [Perronnin 08, Gordoa 12] which poses a problem: the large number of keypoints. If the previous results are confirmed then the visual attention can be used to filter local keypoints regardless the features detector for CMI.

For our dense selection we picked a pixel on a grid of 15*15 every 6 pixels producing 8 190 local features. The results are presented in Figure 2.15, page 36.
Filter dense local features according to their visual saliency values has the same impact as previous filtering (Figure 2.14). We can conclude that using CMI, on UKB saliency filtering does not impact in a negative way the retrieval results respecting a adequate threshold.
The previous study highlight the importance of salient local features for a correct

retrieval on UKB both with Harris-Laplace detection and dense selection. So we decided to replace the less salient features by the more salient one from dense detection. The results are shown in Figure 2.16, page 37.
Replace less salient local features by the most salient ones from dense detection seems to be a good compromise to use visual saliency in order to improve the retrieval. In fact the retrieval is improved by 3.75% for 20%. Of course, this improvement is small but it shows that this operation does not degrade at all the results and tends to improve them. It confirms that, for the considered database and local descriptors visual saliency is very important to keep a very good accuracy.

Summary

In Chapter 1, we evoked certain solutions for feature detection and description. We do not insist on global description because our contributions are about local description. In the second chapter, we have presented an algorithm for visual codebook construction based on random visual word selection. Our method proves its performance compared to *K-Means* based codebook. Spite of the simplicity of our algorithm, the results are really hopeful. On UKB, the results obtain with BoVW representation are close to those obtained with VLAD and Fisher Vector in the literature. Our algorithm can be improved on several aspects for example distance used to compared visual signatures. We have shown that χ^2 distance is better than L2 distance for visual signature comparison.
The local feature detector saliency evaluation informs that very few local features detected are visually salient. However the salient local features detected with Harris-Laplace detector are important for a good retrieval. The less salient ones can be filtered in a certain proportion without affecting the retrieval results.



(a) LeMeur: 1.48% of outliers

(b) Kootstra: 1.58% of outliers

(c) UKB: 2.28% of outliers

(d) Pascal VOC2012: 1.99% of outliers

Figure 2.11: The pixels saliency values repartition for the four selected image databases.

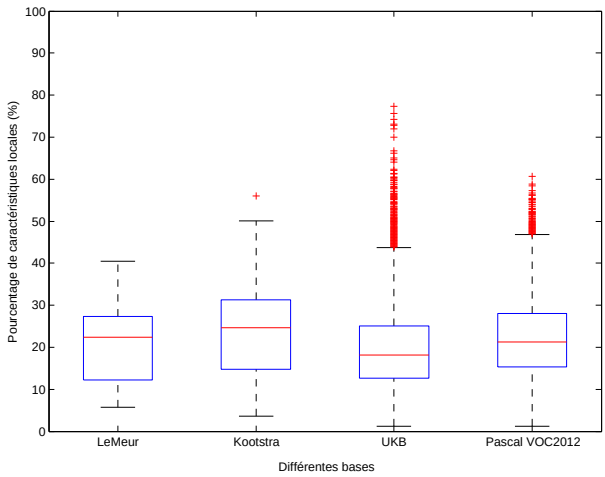
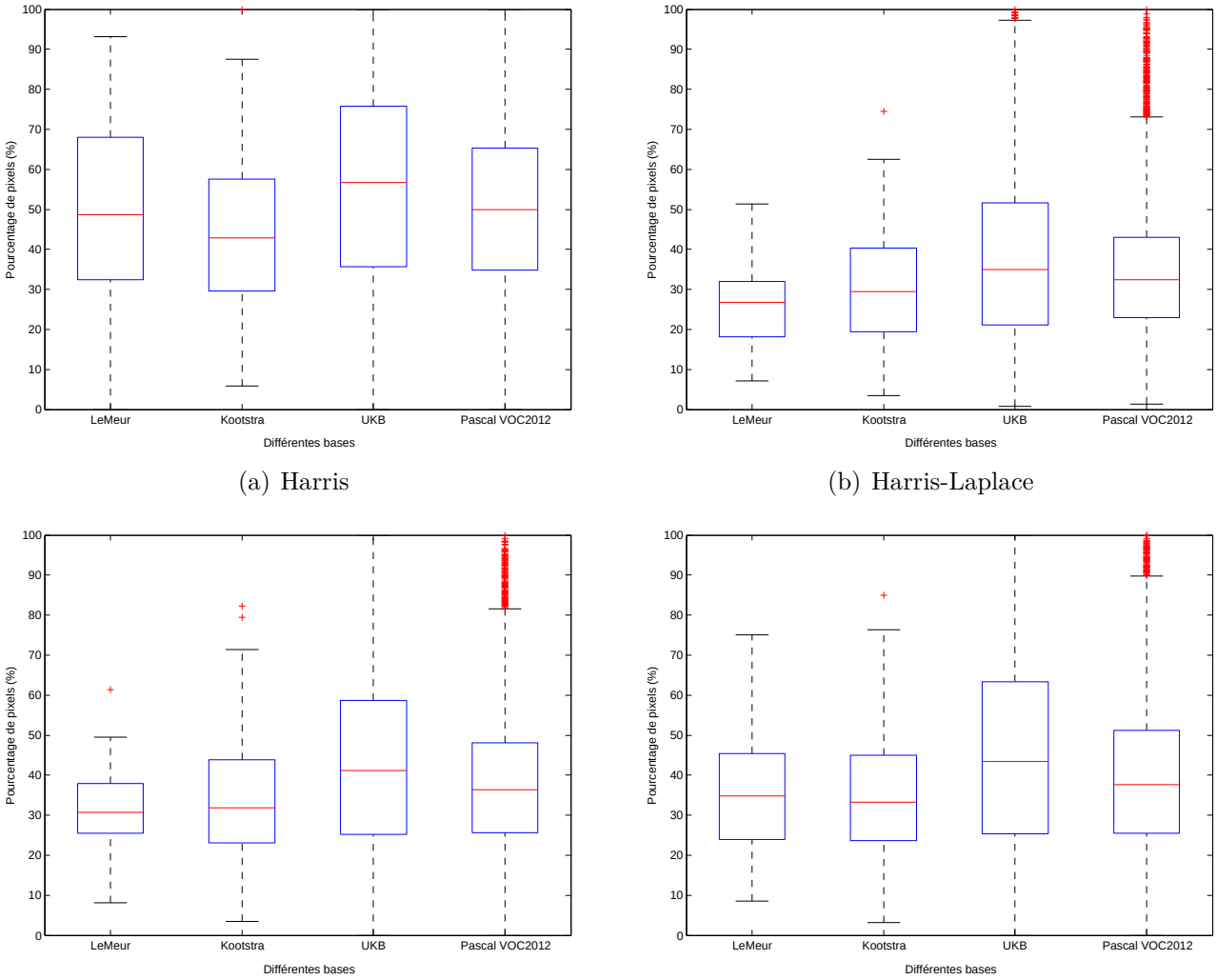


Figure 2.12: The average repartition of the pixels with saliency values ≥ 0.4.



(a) Harris

(b) Harris-Laplace

(c) DOG

(d) FAST

Figure 2.13: The average repartition of the local features with saliency values ≥ 0.4 for UKB and PASCAL VOC2012.

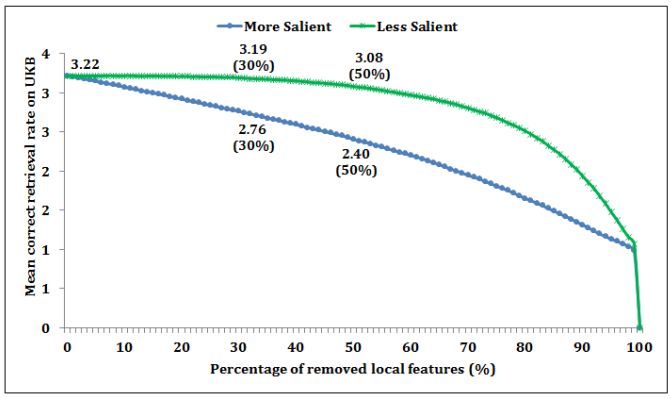


Figure 2.14: Local features detected by Harris-Laplace filtered according to their saliency value.

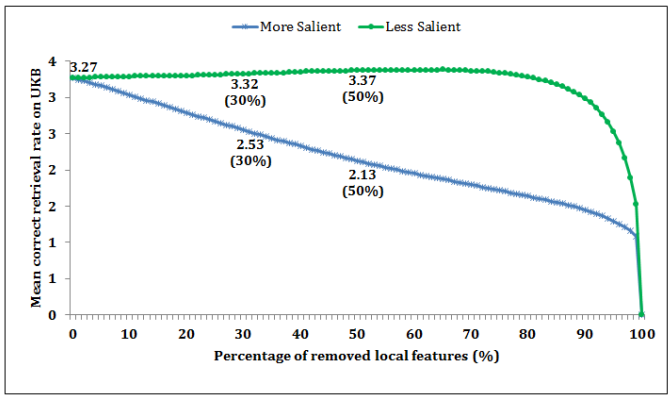


Figure 2.15: Filtering dense selected local features according to their saliency value.

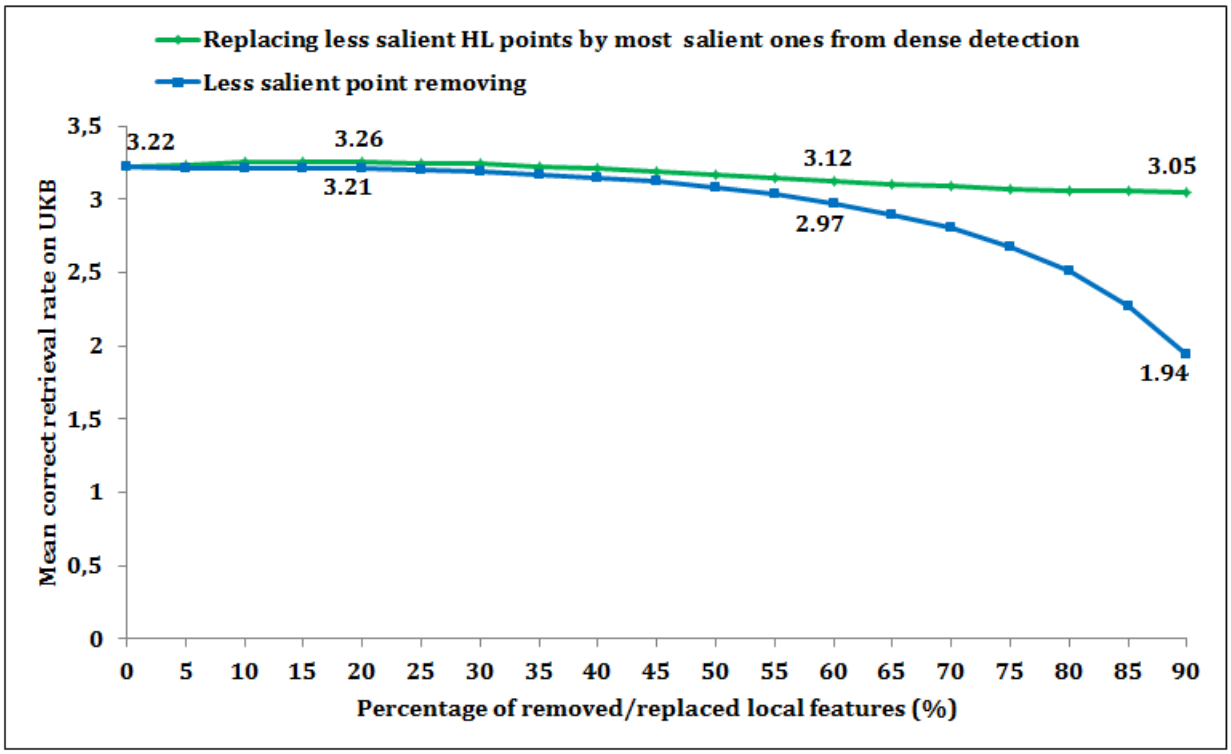


Figure 2.16: Replacing the less salient points detected buy Harris-Laplace by the most salient selected with dense quantization.

Part II

Image emotional impact recognition

Chapter 3

Emotion recognition in the literature

Contents

Emotion is a complex notion that can be defined as a psychological state that arises spontaneously rather than through conscious effort. It is sometimes accompanied by physiological changes. There are many other definitions of emotion according to the different schools of psychology. In fact, the concept of emotion is used in different ways as it is considered in reference to the stimulus aspect, the subjective experience, a phase of process, an intermediate variable or a response. This complexity induces different emotion classification.

3.1 Emotion classification

Usually two methodologies of emotion classification are used in the literature [Liu 11]:

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 modelling is usually preferred in emotions extraction based on facial expressions.

2. Dimensional approach: contrary to previous one, the emotions are considered as the result of fixed number of concepts represented in a dimensional space. The dimensions can be an axis of pleasure, arousal and power. These dimensions vary depending to the needs of the model. The most used dimensional model is Russel's represented in Figure 3.1 with the dimensions valence and arousal:

- The valence corresponds to the way a person feels when she looks at a picture. This dimension varies from negative to positive and allows to distinguish between negative and pleasant emotions.
- The arousal represents the activation level of the human body.

The advantage of these models is to define a large number of emotions.

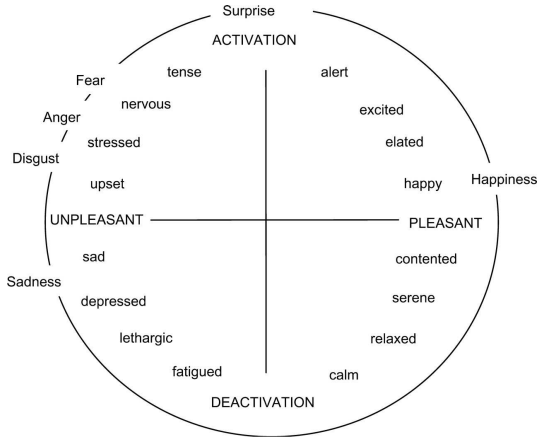


Figure 3.1: Russel's emotions modelling. The axe Unpleasant/Pleasant corresponds to the arousal and the second one to the valence.

In the literature, a lot of works are based on the discrete modelling of the emotions, for example those of Paleari and Huet [Paleari 08], Kaya and Epps [Kaya 04], Wei et al. [Wei 08], Ou et al. [Ou 04a, Ou 04b, Ou 04c].

Try to extract the emotional impact is an ambitious task, since different informations contained in an image (textures, colours, semantic, ...) can be emotional vector. More, many factors, including cultural aspects, more complex than the content are considered in our emotional interpretation of an image.

3.2 Some solutions about emotion recognition

A large part of the literature has long been devoted to the links between emotions and colours [Wei 08, Ou 04a, Ou 04b, Ou 04c, Boyatziz 93, Lucassen 10, Bradley 01, Beke 08]. Several studies have been conducted on the study of emotions associated with particular colours through culture, age, gender, social status influences. The authors agree on the fact that the colours convey particular emotions. As stated by Ou et al. [Ou 04a], colours play an important role in decision-making, evoking different emotional feelings. The research on colour emotion or two colours combination emotion is now a well-established area of research. Indeed, in a series of publications, Ou et al. [Ou 04a, Ou 04b, Ou 04c] studied the relationship between emotions, preferences and colours. They have established a model of emotions associated with colours from psychophysical experiments.
Another part of the literature concerns facial expression interpretations [Paleari 08].

Emotions are then associated with facial features (such as eyebrows, lips). It seems to be the easiest way to predict emotions. Indeed some facial expressions are common to human to express basic emotional feelings (happy, fear, sadness). In this case, the system detects emotions carried in the images and the videos and not really the emotions felt by someone looking at these pictures or videos.
To finish some authors considered the emotion recognition as a CBIR task [Solli 10, Machajdik 10, Yanulevskaya 08]. The underlying idea is to use the traditional techniques of image recognition. So image features have been extracted and used combined with a classification system to find the emotional impact. The most used features are: colours, textures and shapes. For example, Wang and Yu [Wang 05], used the semantic description of colours to associate an emotional semantic with an image. The orientation of the different lines contained in the images is sometimes considered. According to Dellandréa et al. [Liu 11], oblique lines could be associated with dynamism and action; horizontal and vertical ones with calm and relaxation.

3.3 Image databases for emotion recognition

Image datasets used for emotion study are often different according to their content, from abstract images to photography. We focused on three databases: the two datasets proposed by Machajdik and Hanbury [Machajdik 10] and International Affective Picture System (IAPS) [Lang 08].

The two datasets proposed by Machajdik and Hanbury [Machajdik 10]
Machajdik and Hanbury published[1] two image datasets:

1. Abstract paintings which consist only of combinations of colours and textures, without any recognisable objects. To obtain a ground truth, the images were peer rated in a websurvey where the participants could select the best fitting emotional category from the ones mentioned above for 20 images per session. 280 images were rated by approximately 230 people, so each image was rated about 14 times. For each image the category with the most votes was selected as the ground truth. Images where the human votes were inconclusive were removed from the set, resulting in 228 images.

2. Artistic photographs downloaded from an art sharing site[2]: for these images the emotion category was determined by the artist who uploaded the photo. These photos are taken by people who attempt to evoke a certain emotion through the conscious manipulation of the image composition, lighting, colours, etc. This dataset therefore allows them to investigate whether the conscious use of colours and textures by the artists improves the classification.

They chose discrete modelling of emotions in their evaluations. To generate the output categories of emotions they used the emotional word list defined by Mikels et al. [Mikels 05] in a psychological study on adjective images. Their emotional output

[1] http://www.imageemotion.org
[2] deviantart. www.deviantart.com

categories are: *Amusement, Awe, Contentment, Excitement* as positive emotions, and *Anger, Disgust, Fear, Sad* to represent negative emotions.

International Affective Picture System (IAPS) [Lang 08]
This dataset is composed of photographs in emotion research, developed since the late 1980s at NIMH Center for Emotion and Attention (CSEA) at the University of Florida. 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. Each of the 1182 images[3] from their dataset was evaluated by about 100 participants. The image was displayed to be assessed for 6 seconds. Then, observers had 15 seconds for adults, 20 for children, to assess their emotions in the scoring system SAM (Self Assessment Mannequin) [Lang 08], which is a graphic figure that ranges:

- From smiling and happy to frowning and unhappy in representing the hedonic valence dimension;
- From excited and wide eyed to relaxed and sleepy for the arousal dimension;
- From a large figure (in control) to a small figure (dominated) for the dominance dimension.

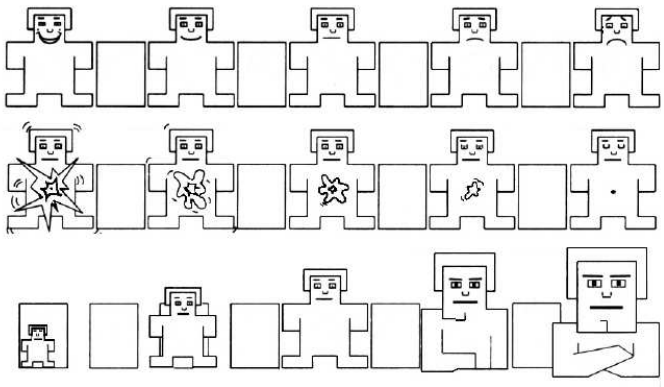


Figure 3.2: An example of SAM used during the IAPS evaluation. At the top there are the representations to assess the pleasure, at the middle, the arousal and at the bottom, the dominance.

Figure 3.2 is an illustration of a SAM used during the IAPS evaluation. During this evaluation, the participant can fill in any of the five figures depicting each scale or the box between any two figures, resulting in a 9-point rating scale for each dimension.

The IAPS is considered as a reference in psychological studies and many papers on emotions study domain present their results on this dataset [Liu 11, Machajdik 10, Yanulevskaya 08].

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

Chapter 4

Our approach for emotion recognition

Contents

In this chapter, firstly we present a new taxonomy for the description of the databases used for emotion study. The set of criteria proposed is based on different deficiencies of the literature. We share those reported by Machajdik and Hanbury [Machajdik 10] by evoking the fact that:

- The datasets are in most cases unknown (unpublished);
- In many cases, no information is given on how the images were selected, for example if there was a manual filtering process that could potentially be biased;
- The evaluation measures are often poorly described ([Yanulevskaya 08]).

Secondly we propose a new image database widely assessed using visual saliency information. Our approach for emotion recognition is part of features based methods. We evaluated some low level features that offer good results in object recognition and image retrieval.

4.1 The new set of criteria proposed

The criteria we propose allow to describe briefly the database according to three kinds of information: inherent, extrinsic information and physiological evaluations available.

4.1.1 Inherent criteria

Inherent information describe the first selective details. They concern:

- *The number of images* in the database;
- *The average evaluation per image* that indicates how many people in average assessed one image;
- *The "free to use"* aspect of the images that allows to know if the database can be modified[1] to explore another way to study the emotions.

4.1.2 Extrinsic criteria

Extrinsic information concern:

- *The database availability* which corresponds to the rapid availability of the database for the community. For example the two databases used by Machajdik and Hanburry [Machajdik 10] can be downloaded on line and IAPS needs a request;
- *Emotions modelling* used: in fact, the need of dimensional or discrete modelling of emotions depends to the applications;

[1]Obviously we are talking about different potential modification or transformation for scientific research.

- *The heterogeneity of ratings* according to the gender, to the age of participants;
- *The nature of the emotional impact* of the different images;
- *The evaluation complexity* that defines the ease of annotation for the general public. The dimensional emotions modelling used to assess IAPS seems less easier than a discrete model. In fact, the arousal and the dominance assessment can be confused.

4.1.3 Physiological evaluations available

In addition to "conventional" methods[2] other ways of understanding emotions were tested. Among them we can mention here the Steady-state Visually Evoked Potentials (SSVEP) response. An Evoked Potential (EP), in the context of EEG signals, is an electrical potential elicited by the presentation of a stimulus that can be recorded from the nervous system. In particular, in the case of non-invasive EEG recordings, it can be acquired from electrodes positioned on the surface of the scalp. Visually Evoked Potentials (VEP) are EP elicited by a visual stimulation. Steady-state VEP (SSVEP) are a particular case of VEP, where the stimulus is presented multiple times at a frequency at least higher than 3.5Hz, but more commonly higher than 6Hz [Friman 07, Kemp 02]. In this case, a periodic response called SSVEP can be observed in the recorded scalp EEG signal, in particular in the occipital brain region, where the visual cortex resides.
Various evidences exist suggesting the SSVEP response not being only a mechanical reaction of the brain to a flickering stimulus. Indeed it is known to be modulated by the user's attention and affective state [Kemp 02, Keil 03]. In particular, in previous works [Kemp 02, Keil 03, Wang 13], flickering pictures from the IAPS [Lang 08] have been showed to a group of users during the acquisition of their EEG. Amplitude, latency and topography of the SSVEP response have been shown to be correlated to the arousal and valence of the shown pictures.

4.1.4 Comparison of the three databases presented in the previous chapter based on our criteria

With those new criteria, it is now rather easy to compare the datasets, to have a interesting overview about them. We can see in Table 4.1 that the different datasets used by Machajdik and Hanburry [Machajdik 10] lack of evaluations quality. The number of participants is not suitable compare to the IAPS. No information about the heterogeneity of the observers is mentioned. However they are interesting for applications that do not need high emotional images. The main weakness of the IAPS is the restriction about the evaluation protocol and the usage of the images in addition to the high emotional impact of some images. Anyway for the three databases mentioned in the Table 4.1, any wish of modification of the dataset must be address to the authors.

[2]We call conventional methods all methods that allow a classification of emotions according to a given model.

Table 4.1: Comparison of three data sets of the literature according to the new criteria. Machajdik1 is the abstract paintings used by Machajdik et al.[Machajdik 10] and Machajdik2 the artistic photographs they also used.

		Databases		
		Machajdik1	***Machajdik2***	***IAPS***
Inherent information	*Number of images*	228	807	> 1000
	Average evaluation per image	14	1	~ 100
	Free to use database	Yes*	Yes*	Yes*
Extrinsic information	*Database availability*	+++	+++	++
	Emotions modelling	Discrete	Discrete	Discrete** Dimensional
	Rating heterogeneity	Unknown	No	Yes
	Emotional impact nature	+	++	+++
	Evaluation complexity	++	++	+++
Physiological evaluations available		-	-	EEG (SSVEP)

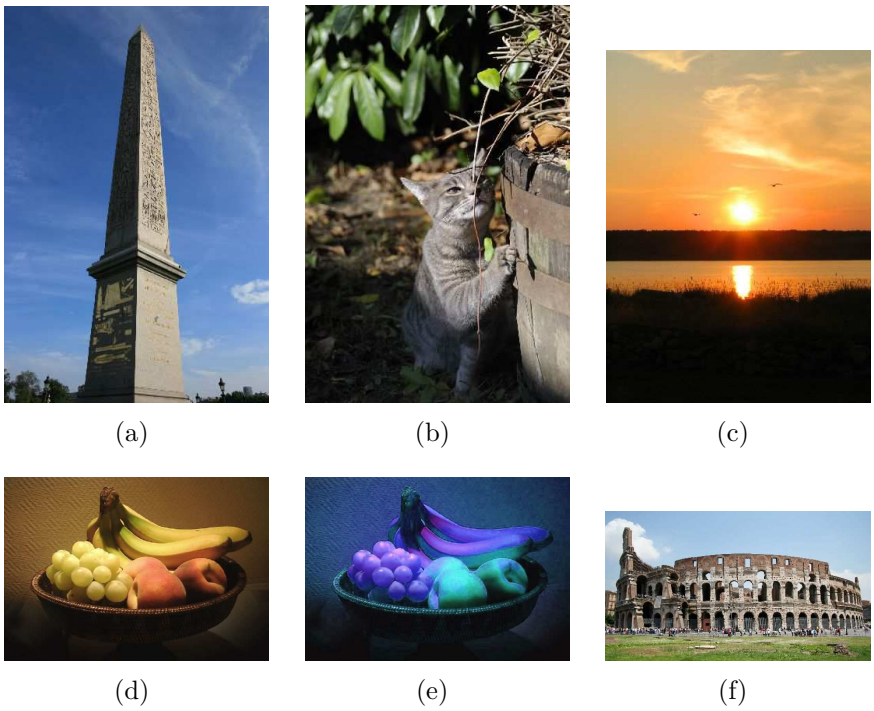
* For academic research, not for profit research. For IAPS there is some specific terms about the evaluations of the image set. For example, not place them on the Internet. For the two other databases, according to the authors they are provided solely for scientific use, to allow results to be compared to those in their paper.

**Just 384 images were assessed according to a discrete model

In our case we need low semantic images to easily extend our research to daily life images so IAPS is not interesting for us. We cannot use one of the datasets used by Machajdik and Hanburry [Machajdik 10] because of the number of people who rated the images. These are the two main reasons of the building of a new database. The last reason is our wish to evaluate the images introducing visual attention information.

4.2 Presentation of our image dataset

Our studies on emotions are essentially focused on low semantic images, whose are images of daily life. When we talk about low semantic images it seems impossible since every image has a semantic. In our case, we address images which do not provoke some high emotions. We chose to deal with "primary emotions" which can be considered as the first feelings, the instinctive judgment. We also choose low semantic images to 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.



(a) (b) (c)

(d) (e) (f)

Figure 4.1: Images from SENSE.

Our database is composed of 350 low semantic, natural and diversified images (Some example in Figure 4.1) and is called SENSE (Studies of Emotion on Natural image databaSE). This set of images is free to use for academic research purpose. The images can be modified and another kind of evaluations can be organized. The only constraint is to mention the name of the authors. For the moment the different results of the evaluations on SENSE are available on request.
SENSE has also the advantage to be mainly composed of natural images except some non-natural transformations (rotations and colour balance modification) on few images. These transformations are performed to measure their impact on emotions recognition system based on low-level image features. This new image dataset contains only 4.86% human faces. The few rate of this kind of images is explained by our wish not to influence observers by judging the faces emotions but only the emotional impact of the whole image.

4.3 Evaluations of SENSE

Our goal during psycho-visual evaluations is to assess the different images according to the nature of the emotional impact during a short viewing duration. For these ratings, viewing duration is really important. In fact, if the observation time extends observers access more to the semantic and their ratings are semantic interpretations and not really "primary emotions".
During our tests the observers assessed nature and power of emotional impact of the images. For the nature, they had choice between "Negative", "Neutral" or "Positive" and the power varies from "Low" to "High" as shown in Figure 4.2. We chose these

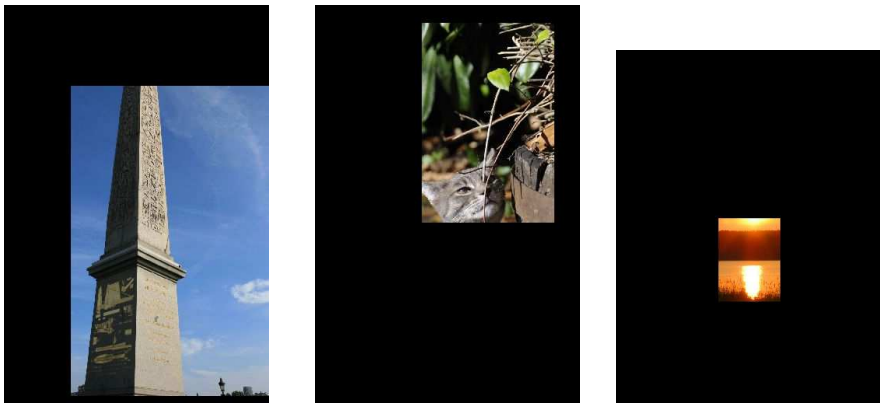
Figure 4.2: Screen shot of test application.

information to define emotions because according to us, it is the best way to evaluate globally a "primary" emotion for low semantic images. Discrete modelling is not adapted in our case. In fact in a discrete representation of emotions, emotional process is explained with a set of basic or fundamental emotions, innate and common to all human. Our database assessment according to this approach can be difficult. For example, scoring an image like "Happy" or "Sad" on a low semantic database needs a real semantic interpretation after a short observation time.
We organized two kind of evaluations, made several months apart:

1. During the first experimentations called SENSE1, observers assessed the full images of the database;

2. During the second evaluations, participants assessed regions of interest obtained with a visual saliency model. The evaluation of this new set of images is called SENSE2 in this paper. Figure 4.3 shows some images assessed during SENSE2. The size of the images evaluated during this test varies from 3% to 100% of the size of the original ones. The thumbnails are built with the bounding rectangle of the salient regions.

For the different subjective evaluations, we decided to use the Internet in order to have suitable number of observers giving statistically significant results for the full database rating. This media also offers the advantage that the participants take voluntarily the test and that remains pleasant and without constraints. These factors are very important for our studies. During the evaluations 24 images were randomly selected and the observation time was not imposed. The observers can move to the next picture when they want or stop the test. We just asked them to answer as quickly they can to limit the semantic interpretation.

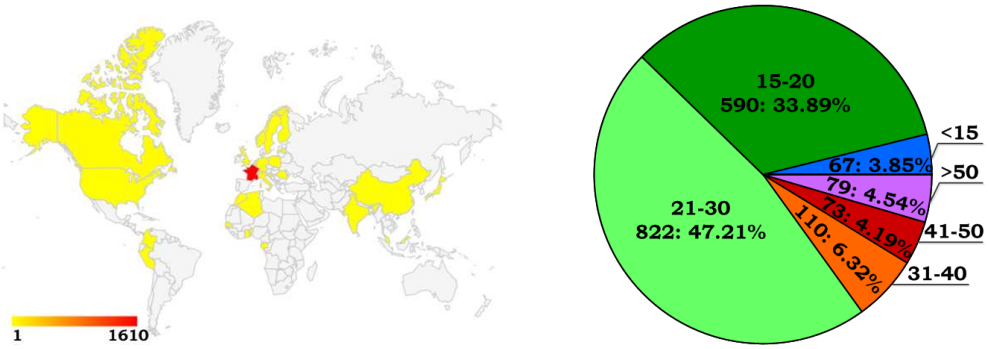


(a) 61% (b) 27% (c) 6%

Figure 4.3: Thumbnails corresponding to the images 4.1(a)-4.1(c) scored during SENSE2. The size of the regions of the interest is given as a percentage of the size of the original image.

4.3.1 Experimentations SENSE1

1741 participants including 893 women (51.29%), took the test around the world (28 different countries) as shown in Figure 4.4(a). The majority of the participants lives in France.



(a) Observers around the world. (b) Age distribution of observers.

Figure 4.4: Description of the subjective evaluations SENSE1.

The database was evaluated by people of all ages, from under 15 to more than 50 as shown on Figure 4.4(b). The large part of them were aged 15 to 30 years as they represent 81.1%.

During the analysis of the results after SENSE1 we only considered scoring with a duration time between 3 and 8 seconds. The average time of observation is 6.6 seconds. Each image was assessed by an average of 104.81 observers. Only 6% of the database was scored by less than 100 persons (the less assessed image was evaluated by 86 different participants).
Despite the fact that we use a low semantic database, observers are really consistent in their scoring. On Figure 4.5 we represent the average percentage of observers

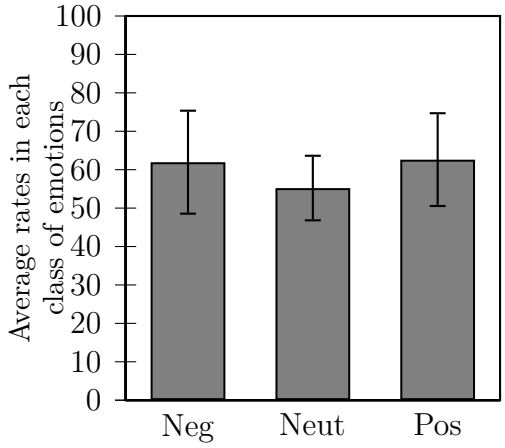


Figure 4.5: Average rate for each nature of emotions during SENSE1. The average rate is represented with the standard deviation.

which gave the class of emotions. We considered that an image is categorized in some emotion nature class Negative, Neutral or Positive, if the difference of the percentages of observers between the two most important emotions is greater than or equal to 10%[3]. In the case of positive or negative emotions observers are more unanimous than for neutral choice. The average rates for negative and positive emotions can be explained by the fact that, for many images the majority of participants has the same emotional impact. In fact, neutral emotions have not the same definition for all observers. Some of them choose "Neutral" for images without particular feelings; the others because they do not know how to describe their feelings. This ambiguity does not really pose a problem in our case. We want to have the feeling for a maximum of participants. One solution to avoid the problem of the heterogeneity of neutral emotions could be to ask observers just two natures of emotions (Positive and Negative) but it would force them to consider a specific class. So we use "Neutral" as a reject class.
In spite of the number of participants, 61 images (17.43%) were not clearly categorized. We think that this is related to their content. Even if we try to reduce the semantic interpretation with a short viewing duration, it does not work every time. So we think we can improve our evaluation by finding the way to reduce the access to the semantic. The idea is to resume the full image with a thumbnail containing the attractive information for the content understanding. To do this we choose saliency which appears to be a good strategy to reduce the amount of data and to conserve the more visual attractive information.

4.3.2 SENSE2: Visual saliency usage to reduce the size of viewed regions

Saliency model used
We choose an hybrid model proposed by Perreira Da Silva et al. [Perreira Da Silva 10] which allows to model the temporal evolution of the visual focus of attention. It is

[3]If an image has the following percentage of classification 43, 17, 40 respectively for Negative, Neutral and Positive emotion, it will be considered as uncategorized because (43-40)<10.

based on the classical algorithm proposed by Itti [Itti 98], as shown in Figure 4.6.

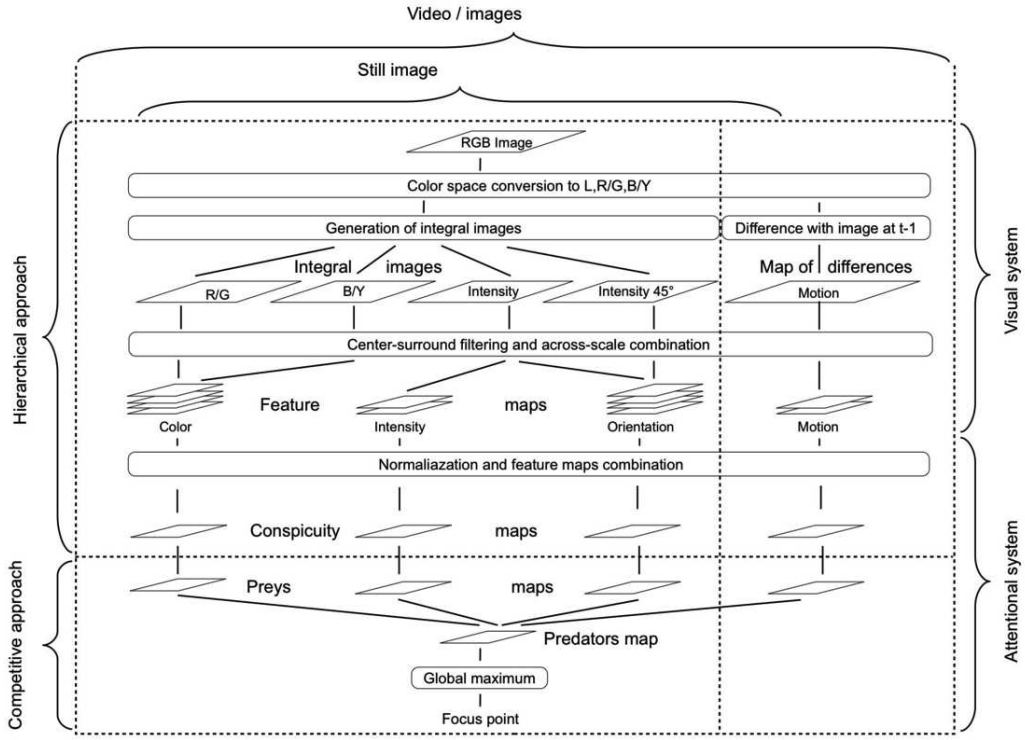


Figure 4.6: Architecture of the used model of attention.

The first part of this model architecture relies on the extraction of three conspicuity maps based on low level characteristics computation. These three conspicuity maps are representative of the three main human perceptual channels: colour, intensity and orientation. Perreira Da Silva et al. [Perreira Da Silva 10] proposed to substitute the second part of Itti's model by an optimal competitive approach: a preys/predators system. They have demonstrated that it is an optimal way to extract information. Besides, this optimal criteria, preys/predators equations are particularly well adapted for such a task:

- Preys/predators systems are dynamic, they include intrinsically time evolution of their activities. Thus, the visual focus of attention seen as a predator, can evolve dynamically;
- Without any objective (top-down information or pregnancy), choosing a method for conspicuity maps fusion is difficult. A solution consists in developing a competition between conspicuity maps and waiting for a natural balance in the preys/predators system. That reflects the competition between emergence and inhibition of elements that engage or not our attention.

The authors show that despite the non deterministic behaviour of preys/preda-tors equations, the system exhibits interesting properties of stability, reproducibility and reactiveness while allowing a fast and efficient exploration of the scene.

We applied the same optimal parameters used by the authors to create the thumbnails of the images of our database.

SENSE2 Evaluations

1166 participants including 624 women (53.49%) and 542 men (46.51%) scored the 350 images. Each image was assessed by an average of 65.39 observers. Only 2 images were rated by less than 50 persons.

If we considered the results of SENSE2 according to the size of the thumbnails[4], we noticed that when the percentage of the image is less than or equal to 7%, the images are "Neutral" or "Uncategorized".

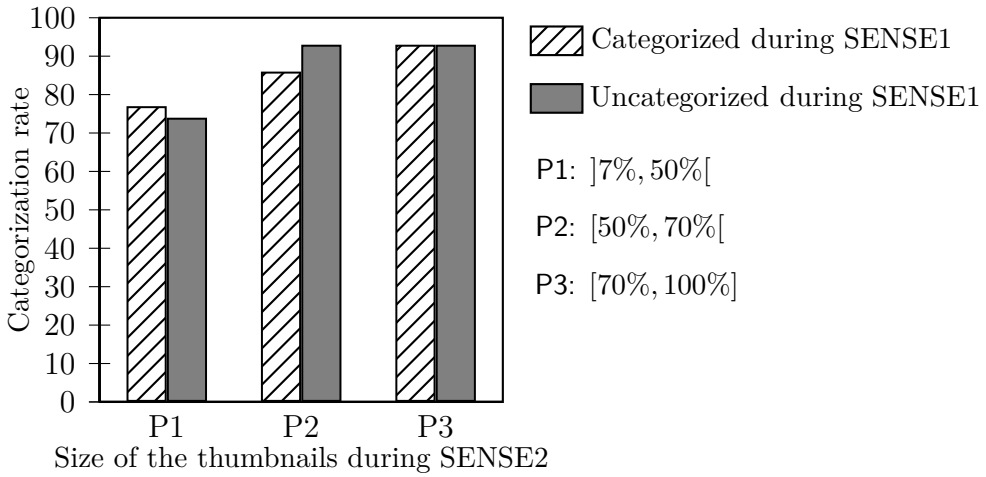


Figure 4.7: Average classification rates during SENSE2. "Categorized during SENSE1" corresponds to the images with the same class during SENSE1 and SENSE2. "Uncategorized during SENSE1" describes the uncategorized images during SENSE1 and now categorized during SENSE2.

In Figure 4.7, we represent the rate of images categorized during SENSE1 and SENSE2 in the same class of emotions. There are also uncategorized images during SENSE1 and definitively classified during SENSE2. Regarding the different results during SENSE1, the reduction of the viewed area according to a visual saliency model is a good solution to evaluate the primary emotions. This conclusion is confirmed for categorized and uncategorized images during SENSE1. During SENSE1, 61 images are "Uncategorized", Figure 4.7 shows that a large part of them (79%) is categorized during SENSE2. Reduce the viewing region has certainly reduced the semantic and the analysis time.

Figure 4.8 shows that for the three classes of emotions, when the viewed area has a size equal to at least 50% of that of the original image, 77% of the images are correctly categorized. This notice answers to our hypothesis that the idea of reduction of the images with a visual attention model can offer similar results compare to the full images.

The hypothesis of semantic interpretation reduction by assessing the bounding rectangle of the most salient regions could be very helpful for evaluation if the interest

[4]Which corresponds to ratio of the original image represented by the visual region of interest

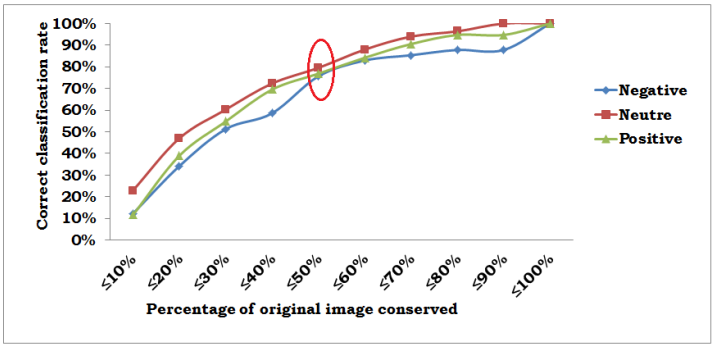


Figure 4.8: Rate of good categorization during SENSE2 according to the percentage of original image viewed.

regions are not too small[5].

4.3.3 SENSE description according to our criteria

Our database description is resumed in Table 4.2 according to the new criteria we proposed. Compared to IAPS, SENSE contains much less images. But it is

Table 4.2: Description of SENSE according to the criteria defined in Section 4.1.

Inherent information	*Number of images*	350
	Average evaluation per image	∼ 100 for SENSE1 ∼ 65 for SENSE2
	Free to use database	Yes*
Extrinsic information	*Database availability*	+++
	Emotions modelling	∼ Dimensional**
	Rating heterogeneity	Yes
	Emotional impact nature	++
	Evaluation complexity	+
Physiological evaluations available		-

* For academic research, not for profit research.

** Our emotions modelling is equivalent to a dimensional.

a widely assessed low semantic database. This database has been evaluated using saliency information and the results show that this HSV information can be usefull for emotion recognition assessment.
We mention that there is no physiological evaluations available but, in fact, we start EEG recording and we have not enough data to propose them. For the moment only 12 images have been shown to 4 observers for EEG recording. The result analysis

[5]In our case $\leq 7\%$ of the original image size

highlights that the obtained SSVEP response is related to the content of our image but we cannot say exactly which low level features are discriminative.

4.4 Low level feature evaluation for emotion recognition

Our goal during this study is to evaluate some traditional indexation image features during an emotional impact prediction task on a low semantic database. Obviously, we also chose our features according to some hypothesis between the emotion related to an image and its content.
Many works in psychology make hypothesis about relationship between colours and emotions. Colours are the first discriminated characteristic of images for the extraction of the emotions. Often, colours reflect the interpretation of the semantic linked to some situations, phenomena and also culture. Textures are also important for emotional analysis of an image. For example, a grid regardless of its colour has a semantic of confinement. We have finally supposed that local descriptors could also implicitly encode high-level data.
We used in our classification process two classes of features; one computed on global information and the others on local.

4.4.1 Features based on global information

The global features we used are related to colours, textures and global scene description.
To identify the different colours, we used colour segmentation by region growing [Fernandez-Maloigne 12]. For the initialization of the seeds, we performed an analysis of greyscale histogram. The analysis of a histogram was made in greyscale to save time in homogeneous areas. To convert colour images to greyscale we have used the equation (4.1) according to the NTSC standard.

$$gray = 0.299R + 0.587G + 0.114B \tag{4.1}$$

The seeds are the maxima of the greyscale histogram. The region growing was performed in CIE Lab colour space in order to have a Euclidean distance correlated with the perceptual distance. The distance between colours is computed with ΔE obtained with the equation (4.2) (for two colours C_1 and C_2). We have retained only the average colour of different regions.

$$\Delta E = \sqrt{(L_1^* - L_2^*)^2 + (a_1^* - a_2^*)^2 + (b_1^* - b_2^*)^2} \tag{4.2}$$

For textures extraction we converted images to greyscale also in accordance with NTSC standard. Our textures features are based on Wave Atoms transform introduced by Ying and Demanet [Demanet 09]. The Wave Atoms are in first approximation a variant of 2D wavelet packets with a parabolic wavelength scale. Like all multi-scale transforms (the wavelet transform, for example), there are several

information from different levels. The number of coefficients for each orientation depends on the decomposition level. Before applying Wave Atoms transform we resized all image to $256 * 256$ with zero padding if it is needed. With this new size, we had 5 levels of decomposition. We just worked with the scales 4 and 5. Scale 4 is composed of 91 orientations. Each orientation contains $2^4 * 2^4(256)$ coefficients. Scale 5 contains 32 orientations and 1024 coefficients per orientation.

GIST introduced by Oliva and Torralba in 2001 [Oliva 01] is computed for the global description of the scene. It allows to have a low dimensional representation. These descriptors are 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 computed GIST on images resized to $256 * 256$ with zero padding (if it is needed) because the authors advise to use images with same dimensions to compare GIST.

4.4.2 Features based on local information

We opted for the same descriptors used in Chapter 2:

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

These features were computed on local points detected with the Harris-Laplace point detector for SENSE1 images. For SENSE2 images we used a dense detection because some thumbnails are to small.

4.4.3 Experimental protocol

We used BoVW and VLAD for the image signatures. Except for GIST the visual codebooks are obtained with *K-Means* algorithm and IteRaSel. For GIST to obtain the visual vocabulary we used PCA as advise Oliva and Torralba [Oliva 01]. The size of the codebook in this case is determine according to the percentage of information conserved during the PCA: we fixed it at 98%.
For image classification, we defined three classes according to the nature of emotion asked during our subjective evaluations. For IAPS we form these classes according to the valence values. We used SVM classifier with a linear kernel in its multiclass extension "One against one".
We consider emotion recognition as a content based image recognition task. So we use different tools from this task: features, visual signatures, ... Our aim is to study the impact of these step for emotion recognition accuracy.

4.4.4 Study of the visual codebook impact

For this we use BoVW as the visual signature and two different visual codebooks:

1. A codebook obtained with a *K-Means algorithm* with:

$$K = \sqrt[4]{N * d} \tag{4.3}$$

In the equation (4.3), K is the number of visual words, N the number of descriptors and d the size of vector of characteristics.
We used in this case the two databases SENSE and IAPS, so there is two dictionaries and they are resumed by the notation Dataset_Visual codebook. 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 the results are dependent on the image database used to create the visual dictionary.

2. A codebook obtained with IteRaSel algorithm: only local features are considered for this configuration. The visual vocabulary is computed from Pascal VOC2012.

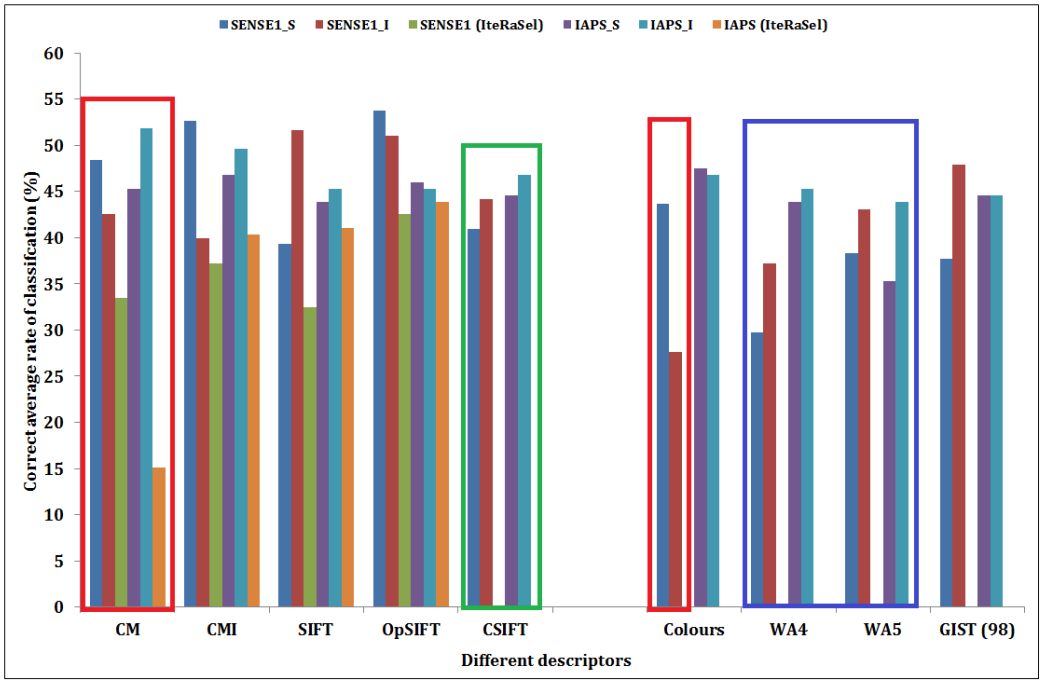
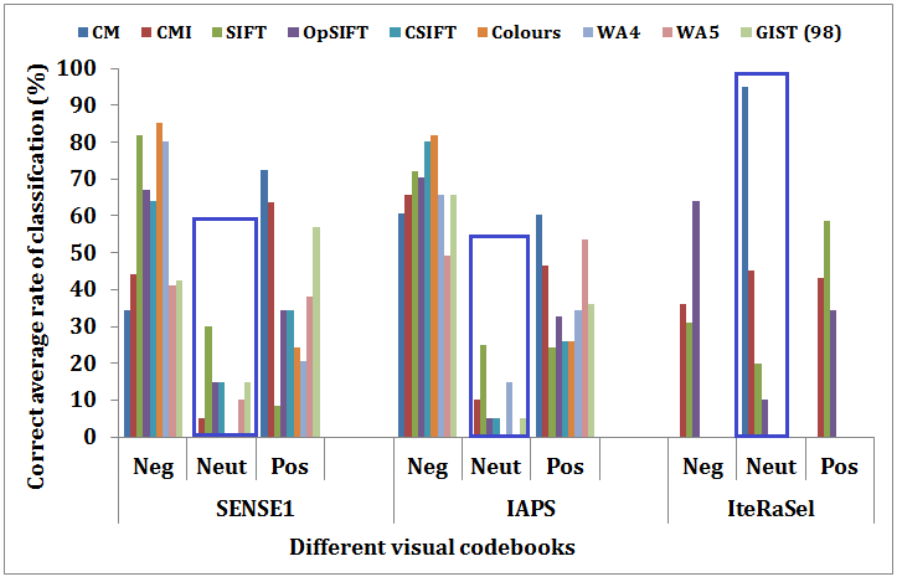
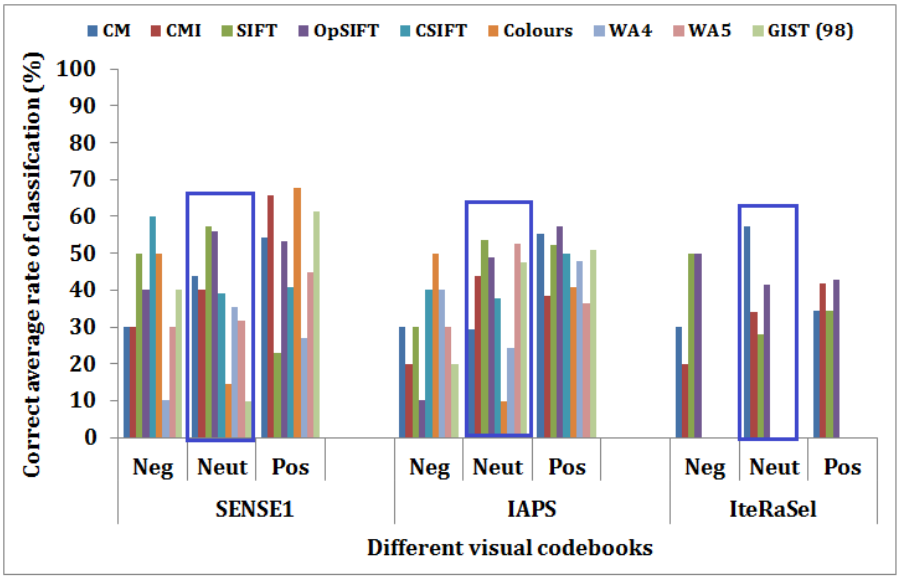


Figure 4.9: Average classification rates for SENSE1 and IAPS.

Figure 4.9 shows the average classification rates for SENSE and IAPS. The first conclusion concerns the different results function to the descriptors. CSIFT results are stable regardless of the database and the visual codebook. It is not the same for the others descriptors. CM results depend on the database and the vocabulary. Colour codebook obtained with SENSE performs better on SENSE and IAPS: +15% for SENSE1_S. For better understanding of the results we analyzed the results for each emotion class with the Figure 4.10. This figure shows that neutral images are difficult to recognize with our system. For example on IAPS, no descriptor, except CM does better than random selection. We can also conclude that colour descriptors



(a) IAPS: the results of class "Neutral" are highlight.



(b) SENSE1: neutral images are better recognized compared to IAPS.

Figure 4.10: Classification rate in each emotion class for the different descriptor. Note that IteRaSel is only computed for local descriptor. Gist (98) means that 98% of the information is conserved during the PCA.

are well adapted for negative and positive images.
This first study informs that the results differ depending the base used for visual codebook and the descriptors.
After visual codebook impact we also study the visual signature impact. The aim of this study is to help us to make the good choice for the image representation according to the kind of descriptor. We consider again here the three codebook used previously.

Our analysis shows that VLAD are more adapted for local features regardless the visual codebook to preserve an acceptable results (> 33%). So for the presentation of our final approach we use BoVW for global descriptors and VLAD for local descriptors. We also remove the neutral class and just considered a two classes problem: positive and negative images.

4.4.5 Presentation of our results for positive and negative emotions

Our results are presented in Table 4.3. The different features have not the same

Table 4.3: Classification rates after classification for each descriptor.

Descriptors		Nature of emotions	Configuration base de test_Dictionnaire visuel				Average
			SENSE1_S	*SENSE1_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%**

behaviours on predicting emotions in the different configurations tested. For example, SIFT have approximately the same results for negative and positive emotions on IAPS and SENSE regardless the vocabulary changes. On the contrary, CMI and WA4, for example, seem more adequate for negative images with at least 50%.

Table 4.4: Comparison of correct average classification rates on SENSE and IAPS before and after fusion with Majority Voting.

		Before fusion	After fusion
SENSE1_S	Negative	55.56%	60%
	Positive	54.17%	57.29%
	Average	54.86%	57.55%
SENSE1_I	Negative	64.44%	90%
	Positive	54.86%	64.58%
	Average	59.65%	66.98%
IAPS_S	Negative	61.75%	75.41%
	Positive	47.13%	41.38%
	Average	54.44%	58.82%
IAPS_I	Negative	65.58%	77.05%
	Positive	45.02%	46.55%
	Average	55.30%	62.18%

Overall, the visual dictionary has little impact on the behaviour of descriptors for classification for SENSE and IAPS. However, CM descriptors for example, are affected. The rate of recognized negative images is significantly higher with codebook from IAPS (+ 70% for SENSE images and + 20% for IAPS images). The opposite effect is observed for positive images: -34% for SENSE images and -17% for IAPS images. This illustrates very well the impact of the variability of the database. Indeed, IAPS contains a lot of negative images: the dictionary built with this dataset allows to better recognize negative emotions. Building the visual dictionary with SENSE improves recognition of positive images since this base contains a lot.
To summarize, the features based on SIFT offer good prediction rates using a codebook only composed of 64 words and VLAD as visual signature. The best recognition of negative images is done through CSIFT with 90% of the images of SENSE1 recognized with the dictionary of IAPS. Global descriptors show great complementarity between the characterizations of images we have chosen. This is illustrated very well by the results of WA4 and WA5. The first one is more suitable for the negative images while the second will be preferred for the positive. We could also conclude that the negative images are much easier to recognize in the two databases that we have chosen.
In Table 4.4 we compare the classification rates before and after fusion with Majority Voting. There is a significant improvement after the fusion. For example, the recognition of negative images is impacted positively by 15% on average. Besides the best classification rates are obtained after merging using the dictionary built from IAPS. This conclusion is also valid for positive images. For both configurations (SENSE1_I and IAPS_I) before the fusion, 54.86% and 45.02% positive images were recognized against 64.58% and 46.55% after. If we generally consider

these results after fusion, we see that they have been improved especially on our image database, 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 SENSE1.

Note that for IAPS, positive image average results are lower than a simple random selection. This can be due to the database or simply because negative images are easy to recognize.

4.4.6 Comparison with literature

Before present the results we would like to mention that it is not easy to compare the different works about the extraction of emotions because of the differences in the databases and the features chosen. In fact, this comparison allows us to validate our approach. Indeed, if our results were well below those in the literature, using CBIR approach could be judged inappropriate.
We chose three results to make the comparison:

- Those of Wei et al. [Wei 08]: they used a semantic description of the images for emotional classification of images. They 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."
- Those of Lui et al. [Liu 11]: they proposed 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.
- Those of Machajdik et al. [Machajdik 10] in which colour, texture, composition and content descriptors are used. 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 these three, we can conclude that they are 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. The methodology we have adopted allows us to match the literature even do better in terms of classification rate. Note that this is just one indication and it is not a judgment on the methods due to the eclecticism of work in the field. This comparison is important to validate our approach: emotion recognition task can be achieved with CBIR techniques and features.

4.5 Consideration of the visual saliency: SENSE2 image classification

The evaluations SENSE2 results show that the regions of interest evaluation is equivalent to the full image evaluation. So we decided to replace the SENSE1 images by those used during SENSE2. The results presented here are for the local descriptors. Because of the variable size of these images (from 3% to 100% of the size of the original images) we chose dense selection. For effective comparison, we also consider dense selection for SENSE1. The first results are shown in Figure 4.11.

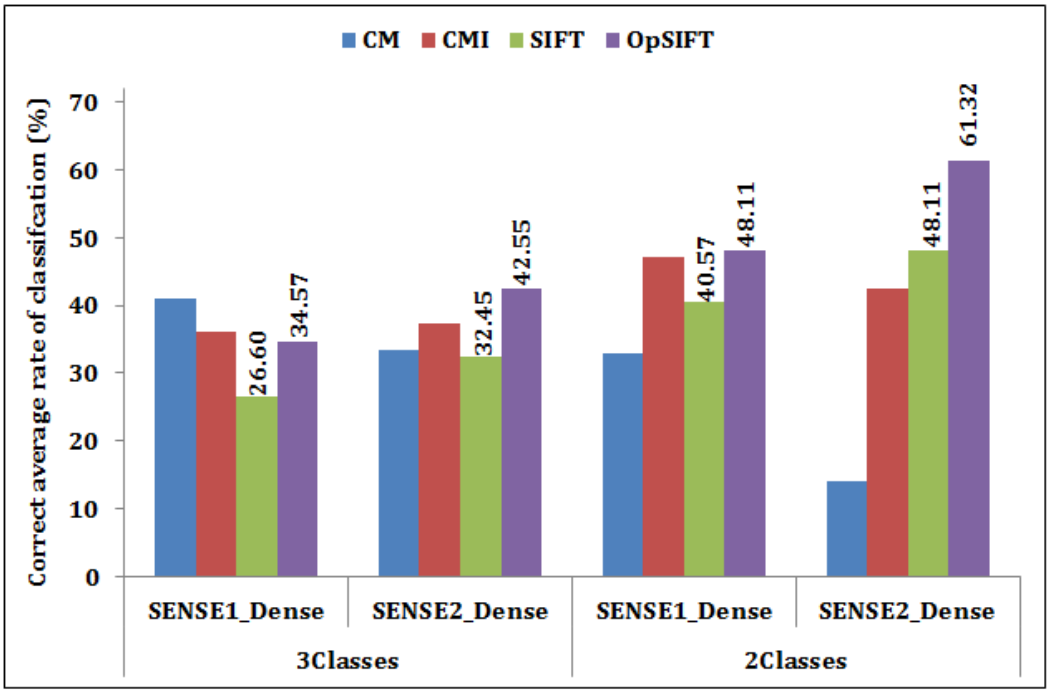


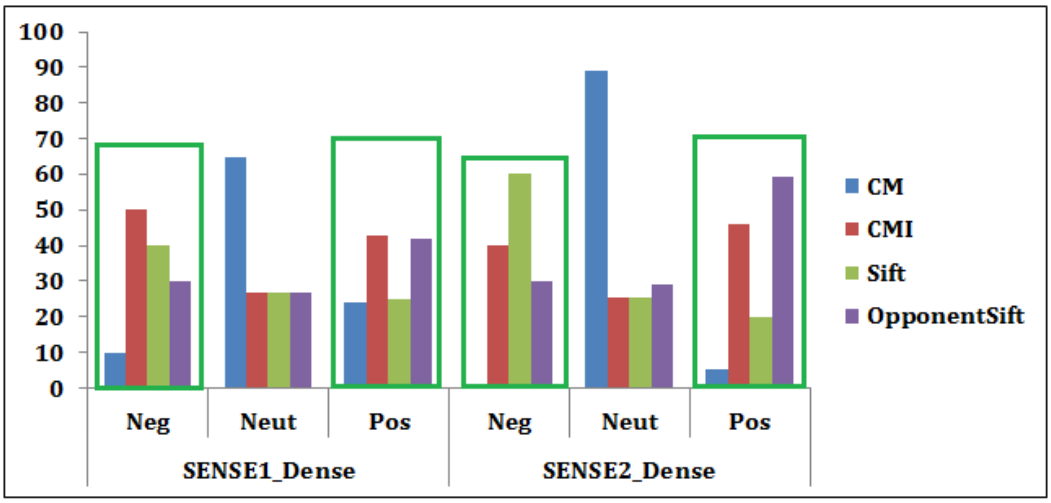
Figure 4.11: Average classification rates obtained for SENSE2 and SENSE1 with a dense selection of local features.

For a majority of descriptors, limit the informative area to the salient region improve the results. CM is the only exception. These results match those of Chapter 2 when we study the impact of filtering local features with the saliency.
As shown in Figure 4.12, the improvements is made for negative and positive classes when using SENSE2 both for 3 and 2 classes. The previous conclusions about SIFT based descriptors remain valid. Note that even with SENSE2 neutral images are difficult to recognize.

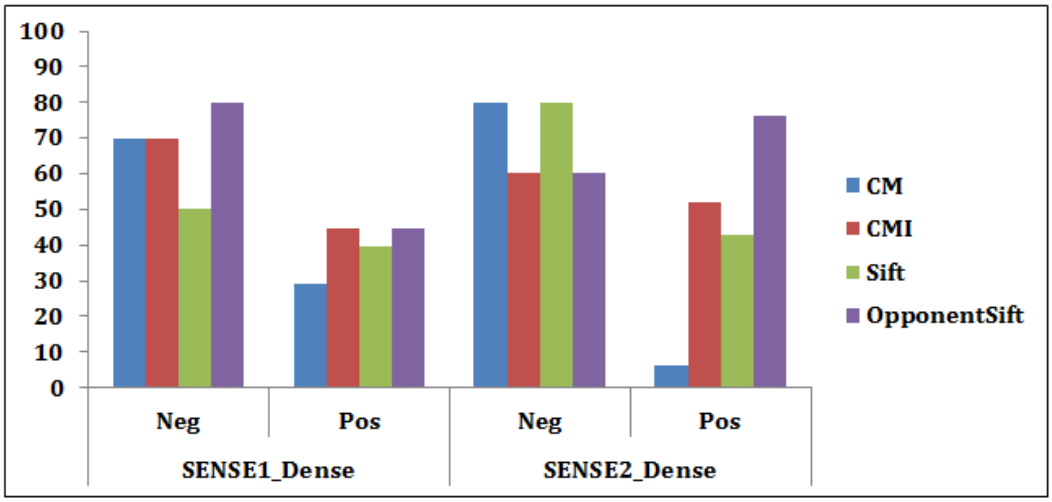
Summary

Extraction of emotional impact of images requires the consideration of many parameters. We modelled the most of them by the following attributes:

- Different colours in the images;



(a) 3 classes of emotions.



(b) 2 classes of emotions.

Figure 4.12: Average classification rates obtained for SENSE2 and SENSE1.

- Different textures;
- Image content with local descriptors.

In this chapter we propose three groups of criteria to describe emotions database for emotions study. They can briefly and efficiently indicated the inherent information about the image set, some interesting extrinsic information and also the different physiological evaluations available. These criteria are mainly based on the literature about the emotions study. We also build a new database of daily life images taking into account the weakness of the available datasets. This database was assessed with general concept. According to us, it is the best way to evaluate globally a "primary" emotion on low semantic databases. In addition to the conventional assessments, we propose to use visual saliency to reduce the semantic interpretation. Our results show that it can be a good alternative. In fact, by using a saliency model, 79% of the inconclusive images has a definitive class.

To finish we presented an evaluation of different features used in indexation for emotions prediction. We opted for a architecture based on content based image retrieval illustrated by Figure 4.13.

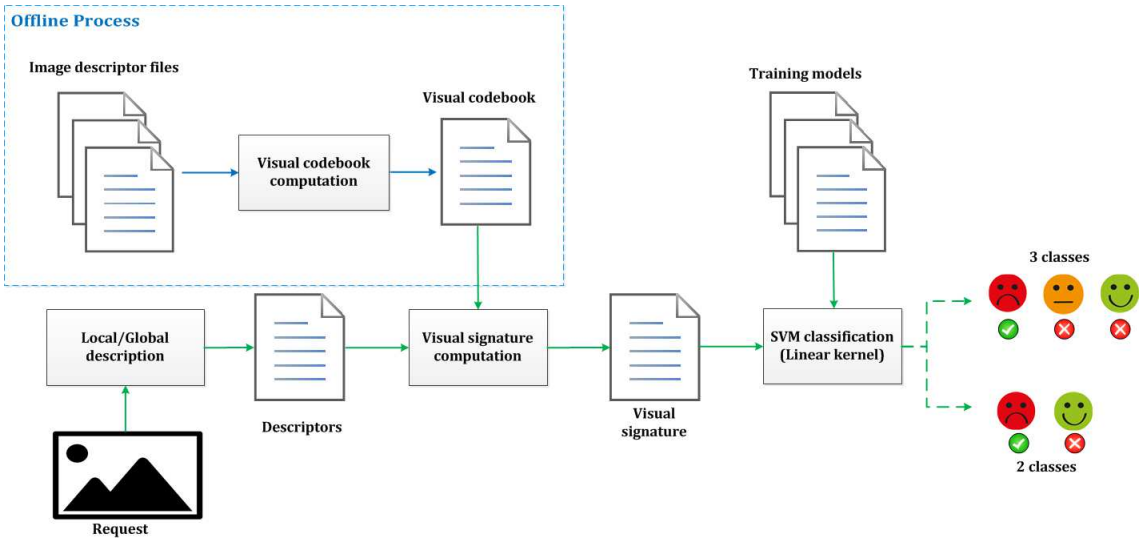


Figure 4.13: Our emotion recognition approach.

Due to the used method, different blocks can influence the results: the visual signature, the visual codebook for example. We have studied their impact and the results show that we cannot conclude to a unique behaviour regardless the descriptors and the database. However:

- The tests with different codebooks from two different databases confirm that emotions can be resumed with a finite images and they can be used for emotions prediction on different database. Note that using finite images as reference need to be coherent with the goals of application. For example, a reference images for natural images are not suitable for emotions detection on faces;
- VLAD representation seems to be adequate for local descriptors and BoVW for global descriptors;
- The chosen descriptors are complementary and the results on IAPS are hopeful;
- The usage of the regions of interest obtained with visual saliency model improves the results for positive and negative images especially for SIFT and OpponentSIFT: respectively $\sim +6\%$ and $+10\%$ for 3 and 2 classes.

Our studies have the distinction of having been made on a new and low semantic database. We tested our algorithm on IAPS in order to make efficient comparison with the other works in this domain. The results on our database and on IAPS are really relevant and they confirm our hypothesis that low level features could encode high level information interesting for emotions is justified regarding the interesting results we obtained.

Conclusion and perspectives

Conclusion

The presented results are based on tradionnal tools for CBIR which were evaluated in this rapport:

- Local and global descriptors;
- Visual codebook;
- BoVW and VLAD visual signatures.

We have introduced a new algorithm for dictionary computation based on a random visual word selection combined to an iterative process. This solution is very effective when using BoVW for very small sizes of codebook (≤ 256) compared to *K-Means*. In our work the results of CMI (24 dimensions) are equivalent to those of SIFT (128 dimensions) on UKB.

Throughout this manuscript, the contribution of saliency was assessed at different stages of the process from the detection of local features to image evaluation for emotion recognition. Regarding the detection of local features, we noticed that very few detectors, among the four most commonly used in the literature we reviewed, produced salient local features. Despite this the salient local features detected with Harris-Laplace are important for the retrieval accuracy. In fact on UKB, removing 20% of the most salient local features impact the results by -25%. Visual saliency has also show a interest for emotion recognition. For this high level task, we introduced a new database widely assessed. Visual saliency has allowed us to improve the evaluation of our image database by reducing the semantic interpretation. Our results show that it is useful as long as the size of the observed region is not too small. This implies that the object/scene should be recognizable. The classification results of the regions of interest obtained with saliency model are equivalent or better depending the descriptors.

Perspectives

We studied the visual saliency of the detectors and conclude that they do not produce salient local features. Also for Harris-Laplace we have shown that the less salient local features are not the most important for the retrieval accuracy. The

first perspective of these works has been started by replacing the less salient local features by the most ones from the dense detection. The results are hopeful and we think that we can study the impact of replacing a certain proportion of less salient point by the most salient ones of the images. The threshold of saliency must be defined function to the image and can differ from their content.
We shown that visual saliency can be useful for CBIR and emotional impact recognition. But we consider a bottom-up saliency models because of the reduced number of Top-down models. Another perspective could be to study the impact of this kind of attention modelling because of the high level aspect of the different tasks.
For SENSE2, we used a bounding box of the different salient areas, we think that a more precise region definition must be studied: defining different regions of interest by image and determine the emotion of each region. The final emotion of the image could be a combination of the negative and positive areas thereby resuming the idea of the harmony of a multi-coloured image from Solli et al. [Solli 09]. The fusion method could be found based on subjective evaluations to find the correct weighting between negative and positive "patches" to form the final emotional impact.

Bibliography

Bibliography

[Abdel-Hakim 06] 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. Page 9

[Bay 06] H. Bay, T. Tuytelaars & L. Van Gool. *SURF: Speeded Up Robust Features*. vol. 3951 of *Lecture Notes in Computer Science*, pages 404–417. Springer Berlin Heidelberg, 2006. Pages 9 and 11

[Beke 08] 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, vol. 33, no. 5, pages 381–394, 2008. Page 42

[Beresniak 90] D. Beresniak. Abc des couleurs leurs incidences dans votre vie quotidienne. 1990. Page 1

[Borji 13] A. Borji, D. Sihite & L. Itti. *Quantitative Analysis of Human-Model Agreement in Visual Saliency Modeling: A Comparative Study*. IEEE Transactions on Image Processing, vol. 22, no. 1, pages 55–69, 2013. Page 13

[Boyatziz 93] C.J. Boyatziz & R. Varghese. *Children's Emotional Associations With Colors*. The Journal of Genetic Psychology, vol. 155, pages 77–85, 1993. Pages 1 and 42

[Bradley 01] M. M. Bradley, M. Codispoti, D. Sabatinelli & P. J. Lang. *Emotion and Motivation II: Sex Differences in Picture Processing*. Emotion, vol. 1, no. 3, pages 300–319, 2001. Page 42

[Busso 04] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, U. Neumann & S. Narayanan. *Analysis of Emotion Recognition Using Facial Expressions, Speech and Multimodal Information*. In Proceedings of the 6th International Conference on Multimodal Interfaces, pages 205–211. ACM, 2004. Page 1

[Csurka 04] G. Csurka, C. Bray, C. Dance & L. Fan. *Visual categorization with bags of keypoints.* Workshop on Statistical Learning in Computer Vision, ECCV, pages 1–22, 2004. Pages 12 and 19

[De Silva 97] L. C. De Silva, T. Miyasato & R. Nakatsu. *Facial emotion recognition using multi-modal information.* In Proceedings of International Conference on Information, Communications and Signal Processing, vol. 1, pages 397–401, Sept. 1997. Page 1

[Demanet 09] L. Demanet L.and Ying. *Wave atoms and time upscaling of wave equations.* Numerische Mathematik, vol. 113, pages 1–71, 2009. Page 56

[Douze 09] M. Douze, H. Jégou, H. Sandhawalia & C. Amsaleg L.and Schmid. *Evaluation of GIST Descriptors for Web-scale Image Search.* In Proceedings of the ACM International Conference on Image and Video Retrieval, pages 1–19. ACM, 2009. Pages 2 and 8

[Ekman 92] P. Ekman. *Facial expressions of emotions.* Psychological science, vol. 3, no. 1, pages 34–38, 1992. Page 1

[Everingham 07] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn & A. Zisserman. *The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results*, 2007. Page 18

[Everingham 12] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn & A. Zisserman. *The PASCAL Visual Object Classes Challenge 2012 (VOC2012)Results*, 2012. Pages 2, 18, and 28

[Farquhar 05] J. Farquhar, S. Szedmak, H. Meng & J. Shawe-Taylor. *Improving "bag-of-keypoints" image categorisation: Generative Models and PDF-Kernels.* PASCAL Eprint Series, 2005. Page 12

[Fernandez-Maloigne 12] C. Fernandez-Maloigne. Advanced color image processing and analysis. Springer, July 2012. Page 56

[Friman 07] O. Friman, I. Volosyak & A. Graser. *Multiple Channel Detection of Steady-State Visual Evoked Potentials for Brain-Computer Interfaces.* Biomedical Engineering, IEEE Transactions on, vol. 54, no. 4, pages 742–750, 2007. Page 47

[Gao 08a] K. Gao, S. Lin, Y. Zhang, S. Tang & H. Ren. *Attention Model Based SIFT Keypoints Filtration for Image*

Retrieval. In Proceedings of IEEE International Conference on Computer and Information Science, pages 191–196, 2008. Pages 13 and 14

[Gao 08b] K. Gao, S. Lin, Y. Zhang, S. Tang & H. Ren. *Attention Model Based SIFT Keypoints Filtration for Image Retrieval.* In Proceedings of the 7th IEEE/ACIS International Conference on Computer and Information Science, pages 191–196, May 2008. Page 14

[Gordoa 12] A. Gordoa, J. A. Rodriguez-Serrano, F. Perronnin & E. Valveny. *Leveraging category-level labels for instance-level image retrieval.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3045–3052, June 2012. Pages 9 and 31

[Harris 88] C. Harris & M. Stephens. *A Combined Corner and Edge Detector.* In Proceedings of the 4th Alvey Vision Conference, pages 147–151, 1988. Pages 2 and 9

[Hays 07] A. A. Hays J.and Efros. *Scene Completion Using Millions of Photographs.* In ACM SIGGRAPH Papers, 2007. Pages 2 and 8

[His] `https://sites.google.com/site/int3llig3nc3artifici3ll3/retrospective-de-l-histoire-de-l-ia`. Page 1

[Hong 06] S. Hong & H. Choi. *Color image semantic information retrieval system using human sensation and emotion.* In Issues in Information Systems, vol. 7, pages 140–145, 2006. Page 1

[Huiskes 08] M. J. Huiskes & M. S. Lew. *The MIR Flickr Retrieval Evaluation.* In Proceedings of the ACM International Conference on Multimedia Information Retrieval. ACM, 2008. Page 18

[Huiskes 10] M. J. Huiskes, B. Thomee & M. S. Lew. *New Trends and Ideas in Visual Concept Detection: The MIR Flickr Retrieval Evaluation Initiative.* In Proceedings of the ACM International Conference on Multimedia Information Retrieval, pages 527–536. ACM, 2010. Page 18

[Itti 98] L. Itti, C. Koch & E. Niebur. *A Model of Saliency-Based Visual Attention for Rapid Scene Analysis.* IEEE Trans. Pattern Anal. Mach. Intell., vol. 20, no. 11, pages 1254–1259, 1998. Pages 13, 14, 15, 20, 26, and 53

[Jégou 10] 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, pages 3304–3311. IEEE Computer Society, 2010. Pages 2, 10, 12, 18, 25, and 26

[Jégou 12] H. Jégou, F. Perronnin, M. Douze, J. Sanchez, P. Perez & C. Schmid. *Aggregating Local Image Descriptors into Compact Codes.* Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 34, no. 9, pages 1704–1716, Sept. 2012. Pages 25 and 26

[Kaya 04] Naz Kaya & Helen H. Epps. *Color-Emotion associations: Past experience and personal preference.* AIC Colors and Paints, Interim Meeting of the International Color Association, 2004. Pages 1 and 42

[Ke 04] 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, pages 506–513, 2004. Pages 10 and 11

[Keil 03] A. Keil, T. Gruber, M.. Müller, S. Moratti, M. Stolarova, M. Bradley & P.J. Lang. *Early modulation of visual perception by emotional arousal: Evidence from steady-state visual evoked brain potentials.* Cognitive, Affective, & Behavioral Neuroscience, vol. 3, no. 3, pages 195–206, 2003. Page 47

[Kemp 02] A. H. Kemp, M. A. Gray, P. Eide, R. B. Silberstein & P. J. Nathan. *Steady-State Visually Evoked Potential Topography during Processing of Emotional Valence in Healthy Subjects.* NeuroImage, vol. 17, no. 4, pages 1684–1692, 2002. Page 47

[Kootstra 11] G. Kootstra, B. de Boer & L. Schomaker. *Predicting Eye Fixations on Complex Visual Stimuli Using Local Symmetry.* Cognitive Computation, vol. 3, no. 1, pages 223–240, 2011. Page 28

[Lang 08] 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.* Rapport technique, University of Florida, 2008. Pages 2, 43, 44, and 47

[Le Meur 06a] O. Le Meur, P. Le Callet, D. Barba & D. Thoreau. *A coherent computational approach to model bottom-up visual attention.* Pattern Analysis and Machine Intelligence,

IEEE Transactions on, vol. 28, no. 5, pages 802–817, May 2006. Page 13

[Le Meur 06b] O. Le Meur, P. Le Callet, D. Barba & D. Thoreau. *A coherent computational approach to model bottom-up visual attention.* Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 28, no. 5, pages 802–817, May 2006. Page 28

[Li 08] X. Li, C. Wu, C. Zach, S. Lazebnik & J. Frahm. *Modeling and Recognition of Landmark Image Collections Using Iconic Scene Graphs.* In Proceedings of the 10th European Conference on Computer Vision: Part I, pages 427–440. Springer-Verlag, 2008. Pages 2 and 8

[Liu 08] W. Liu, W. Xu & L. Li. *A tentative study of visual attention-based salient features for image retrieval.* In Proceedings of the 7th World Congresson Intelligent Control and Automation, pages 7635–7639, June 2008. Page 14

[Liu 11] 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. Pages 1, 41, 43, 44, and 62

[Lowe 99] D. G. Lowe. *Object Recognition from Local Scale-Invariant Features.* International Conference on Computer Vision, vol. 2, pages 1150–1157, 1999. Pages 2 and 10

[Lowe 04] D. G. Lowe. *Distinctive Image Features from Scale-Invariant Keypoints.* International Journal of Computer Vision, vol. 60, pages 91–110, 2004. Page 10

[Lucassen 10] M. P. Lucassen, T. Gevers & A. Gijsenij. *Adding texture to color: quantitative analysis of color emotions.* In Proceedings of CGIV, 2010. Page 42

[Machajdik 10] J. Machajdik & A. Hanbury. *Affective image classification using features inspired by psychology and art theory.* In Proceedings of the international conference on Multimedia, pages 83–92, 2010. Pages 1, 2, 43, 44, 46, 47, 48, and 62

[Matas 02] J. Matas, O. Chum, M. Urban & T. Pajdla. *Robust Wide Baseline Stereo from Maximally Stable Extremal Regions.* In Proceedings of the British Machine Vision Conference, pages 1–10, 2002. Page 2

[Mikels 05] J. A. Mikels, B. L. Fredrickson, G. R. Larkin, C. M. Lindberg, S. J. Maglio & P. A. Reuter-Lorenz. *Emotional category data on images from the international affective picture system.* Behavior Research Methods, vol. 37, no. 4, pages 626–630, 2005. Page 43

[Mikolajczyk 01] K. Mikolajczyk & C. Schmid. *Indexing based on scale invariant interest points.* In Proceedings of the 8th IEEE International Conference on Computer Vision, vol. 1, pages 525–531, 2001. Pages 2, 9, and 10

[Mikolajczyk 02] K. Mikolajczyk & C. Schmid. *An Affine Invariant Interest Point Detector.* In Computer Vision-ECCV, vol. 2350 of *Lecture Notes in Computer Science*, pages 128–142. Springer Berlin Heidelberg, 2002. Page 10

[Mikolajczyk 05a] K. Mikolajczyk & C. Schmid. *A performance evaluation of local descriptors.* IEEE Transactions on Pattern Analysis & Machine Intelligence, vol. 27, no. 10, pages 1615–1630, 2005. Page 11

[Mikolajczyk 05b] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir & L. Van Gool. *A Comparison of Affine Region Detectors.* Int. J. Comput. Vision, vol. 65, no. 1-2, pages 43–72, 2005. Page 9

[Mindru 04] F Mindru, T. Tuytelaars, L Van Gool & T. Moons. *Moment invariants for recognition under changing viewpoint and illumination.* Computer Vision and Image Understanding, vol. 94, no. 1–3, pages 3–27, 2004. Page 11

[Moravec 77] H. P. Moravec. *Towards Automatic Visual Obstacle Avoidance.* In Proceedings of the 5th International Joint Conference on Artificial Intelligence, vol. 2, pages 584–584. Morgan Kaufmann Publishers Inc., 1977. Page 9

[Nistér 06] D. Nistér & H. Stewénius. *Scalable Recognition with a Vocabulary Tree.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, pages 2161–2168, 2006. Pages 2, 8, 10, 18, and 28

[Oliva 01] A. Oliva & A. Torralba. *Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope.* International Journal of Computer Vision, vol. 42, pages 145–175, 2001. Pages 2, 8, and 57

[Ou 04a] 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, vol. 29, no. 3, pages 232–240, 2004. Pages 1 and 42

[Ou 04b] 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, vol. 29, no. 4, pages 292–298, 2004. Pages 1 and 42

[Ou 04c] 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, vol. 29, no. 5, pages 381–389, 2004. Pages 1 and 42

[Ou 06] L. C. Ou & M. R. Luo. *A colour harmony model for two-colour combinations.* Color Research & Application, vol. 31, no. 3, pages 191–204, 2006. Page 1

[Ou 11] L. C. Ou, P. Chong, M. R. Luo & C. Minchew. *Additivity of colour harmony.* Color Research & Application, vol. 36, no. 5, pages 355–372, 2011. Page 1

[Paleari 08] M. Paleari & B. Huet. *Toward emotion indexing of multimedia excerpts.* Proceedings on Content-Based Multimedia Indexing, International Workshop, pages 425–432, 2008. Page 42

[Parsons 04] L. Parsons, E. Haque & H. Liu. *Subspace clustering for high dimensional data: a review.* In Proceedings of the ACM SIGKDD, vol. 6, pages 90–105. Explorations Newsletter, 2004. Pages 2 and 19

[Perreira Da Silva 10] M. Perreira Da Silva, V. Courboulay, A. Prigent & P. Estraillier. *Evaluation of preys/predators systems for visual attention simulation.* In Proceedings of the International Conference on Computer Vision Theory and Applications, pages 275–282. INSTICC, 2010. Pages 52 and 53

[Perronnin 06] F. Perronnin, C. Dance, G. Csurka & M Bressan. *Adapted vocabularies for generic visual categorization.* In Proceedings of the ECCV, pages 464–475, 2006. Page 12

[Perronnin 07] 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. Pages 2, 12, and 13

[Perronnin 08] F. Perronnin. *Universal and Adapted Vocabularies for Generic Visual Categorization.* Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 30, no. 7, pages 1243–1256, July 2008. Pages 9 and 31

[Rosten 05] E. Rosten & T. Drummond. *Fusing points and lines for high performance tracking.* In Proceedings of the IEEE International Conference on Computer Vision, vol. 2, pages 1508–1511, Oct. 2005. Page 10

[Rosten 06] E. Rosten & T. Drummond. *Machine learning for high-speed corner detection.* In Proceedings of the European Conference on Computer Vision, vol. 1, pages 430–443, May 2006. Page 10

[Sander 13] D. Sander. *Vers une définition de l'émotion.* Cerveau&Psycho, no. 56, 2013. Page 1

[Scherer 84] K. R. Scherer & P. Ekman. Approaches to emotions. Lavoisier, Jan. 1984. Page 1

[Schmid 00] C. Schmid, R. Mohr & C. Bauckhage. *Evaluation of Interest Point Detectors.* Int. J. Comput. Vision, vol. 37, no. 2, pages 151–172, 2000. Page 9

[Sivic 03] J. Sivic & A. Zisserman. *Video Google: A Text Retrieval Approach to Object Matching in Videos.* In Proceedings of the International Conference on Computer Vision, pages 1470–1477, 2003. Pages 2 and 12

[Smith 97] S. M. Smith & J. M. Brady. *SUSAN&Mdash;A New Approach to Low Level Image Processing.* Int. J. Comput. Vision, vol. 23, no. 1, pages 45–78, May 1997. Page 10

[Solli 09] M. Solli & R. Lenz. *Color harmony for image indexing.* In Proceedings of the 12th International Conference on Computer Vision Workshops, pages 1885–1892, Sept. 2009. Pages 1 and 68

[Solli 10] M. Solli & R. Lenz. *Emotion Related Structures in Large Image Databases.* In Proceedings of the ACM International Conference on Image and Video Retrieval, pages 398–405. ACM, 2010. Pages 1 and 43

[Swain 91] M. J. Swain & D. H. Ballard. *Color indexing.* International Journal of Computer Vision, vol. 7, pages 11–32, 1991. Pages 2 and 8

[Tomkims 62] S. S. Tomkims. Affect imagery consciousness: The positive affects, vol. 1. Springer Publishing Company, 1962. Page 1

[Turing 50] A. M. Turing. *Computing Machinery and Intelligence*, 1950. Page 1

[Tuytelaars 08] T. Tuytelaars & K. Mikolajczyk. *Local Invariant Feature Detectors: A Survey.* Foundations and Trends in Computer Graphics and Vision, vol. 3, no. 3, pages 177–280, 2008. Page 9

[van de Sande 10] 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, vol. 32, no. 9, pages 1582–1596, 2010. Pages 9, 10, 13, 19, and 29

[Wang 05] 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, pages 4571–4576, 2005. Pages 1 and 43

[Wang 13] S. Wang, G. Wu & Y. Zhu. *Analysis of Affective Effects on Steady-State Visual Evoked Potential Responses.* In Intelligent Autonomous Systems, vol. 194 of *Advances in Intelligent Systems and Computing*, pages 757–766. Springer Berlin Heidelberg, 2013. Page 47

[Wei 08] K. Wei, B. He, T. Zhang & W. He. Image emotional classification based on color semantic description, vol. 5139 of *Lecture Notes in Computer Science*, pages 485–491. Springer Berlin / Heidelberg, 2008. Pages 42 and 62

[Yanulevskaya 08] 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, pages 101–104, 2008. Pages 1, 43, 44, and 46

[Zdziarski 12] Z. Zdziarski & R. Dahyot. *Feature selection using visual saliency for content-based image retrieval.* In Proceedings of the IET Irish Signals and Systems Conference, pages 1–6, 2012. Page 14

[Zhang 08] L. Zhang, M. H. Tong, T. K. Marks, H. Shan & G. W. Cottrell. *SUN: A Bayesian framework for saliency using natural statistics.* J Vis, vol. 8, no. 7, pages 1–20, 2008. Page 13

List of publications

List of publications

National journals

- S. Gbèhounou, F. Lecellier, C. Fernandez-Maloigne, Extraction et analyse de l'impact émotionnel des images, *Traitement de Signal num. 3-4-5/2012, p. 409-432.*

International Conferences

- T. Urruty, S. Gbèhounou, H. T. Le, J. Martinet, C. Fernandez-Maloigne, Iterative Random Visual Word Selection *4th International Conference on Multimedia Retrieval, 1-4 April 2014.*

- S. Gbèhounou, F. Lecellier, C. Fernandez-Maloigne, V.Courboulay, Can salient interest regions resume emotional impact of an image?, *15th International Conference on Computer Analysis of Images and Patterns, 27-29 August 2013, LNCS 8047, p. 515.*

- S. Gbèhounou, F. Lecellier, C. Fernandez-Maloigne, Gender influences on subjective evaluations in image, *12th International AIC Colour Congress, 8-12 Juillet 2013.*

- S. Gbèhounou, F. Lecellier, C. Fernandez-Maloigne, Extraction of emotional impact in colour images, *CGIV 2012, Vol. 6, Society for Imaging Science and Technology, 2012, p. 314-319.*

National Conferences

- S. Gbèhounou, F. Lecellier, C. Fernandez-Maloigne, Extraction et analyse de l'impact é motionnel des images, $18^{ème}$ *Congrès francophone sur la Reconnaissance des Formes et l'Intelligence Artificielle, 24-27 Janvier 2012.*

National presentations

- S. Gbèhounou, F. Lecellier, C. Fernandez-Maloigne, V.Courboulay, Les régions saillantes améliorent-elles l'évaluation de l'impact émotionnel des images?, *GDR ISIS, 26 Septembre 2013, Paris.*

- S. Gbèhounou, F. Lecellier, C. Fernandez-Maloigne, V.Courboulay, Extraction et analyse de l'impact émotionnel des images, *Séminaire École Doctorale S2IM, 10-12 Avril 2013, Poitiers.*

Indexation de bases d'images : Évaluation de l'impact émotionnel

Résumé: L'objectif de ce travail est de proposer une solution de reconnaissance de l'impact émotionnel des images en se basant sur les techniques utilisées en recherche d'images par le contenu. Nous partons des résultats intéressants de cette architecture pour la tester sur une tâche plus complexe. La tâche consiste à classifier les images en fonction de leurs émotions que nous avons définies "Négative", "Neutre" et "Positive". Les émotions sont liées aussi bien au contenu des images, qu'à notre vécu. On ne pourrait donc pas proposer un système de reconnaissance des émotions performant universel. Nous ne sommes pas sensible aux mêmes choses toute notre vie: certaines différences apparaissent avec l'âge et aussi en fonction du genre. Nous essaierons de nous affranchir de ces inconstances en ayant une évaluation des bases d'images la plus hétérogène possible. Notre première contribution va dans ce sens: nous proposons une base de 350 images très largement évaluée. Durant nos travaux, nous avons étudié l'apport de la saillance visuelle aussi bien pendant les expérimentations subjectives que pendant la classification des images. Les descripteurs, que nous avons choisis, ont été évalués dans leur majorité sur une base consacrée à la recherche d'images par le contenu afin de ne sélectionner que les plus pertinents. Notre approche qui tire les avantages d'une architecture bien codifiée, conduit à des résultats très intéressants aussi bien sur la base que nous avons construite que sur la base IAPS, qui sert de référence dans l'analyse de l'impact émotionnel des images.

Mots-clés: Recherche d'images par le contenu, Sac de mots visuels, impact émotionnel des images, saillance visuelle, évaluations subjectives

Image databases indexing: Emotional impact assessing

Abstract: The goal of this work is to propose an efficient approach for emotional impact recognition based on CBIR techniques (descriptors, image representation). The main idea relies in classifying images according to their emotion which can be "Negative", "Neutral" or "Positive". Emotion is related to the image content and also to the personnal feelings. To achieve our goal we firstly need a correct assessed image database. Our first contribution is about this aspect. We proposed a set of 350 diversifed images rated by people around the world. Added to our choice to use CBIR methods, we studied the impact of visual saliency for the subjective evaluations and interest region segmentation for classification. The results are really interesting and prove that the CBIR methods are usefull for emotion recognition. The chosen desciptors are complementary and their performance are consistent on the database we have built and on IAPS, reference database for the analysis of the image emotional impact.

Keywords: Content Based Image Retrieval, Bag of Visual Words, image emotional impact, visual saliency, subjective evaluations

Doctorat de l'Université de Poitiers, Spécialité: Traitement du Signal et des images

Thèse préparée et soutenue au Département SIC du Laboratoire XLIM, UMR 7252
Université de Poitiers, Bât. SP2MI, Téléport 2, Bvd Marie et Pierre Curie
BP 30179, 86962 Futuroscope Chasseneuil Cedex France