# Classification multilingue et multimédia pour la recherche d'images dans le projet OMNIA

David Rouquet, Achile Fallaise, Didier Schwab, Hervé Blanchon, Vallérie Bellynck, Christian Boitet, Emmanuel Dellandréa, Ningning Liu, Liming Chen, Alexandre Saidi, Sandra Skaff, Luca Marchesotti, and Gabriela Csurka

XRCE, LIRIS, LIG-GETALP
http://www.omnia-project.org

**Abstract.** Cet article expose différents traitements pour l'indexation automatique d'images accompagnées de textes multilingues. La combinaison de ces traitements a pour but d'obtenir des descripteurs multi-facettes d'images (thématique, affectif, esthétique, etc.). Cette approche est évaluée dans le cadre du projet ANR OMNIA qui rassemble des équipes de XRCE, du LIRIS et du LIG.

**Key words:** classification, sémantique, multimédia, multilingue, images, masse de données

## 1 Introduction

Un des buts du projet OMNIA, financé par l'ANR de 2008 à 2011, est l'indexation automatique d'images, accompagnées de courts textes en langue naturelle, issues de grands entrepôts de données sur le web. Un aspect original est la prise en compte d'attributs affectifs et émotionnels. L'indexation est réalisée grâce aux résultats de différentes analyses visuelles des images, ainsi qu'au traitement des textes multilingues accompagnant les images. Cet article présente les traitements successifs que subissent une image et ses textes compagnons dans le système OMNIA.

Dans une première partie, nous décrivons le Classificateur Visuel Générique et l'identification de caractéristiques esthétiques développés au laboratoire XRCE. Nous verrons ensuite les travaux du LIRIS qui permettent l'utilisation de descripteurs visuels pour attribuer une émotion dans un modèle dimentionnel. Nous présentons enfin les travaux de l'équipe GETALP du LIG pour l'extraction de contenu dans des textes multilingues. Le défi que présente la fusion des résultats de tels processus hétérogène dans un descripteur unique et les travaux en cours sont décrits dans la conclusion.

## 2 XRCE : Analyse des images pour la classification thématique et esthétique

Nous présentons d'abord les résultats obtenus au XRCE. Les traitements ont été testés sur la base MIRFLICKR [1]. Les images sont classées selon leur contenu et leurs attributs esthétiques.

### 2.1 L'analyse contextuelle

La Catégorisation Visuelle Generique (GVC - Generic Visual Categorization), est un processus qui catégorise automatiquement les images dans un ensemble discret de classes sémantiques. Les classes pourraient être intérieur ou extérieur, naturel ou artificiel, plages ou couchers de soleil ou montagnes. On attribue à l'image des étiquettes (labels) correspondant aux objets ou concepts présents dans l'image. La GVC est un problème multi-classe, ce qui signifie que plusieurs étiquettes peuvent être attribuées à la même image. En ce sens, la GVC est souvent appelé annotation automatique d'image, car les catégories pertinentes peuvent être considérées comme des annotations attribuées automatiquement (tags, labels). Contrairement à la plupart des méthodes de détection et de reconnaissance qui sont spécifiquement développées pour une classe donnée (ex. détection de visage), la technologie GVC est indépendante des classes ou des types d'objets. Elle peut donc être qualifiée de générique et applicable sans modification spécifique des paramètres à des categories très variées comme des classes d'objets, des scènes ou des événements, des peintures, etc.

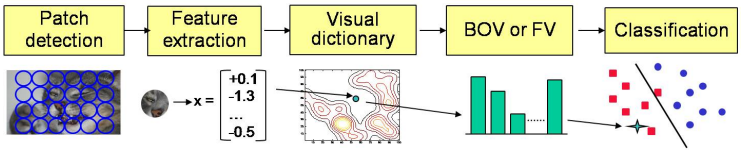


**Fig. 1.** Generic Visual Categorization (GVC) - pipeline

L'approche sac de mots visuels (BOV - Bag Of Visual words) (la figure 1) est constituée par les étapes suivantes :

**Détection de patch :** Premièrement, les régions d'intérêt dans l'image (patches) sont détectées.

**Extraction de caractéristiques :** A partir de chaque patch, les caractéristiques sont extraites. Elles peuvent être ou ne pas être invariantes par transformation géométrique simple (translation, rotation, etc.).

**Vocabulaire visuel :** Toutes les caractéristiques extraites sont mappées à l'espace de caractéristiques et regroupés pour obtenir le vocabulaire visuel.

[1] http://press.liacs.nl/mirickr/

Souvent, un K-means simple est utilisé [Csurka 04], mais les Gaussian Mixture Models [Perronnin 06] peuvent également être utilisés pour obtenir une classification douce.

**Estimation d'histogramme :** Un sac de mots visuels est construit en comptant le nombre de patches assignés à chaque groupe: chaque patch est attribué au mot visuel le plus proche ou à tous les mots visuels de manière probabiliste dans le cas d'un modèle de vocabulaire visuel stochastique. L'histogramme est calculé en accumulant les occurrences de chaque mot visuel.

**Classement :** L'histogramme ainsi obtenu peut être vue comme un vecteur de caractéristiques haut niveau de l'image et classifié par une série de OVA (one-versus-all) classificateurs (e.g. SVMs comme dans [Csurka 04] une ou multi classificateurs multi classes KNN [Bosch 06], pLSA [Bosch 06], pour déterminer quelles catégories à attribuer à l'image.

En résumé, l'entrée de l'analyseur visuel est une image et la sortie est un ensemble de mots-clés (les concepts d'ontologie de OMNIA) associés à des degrés de confiance. Ces degrés représentent la vraisemblance qu'un concept soit effectivement représenté dans l'image.

### 2.2 L'étape d'annotation

Nous avons testé avec succés notre système GVC dans plusieurs compétitions internationales tel que Pascal VOC [Everingham 06] ou ImageCLEF09 Large Scale Visual Concept Detection and Photo Annotation Task [2]. Dans le démonstrateur de ce papier nous avons utilisé les 53 concepts de cette dernière compétition.

Ces concepts sont organisés dans une ontologie comme présentée dans la Figure 2. Une partie de la base MIRFLICKR a été renommée et manuellement annotée par plusieurs annotateurs. La compétition consistait à entrainer les classificateurs sur les 5000 images d'entrainement et à les tester sur les 13000 images de test. Le but était de décider pour chaque image quels concepts sont présents et absents tout en assurant une consistance avec les ontologies.

Nous avons entrainé notre système GVC d'une manière OVA (One Versus All) non hiérarchique, puis pour assurer les constraintes ontologique, nous avons post-traités les degrés de confiances afin d'assurer ces contraintes.

La Figure 3 montre les résultats de la compétition. Deux types de mésures sont utilisés pour évaluer les systèmes. D'un côté les mesures classiques comme EER (Equal Error Rate) et AUC (Area under Curve) mesurent la performance de chaque classificateur individuellement. De l'autre côté, la mesure hiérarchique proposée évalue la performance d'auto annotation des images en considérant l'accord entre les annotations manuelles et les annotations automatiques. Cette dernière mesure est calculée sur chaque image test et moyenne à la fin. Comme le montre la Figure 3 notre système obtient la meilleur selon la mesure de EER.

Dans notre démonstrateur, nous avons appliqué ces classificateurs sur l'ensemble des images MIRFLICKR pour obtenir les annotations de ces images.

[2] http://www.imageclef.org/2009

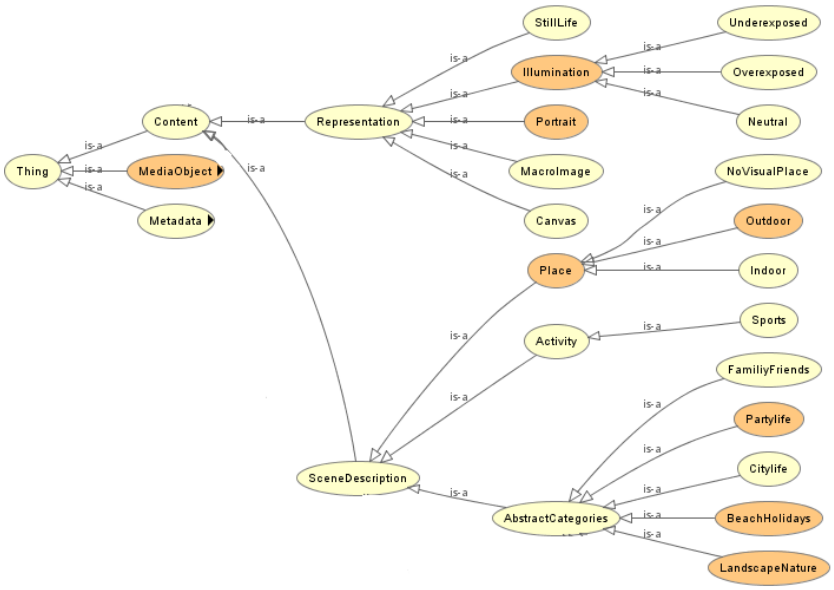


**Fig. 2.** Extrait de l'ontologie des concepts

En plus de la classification basée sur le contenu que nous venons de décrire, nous considérons les aspects esthétiques suivant : luminosité, contraste, netteté, teinte et taille.

## 3 LIRIS : Reconnaissance de la sémantique émotionnelle portée par les images

### 3.1 Introduction

Un des buts de l'informatique, et particulièrement de l'intelligence artificielle est d'élaborer des ordinateurs intelligents qui ont la capacité d'interagir avec des êtres humains de façon naturelle. Dès lors, une des questions essentielles est de permettre aux ordinateurs de reconnaitre, de comprendre et d'exprimer des émotions [Picard 97]. Plusieurs travaux ont été faits depuis plusieurs années sur ces aspects en informatique mais également en robotique. Quand il s'agit de reconnaitre des émotions (voir [Zeng 09] pour un tour d'horizon très complet), les recherches portent principalement sur la reconnaissance d'affects dans des données audio (parole ou musique) et sur la reconnaissance visuelle d'expressions faciales. Très peu de contributions traitent de la reconnaissance de la sémantique émotionnelle portée globalement par les images que ce soit par ses couleurs, sa composition ou tout autre élément qui peut provoquer une émotion. Face à ce sujet de recherche émergeant, un grand nombre de questions doivent être abordées

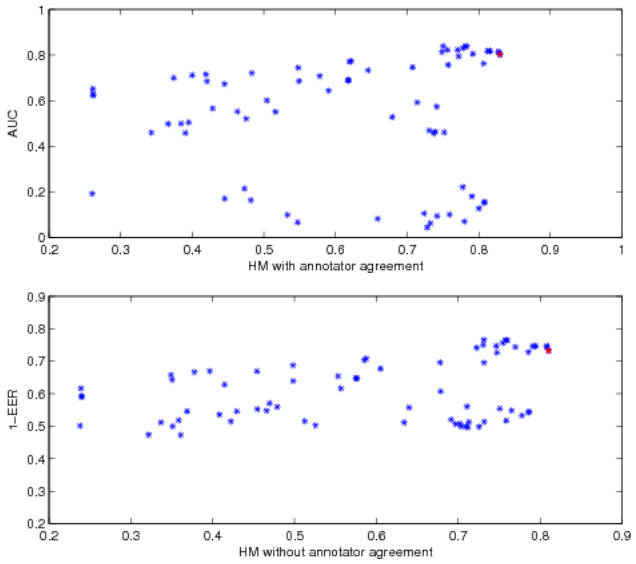


**Fig. 3.** En rouge: notre resultats; en bleu: les resultats des autre systèms participants. Les mesures AUC et EER contre HM (Hierarchical Measure). Le dernier est une mesure hiérarchique devloppée par [ImagCLEF ].

concernant principalement les trois problèmes suivants : la représentation des émotions, l'extraction de caractéristiques visuelles nécessaire à la reconnaissance des émotions et les modèles de classification pour traiter les différentes propriétés des émotions [Wang 05,Weining 06,Wang 08]. En effet, comme dans tous les autres problèmes de vision par ordinateur, la principale difficulté consiste à franchir le fossé sémantique qui existe entre les descripteurs bas-niveau extraits des images et les concepts sémantiques de haut-niveau qui sont dans notre cas les émotions.

Dans cet section, nous nous proposons d'étudier l'efficacité de différents types de descripteurs visuels ainsi que les classificateurs nécessaires à la reconnaissance d'émotions dans les images. De plus, nous proposerons d'utiliser la théorie des fonctions de croyance de Dempster-Shafer [Dempster 68,Smets 90], qui permet la manipulation de connaissances ambigües et incertaines comme celles relatives aux émotions.

### 3.2 Représentation des émotions

Plusieurs modèles ont été étudiés dans la littérature pour représenter les émotions [Zeng 09]. Les deux principales approches sont le modèle discret et le modèle dimensionnel. Le premier modèle consiste à choisir des noms ou des adjectifs pour décrire les émotions, tels que le bonheur, la tristesse, la peur, la colère, le dégoût

et la surprise. Le second modèle décrit les émotions selon une ou plusieurs dimensions où chacune représente une caractérisation de l'émotion, les plus utilisées étant l'appréciation, l'activité ou le contrôle. Ce deuxième modèle permet de représenter un plus large éventail d'émotions que le premier.

Le choix de la représentation émotionnelle est généralement guidé par l'application. Ainsi, les deux approches sont utiles et peuvent même être combinées, car elles peuvent apporter des informations complémentaires. Dans cet section, nous proposons une représentation hybride comme l'illustre la figure 4. Chaque image est ainsi représentée comme un point de l'espace constitué des deux dimensions que sont l'appréciation (variant de très déplaisante à très plaisante) et l'activité (variant de très calme à très dynamique). Cet espace est divisé en quatre quadrants permettant d'obtenir quatre types d'émotions distinctes afin de caractériser la charge émotionnelle de chaque image.

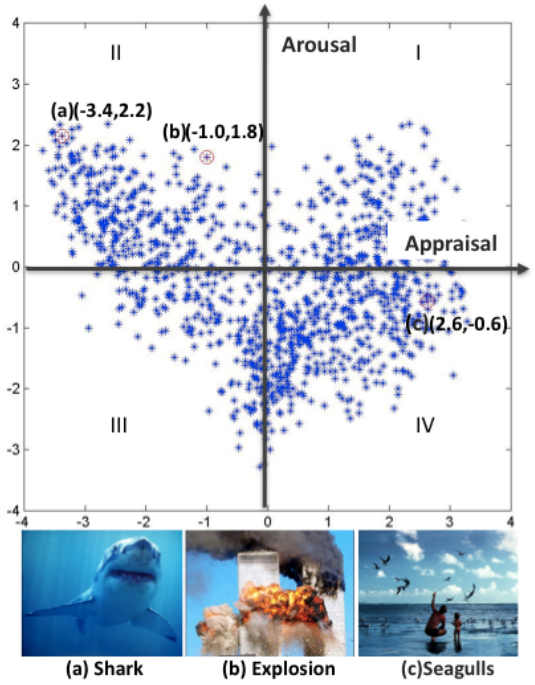


**Fig. 4.** Représentation des images de la base IAPS selon des critères d'activité (arousal) et d'appréciation (appraisal).

### 3.3 Descripteurs d'images pour la reconnaissance des émotions

L'extraction des caractéristiques propres d'une image est une question clé pour la reconnaissance de concepts dans des images, et en particulier, les émotions.

Ces caractéristiques doivent porter les informations nécessaires pour permettre la reconnaissance des différents concepts. Comme la reconnaissance des émotions dans les images est un domaine de recherche émergent, très peu de travaux ont été réalisés pour identifier les caractéristiques de l'image qui sont les plus efficaces dans ce contexte.

**Descripteurs d'images traditionnels** La plupart des travaux traitant de la reconnaissance des émotions utilisent les mêmes descripteurs que ceux généralement exploités pour d'autres problèmes de vision par ordinateur. Les trois principales catégories de descripteurs d'images sont basées sur la couleur, la texture et la forme. En ce qui concerne la couleur, des études ont montré que l'espace HSV (Hue, Saturation, Value) est un espace de couleur qui est mieux adapté à la perception réelle des couleurs par l'homme que d'autres espaces tels que l'espace RGB traditionnel. Ainsi, sur la base de cet espace de couleur, plusieurs façons de décrire le contenu couleur des images peuvent être considérées tels que les moments de couleurs, les corrélogrammes et histogrammes de couleur ainsi que les histogrammes relatifs à la température de la couleur [Dunker 09,Li 07]. En ce qui concerne la texture, la principale caractéristique demeure les matrices de cooccurrences [Li 07,Wu 05]. Toutefois, les descripteurs de Tamura [Wu 05] peuvent également représenter une alternative intéressante. En effet, des descripteurs tels que la granularité, le contraste ou la directionnalité se sont avérés fortement corrélés avec la perception visuelle de l'homme. Enfin, la description des formes peut être envisagée grâce à l'extraction des contours permettant l'obtention de l'histogramme d'orientation des lignes [Columbo 99,Wu 05] ou encore les descripteurs de Haar [Dunker 09,Cho 02].

**Descripteurs d'images pour la reconnaissance des émotions** Certaines tentatives ont été faites pour identifier des descripteurs de plus haut-niveau liés aux émotions. En effet, les études sur les peintures ont mis en évidence la portée sémantique des couleurs et des lignes qui y apparaissent, comme cela est rappelé dans les travaux de [Columbo 99] où sont proposés des descripteurs d'images plus corrélés aux émotions grâce à l'exploitation de ces informations. Ainsi, en utilisant la théorie des couleurs d'Itten, une signification émotionnelle des couleurs peut être dégagée. Tout d'abord, comme mentionné plus haut, les couleurs sont décrites en terme de teinte, de luminance et de saturation grâce à l'espace de couleur HSV, afin de se rapprocher de la perception humaine des couleurs. Ces couleurs sont ensuite projetées sur un cercle chromatique, appelé cercle d'Itten où les couleurs fortement contrastées ont des coordonnées opposées par rapport au centre du cercle. Itten a montré que les combinaisons de couleurs peuvent produire des effets tels qu'une harmonie, une disharmonie, du calme ou de l'excitation. Ainsi, l'harmonie sera détectée sur le cercle d'Itten si les positions des couleurs connectées entre elles constituent un polygone régulier. Le descripteur correspondant à cette hypothèse est obtenu en mesurant la distance entre le centre du cercle d'Itten et le centre du polygone reliant les couleurs dominantes

de l'image. Ces dernières sont préalablement obtenues par un algorithme basé sur les k-means.

Les lignes portent également une information sémantique importante sur les images. En effet, des lignes obliques suggèrent le dynamisme et l'action tandis que les lignes horizontales ou verticales communiquent plutôt le calme et la détente. Pour exprimer cela en terme de descripteurs d'images, les lignes sont d'abord extraites grâce à une transformée de Hough, puis le rapport entre le nombre de lignes obliques et le nombre total de lignes dans une image est calculé.

### 3.4 Modèles de classification pour la reconnaissance des émotions

**Classificateurs traditionnels** La plupart des travaux traitant de la classification des émotions dans les images reposent sur des approches traditionnelles de classification largement utilisée dans d'autres problèmes de vision par ordinateur. Malheureusement, elles ne sont pas toujours appropriées pour traiter de la spécificité des émotions. Parmi ces approches, on peut citer les réseaux de neurones [Kuroda 02], les machines à vecteurs supports (SVM) [Dunker 09,Wu 05] ou les modèles par mélange de gaussiennes[Dunker 09].

**La théorie des fonctions de croyance** Les émotions sont des concepts de haut-niveau sémantique qui sont, par nature, hautement subjectifs et ambigus. Ainsi, afin de s'acquitter efficacement de cette tâche de reconnaissance, il est nécessaire de traiter des informations qui peuvent être incertaines, incomplètes, équivoques et pouvant conduire à des conflits. C'est la raison pour laquelle nous proposons de faire usage de la théorie des fonctions de croyance qui gère naturellement ces difficultés.

**Aperçu de la théorie des fonctions de croyance**

La théorie des fonctions de croyance de Dempster-Shafer [Dempster 68,Smets 90] propose un cadre permettant un raisonnement sur des connaissances qui peuvent être incertaines, incomplètes et conduisant à des conflits. Cette théorie s'appuie sur des fonctions de masse qui sont une généralisation des probabilités et des mesures de possibilité. Soit $\Theta = \{\theta_1, \theta_2, ..., \theta_K\}$ l'ensemble fini des $K$ hypothèses possibles. Cet ensemble est nommé cadre de discernement. Les concepts de base de la théorie sont les suivants :

*Fonction de masse de croyance élémentaire* : la fonction de masse $m$, associée à une source d'information donnée (un type de descripteur dans notre cas), attribue une valeur comprise dans l'intervalle $[0, 1]$ pour toute partie $A$ de $\Theta$ (proposition) et remplit les conditions suivantes : $m(\emptyset) = 0$ et $\sum_{A \subseteq \Theta} m(A) = 1$.

$m(A)$ représente la confiance, ou croyance, que nous pouvons avoir dans la réalisation d'une proposition $A$. Les éléments focaux sont des sous-ensembles $A$ tels que $m(A) > 0$. Si $m(\Theta) = 1$ alors la source est totalement incertaine alors que si $m(\theta_1) = 1$ alors la source est parfaite pour l'hypothèse $\theta_1$.

*Règle de combinaison* : l'une des propriétés les plus intéressantes de la théorie des fonctions de croyance réside dans sa capacité à combiner les fonctions de masse différentes issues de plusieurs sources d'information. Considérons $m_1(.)$ et

$m_2(.)$ deux fonctions de masse provenant de deux sources d'information indépendantes $S_1$ et $S_2$ respectivement. Dès lors, $m_1(.)$ et $m_2(.)$ peuvent être combinées pour obtenir la masse de la croyance engagée sur $C \subseteq \Theta$, $C \neq \emptyset$ selon la formule de combinaison suivante (Shafer, 1976):

$$m(C) = \frac{\sum_{B \cap A = C} m_1(B).m_2(A)}{1 - \sum_{B \cap A = \emptyset} m_1(B).m_2(A)} \quad (1)$$

Une fois que les fonctions de masse des différentes sources d'informations à notre disposition sont combinées en une seule fonction de masse, une décision finale peut être prise en considérant l'hypothèse qui est associée à la valeur la plus élevée.

**Construction de la croyance élémentaire**

Une des principales difficultés rencontrées lors de l'élaboration d'une méthode de classification basée sur la théorie des fonctions de croyance concerne la manière dont les fonctions de masse de croyance élémentaire sont construites à partir des descripteurs d'images. Dans ce travail, nous avons utilisé l'approche proposée dans [Al-Ani 02] qui estime les fonctions de masse à partir de classificateurs en minimisant l'Erreur Quadratique Moyenne entre les résultats de la classification et les sorties attendues.

### 3.5 Expérimentations

Dans nos expérimentations, nous avons utilisé la base de données d'images IAPS qui est une base de référence en psychologie pour l'étude des émotions communiquées par les images [Lang 08]. Elle fournit une caractérisation des images selon trois critères en fonction de l'émotion produite : l'appréciation, l'activité et le contrôle. Cette base comporte 1192 images qui peuvent donc être représentées dans un espace dimensionnel des émotions, selon les axes d'appréciation et d'activité. Par commodité, cette représentation des émotions n'est pas utilisée directement, mais est utilisée pour définir 4 classes d'émotions correspondant aux 4 quadrants de la figure 4. Le corpus IAPS est partitionné aléatoirement en un ensemble d'apprentissage (80% des données, 953 images) et un ensemble de test (20% des données, 239 images). Toutes les expériences sont répétées 10 fois pour obtenir un pourcentage moyen de classification correcte. Pour évaluer la performance des différents classificateurs pour la reconnaissance des émotions dans les images, nous avons examiné quatre classificateurs représentatifs : machines à vecteurs supports (SVM), réseaux de neurones (Feed-Forward Neural Networks), Adaboost et K-plus proches voisins. Le schéma de classification que nous avons retenu consiste à utiliser deux classificateurs binaires. Le premier est entrainé pour identifier l'activité, et le second sert à identifier l'appréciation. Les résultats sont ensuite combinés pour identifier l'une des 4 classes d'émotion.

Les caractéristiques d'entrée sont générées en utilisant les techniques décrites dans la partie 3.3 et alignées en un seul vecteur, ce qui correspond à une fusion précoce. Les résultats de classification sont donnés dans la figure 5. Nous pouvons observer que les classificateurs SVM et Adaboost sont les plus efficaces avec

des performances très proches, leur pourcentage moyen de classification correcte étant respectivement de 62,6% et 63,3%.

| | *NN (%)* | *SVM (%)* | *Adaboost( %)* | *Knn(%)* |
|---|---|---|---|---|
| I | 57.21 | 61.55 | **65.02** | 51.33 |
| II | 63.42 | 60.34 | 62.53 | **64.42** |
| III | 58.21 | **62.61** | 61.31 | 51.52 |
| IV | 61.72 | **65.75** | 64.30 | 61.71 |

**Fig. 5.** Pourcentages moyens de classification correcte pour 4 classes d'émotion obtenus par les 4 classificateurs.

Un autre aspect intéressant consiste à comparer la capacité des différents types de descripteurs d'images à porter l'information relative aux émotions. Ainsi, le système de classification basé sur SVM décrit précédemment a été appliqué indépendamment pour chaque type de descripteurs. Les résultats sont donnés dans la figure 6. Cette figure présente également le pourcentage de bonne classification obtenu avec la fusion de tous les descripteurs en s'appuyant sur l'approche fondée sur la théorie des fonctions de croyance à la section 3.4. La première remarque est que la performance entre les différents descripteurs est très similaire, variant de 53,3% pour les corrélogrammes jusqu'à 58,2% pour LBP (Local Binary Patterns). Toutefois, parmi les différents descripteurs, il semble que la texture (LBP et la matrice de cooccurrences) soit le type de descripteurs le plus efficace. En outre, les descripteurs de plus haut-niveau (dynamisme et harmonie) même s'ils peuvent paraître moins performants au premier abord ne reposent que sur une seule valeur et donc, leur efficacité est tout à fait remarquable. Enfin, il faut mentionner que l'approche proposée pour la fusion de l'ensemble des descripteurs basée sur la théorie des fonctions de croyance, et dont la matrice de confusion est donnée dans la figure 7, donne les meilleurs résultats avec un pourcentage moyen de classification correcte de 64,6%. Cette valeur montre la capacité de la théorie des fonctions de croyance à combiner différentes sources d'information et à exploiter leurs complémentarités.

Ainsi, dans le cadre du projet OMNIA, ces techniques offrent la possibilité non seulement d'attribuer une ou des étiquettes aux images correspondant aux émotions qui leur sont associées mais également de fournir pour chaque image un score selon les dimensions d'appréciation et d'activité afin de faciliter et d'améliorer la recherche dans les collections d'images en intégrant l'émotion comme critère de recherche.

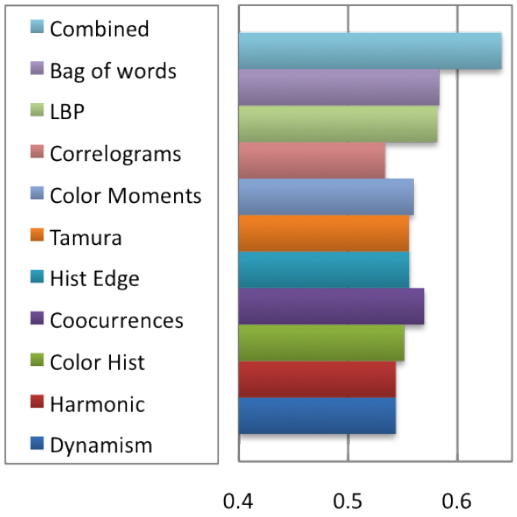


**Fig. 6.** Taux de reconnaissance moyens obtenus pour chaque type de descripteurs et par fusion (combined).

| Réel \ Prédit | I | II | III | IV |
|---|---|---|---|---|
| I | **63.32** | 12.25 | 11.23 | 11.15 |
| II | 11.05 | **61.42** | 12.27 | 11.82 |
| III | 16.21 | 12.53 | **66.19** | 10.52 |
| IV | 10.42 | 13.80 | 10.31 | **67.51** |
| Total | 100 | 100 | 100 | 100 |

**Fig. 7.** Matrice de confusion pour les 4 classes d'émotion en utilisant la théorie des fonctions de croyance.

## 4 GETALP : Extraction de contenu dans des textes multilingues

### 4.1 Introduction

Le contenu visuel des images, auquel on souhaite accéder via le système OMNIA, est indépendant de la langue utilisée dans les textes compagnons. Par exemple, une image de chien accompagnée d'une légende "dog", "perro" ou "cane" reste une image de chien [Popescu 07]. D'autre part, lors du processus de recherche, un utilisateur doit pouvoir formuler librement des requêtes dans une langue naturelle préférée (si possible sa langue maternelle). Ainsi, un traitement multilingue des textes est nécessaire, tant pour la recherche que pour l'indexation.

Lors du processus d'indexation, l'analyse des textes vise à extraire le contenu pertinent pour la description des images associées. Lors du processus de recherche, les requêtes en langue naturelle doivent être transformées en requêtes formelles ne contenant que les informations pertinentes. Ces deux tâches relèvent de l'extraction de contenu (un cas particulier d'extraction d'information) et nécessitent une approche différente de celles employées en traduction automatique. Il est montré dans [Daoud 06] que l'annotation de mots ou locutions avec des items sémantiques (ou "présémantiques") est une approche valide pour commencer une extraction de contenu. En particulier, ce processus ne requiert pas d'analyse syntaxique, une tâche coûteuse et dont la qualité est limitée.

Sans entrer dans le détail des modules qui composent traditionnellement un système d'extraction d'informations [Grishman 97], nous pouvons dire qu'ils sont développés en fixant un certain nombre de paramètres (e.g. le domaine ou le type des entrées). La langue des entrées est bien sûr l'un de ces paramètres et il est montré dans [Hajlaoui 07] que le portage linguistique de tels systèmes est grandement facilité si l'on dispose d'une représentation interne formalisée du contenu des textes.

Ainsi, en vue de permettre une extraction de contenu dans des textes multilingues, nous proposons de les représenter et d'effecter les traitements initiaux avec le formalisme des Systèmes-Q [Colmerauer 70] et d'annoter les mots ou locutions de ces textes avec les lexèmes interlingues (Universal Words, UW) du langage pivot UNL (Universal Network Language). Ce processus peut être vu comme une lemmatisation interlingue qui n'est aujourd'hui proposée par aucun logiciel de lemmatisation. On peut également le qualifier d'annotation présémantique des textes.

Notre méthode est testée sur une base de données, contenant 500K images accompagnées de textes d'une cinquantaine de mots (environ 2,5M mots au total), fournie par l'agence de presse Belge *Belga News* pour la campagne *CLEF09*. Ce corpus n'est disponible qu'en anglais mais permet de tester le passage à l'échelle de notre méthode. Il a été traduit automatiquement dans 5 langues pour de futures expériences sur le multilinguisme.

### 4.2 Ressources et structures de données

**UNL (Universal Networking Language)** [Uchida Hiroshi et al. 09] fait référence à trois choses différentes :

1. un projet international impulsé en 1996 par l'UNU (Université des Nations Unies) à Tokyo ;
2. un langage pivot dans lequel les phrases sont représentées sous forme de graphes sémantiques fondés sur l'anglais ;
3. un format de document multilingue (aligné au niveau des phrases), intégré à HTML.

Le langage UNL [Boitet 09] représente le sens d'une phrase par une structure sémantique abstraite (un hyper-graphe). Chaque arc de l'hyper-graphe est

étiqueté avec une relation sémantique parmi 41 disponibles (agt, obj, aoj, pos, pls, mea, cag, etc.). Chaque nœud contient, soit un lexème interlingue appelé UW (Universal Word) et des attributs sémantiques (cardinal, aspect, intonation, flexion, etc.), soit un sous-graphe (ce qui explique la notion d'hyper-graphe).

Un UW est composé de :

1. un *mot vedette*, si possible tiré de l'anglais, qui peut être un mot, des initiales, une expression ou même une phrase complète. C'est une étiquette pour le concept qu'il représente ;
2. une liste de *restrictions* dont le but est de spécifier précisément le concept auquel l'UW fait référence.

Exemples :

- book(icl>do, agt>human, obj>thing) et book(icl>thing)
- ikebana(icl>flower_arrangement)
- go_down

Un ensemble d'UW constitue un lexique pour UNL. Idéalement, un UW fait référence de façon non ambigüe à un concept partagé par plusieurs cultures. Cependant, les UW sont faits pour représenter les acceptions d'une langue ; nous nous trouvons ainsi des UW différentes (e.g. affection et disease) qui font référence au même concept (maladie). Les UW peuvent ainsi être qualifiées de présémantiques ou préconceptuelles.

Nous utilisons actuellement 207K UW construites automatiquement à partir des synsets du Princeton Wordnet dans le cadre du consortium U++ [Jesus Cardeñosa et al. 09]. Ces UW sont reliés aux langues naturelles par des dictionnaires bilingues dont le stockage et la manipulation en ligne sont assurés par la plateforme PIVAX [Nguyen 07].

**PIVAX** est une plate-forme de gestion de dictionnaires en ligne basée sur JIBIKI, une plate-forme générique pour la gestion et l'édition collaborative de ressources lexicales [Sérasset 05,Serrasset 09].

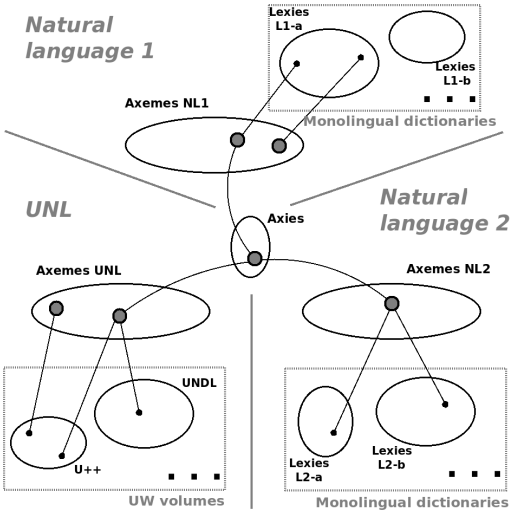


**Fig. 8.** Stockage de dictionnaires

Pour chaque langue supportée, une instance de PIVAX dispose d'un espace dédié contenant :
- un ou plusieurs volumes de *lexies* correspondant aux sens des mots dans un dictionnaire ;
- un unique volume d'*axèmes* (acceptions monolingues) qui sont des liens entre les lexies synonymes d'une langue ;
- un volume partagé d'*axies* (acceptions interlingues) qui sont des liens entre les axèmes synonymes de différentes langues.

La figure ci-contre illustre le stockage de dictionnaires multilingues dans PIVAX.

**Les Systèmes-Q :** Il est possible d'insérer des annotations directement au fil du texte avec des balises (e.g. XML) comme dans la table 1. Cependant, cette approche n'est pas adéquate pour représenter des ambiguïtés de segmentation (dans l'exemple suivant, il est possible de lister les différentes interprétations pour "in", mais pas de représenter "waiting", "room" et "waiting room" comme trois unités lexicales possibles).

| in a waiting room |
|---|
| <tag uw='in(icl-sup-how), in(icl-sup-adj), in(icl-sup-linear_unit, equ-sup-inch)'>**in**</tag> <tag uw='unk'>**a**</tag> <tag uw='waiting_room(icl-sup-room, equ-sup-lounge)'>**waiting room**</tag> |

**Table 1.** Annotation naïve du fragment de texte "in a waiting room"

Pour permettre la représentation des ambiguïtés (et notamment des ambiguïtés de segmentation), nous proposons d'utiliser les Systèmes-Q. C'est une représentation des textes dans une structure de graphe de chaînes (les Q-graphes), dont les arcs sont décorés par des expressions parenthésées (des arbres). De plus, les Systèmes-Q permettent des traitements à l'aide de règles de réécriture (un ensemble de telles règles est appelé système-Q).

Un exemple de ce formalisme est donné dans la figure 11 de la section 4.2.3. La figure présente successivement le code décrivant un graphe-Q, une règle de réécriture et un schéma du graphe-Q obtenu après application d'un système-Q contenant la règle de l'exemple.

Les Systèmes-Q ont été développés par Alain Colmerauer à l'Université de Montréal [Colmerauer 70]. Nous utilisons actuellement une réimplémentation

élaborée par Hong-Thai Nguyen lors de sa thèse dans l'équipe LIG-GETALP en 2007 [Nguyen 09].

### 4.3 Étapes du processus d'annotation

**Vue d'ensemble :** Le processus d'annotation est composé des étapes suivantes :

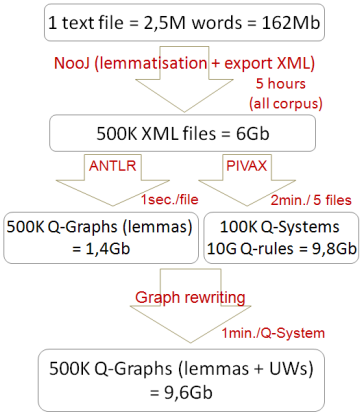


**Fig. 9.** Taille des données et durées des traitements (2,1GHz CPU, 2Gb RAM)

1. Lemmatisation avec un logiciel adéquat
2. Transcription des textes lemmatisés en graphes-Q ;
3. Création par PIVAX de dictionnaires locaux bilingues, sous forme de systèmes-Q dont chaque règle est de la forme lemme → lemme + UW ;
4. Exécution de ces systèmes-Q (dictionnaires) sur les graphes-Q (textes) ;
5. Extraction de contenu sur les graphes-Q obtenus.

Dans la suite, nous décrivons en détail les quatre premières étapes. Nous présentons également des résultats expérimentaux obtenus sur la base de test Belga (500K légendes d'une cinquantaine de mots, soit environ 2,5M mots au total). La taille des données et la durée des traitements sont récapitulées dans la figure ci-contre.

**Lemmatisation :** Nous utilisons des dictionnaires dont les entrées sont des lemmes ; la première étape est donc de lemmatiser le texte d'entrée. Cette étape entraine deux types d'ambiguïtés : d'une part des ambiguïtés de segmentation pour déterminer les unités lexicales, d'autre part la multitude des interprétations possibles pour une unité lexicale.

Pour l'extraction ou la recherche d'information, il est plus judicieux de conserver les ambiguïtés que de mal les résoudre. Nous avons donc besoin de lemmatiseurs qui conservent les ambiguïtés (éventuellement un pour chaque langue). Pour chacun, nous proposons d'utiliser des grammaires ANTLR [Terence Parr et al. 09] pour transformer les résultats en graphes-Q.

Dans notre première expérience sur le corpus Belga, nous avons utilisé le système NooJ. Il représente toutes les ambiguïtés dans un *réseau de confusion* comme illustré dans la figure 10.

**Dictionnaires locaux sous forme de Systèmes-Q :** Une fois les textes annotés avec des lemmes, sous forme de graphes-Q, nous utilisons les possibilités

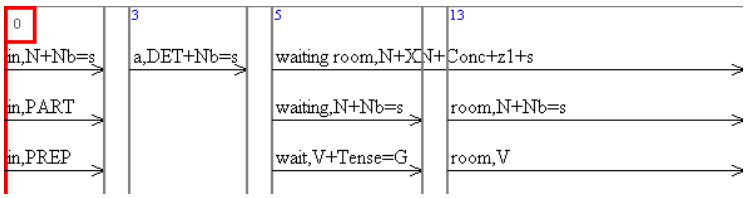


**Fig. 10.** Sortie de NooJ pour l'exemple"waiting room"

de réécriture des Systèmes-Q pour les annoter avec des UW. Grâce à PIVAX, nous exportons les entrées des dictionnaires bilingues sous forme de règles-Q (lemme → lemme + UW). Afin que les systèmes-Q produits soient exécutables avec des ressources raisonnables, nous créons un système-Q par texte (appelé dictionnaire local), qui ne contient que les entrées de ce texte (une cinquantaine d'entrées au maximum pour les textes de Belga). La création de ces dictionnaires locaux est la tâche la plus coûteuse en temps dans notre processus d'annotation et devra être optimisée (voir 4.3). Cependant, la première expérience a montré que notre méthode passe déjà bien à l'échelle.

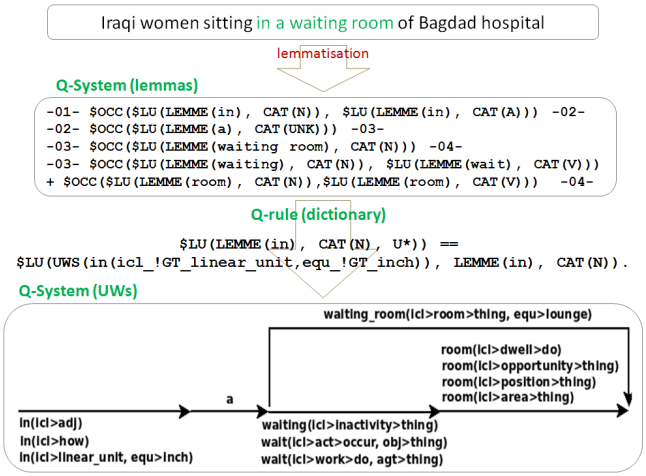


**Fig. 11.** Creation et execution d'un Système-Q

### 4.4 Vers une extraction de contenu interlingue

Notre processus d'annotation donne lieu à de nombreuses ambiguïtés qui viennent tant de la lemmatisation que de l'interprétation des lemmes comme UW. Par exemple, la figure 11 montre que l'unité lexicale "waiting room" peut être interprétée de 13 façons différente avec des UW. Nous travaillons donc sur un

processus de désambiguïsation. Il est basé sur l'utilisation de *vecteurs conceptuels* [Schwab 05] et associe des scores de vraisemblance à chaque UW pouvant correspondre à une unité lexicale du texte.

L'extracteur de contenu que nous développons dans le cadre du projet OMNIA est guidé par une base de connaissances pour déterminer quelles sont les informations pertinentes à extraire des textes, en vue de classifier les images associées. Cette base de connaissances prend la forme d'une ontologie avec une faible expressivité logique. Comme l'extraction de contenu est faite sur des graphes-Q étiquetés par des UW (une représentation interlingue des textes), il est nécessaire que notre base de connaissances soit reliée aux lexiques d'UW. La construction et le maintien d'une correspondance entre une ontologie et un lexique interlingue est un défi, intéressant également pour la multilinguïsation d'ontologies [Rouquet 09].

## 5 Conclusion

Nous avons vu dans cet article plusieurs traitements, visuels ou textuels, permettant de recueillir des informations variées et complémentaires pour décrire différentes facettes d'une image (thématique, esthétique, affective, etc.). Chaque traitement donne lieu à un descripteur spécifique de l'image qui n'a pas été décrit en détail ici. Grossièrement, ces descripteurs ont la forme de vecteurs donc chaque composante est associée à un aspect ou une classe pour l'image (mer, montagne, active, etc.). Les valeurs scalaires stockées dans les vecteurs peuvent avoir différentes interprétations : score de vraisemblance, intensité, etc. La fusion des descripteurs spécifiques au sein d'un descripteur unique constitue la prochaine tache difficile du projet OMNIA. Pour relever ce défi, nous devrons notamment spécifier précisément la sémantique des scalaires contenus dans un descripteur et déterminer des stratégies de résolution quand les informations issues de différents descripteurs s'avèrent contradictoires.

Un autre travail important en cours est l'intégration des différents processus dans une chaine de traitement (*workflow*) pour l'indexation des images. Nous souhaitons enfin exploiter les résultats de l'indexation dans une interface graphique permettant à l'utilisateur d'exprimer des requêtes spontanées dans sa langue maternelle.

## 6 Remerciements

Les auteurs remercient l'ANR et le projet OMNIA qui leur permet de développer ces recherches.

## References

[Al-Ani 02] A. Al-Ani & M. Deriche. *A new technique for combing multiple classifier using the Dempster Shafer theory of evidence.* J. Artif. Intell. Res, vol. 17, pages 333–361, 2002.

[Boitet 09] Christian Boitet, Igor Boguslavskij & Jesus Cardeñosa. *An Evaluation of UNL Usability for High Quality Multilingualization and Projections for a Future UNL++ Language*. In Computational Linguistics and Intelligent Text Processing, pages 361–373. 2009.

[Bosch 06] A. Bosch, A. Zisserman & X. Munoz. *Scene classification via pLSA*. In ECCV, 2006.

[Cho 02] S.-B. Cho & J.-Y. Lee. *A human-oriented retrieval system using interactive genetic algorithm*. IEEE Transactions on systems, man and cybernetics, vol. 32, no. 3, pages 452–458, 2002.

[Colmerauer 70] A. Colmerauer. *Les systèmes-q ou un formalisme pour analyser et synthétiser des phrases sur ordinateur*. département d'informatique de l'Université de Montréal, publication interne, vol. 43, September 1970.

[Columbo 99] C. Columbo, A. Del Bimbo & P. Pala. *Semantics in visual information retrieval*. IEEE Multimedia, vol. 6, no. 3, pages 38–53, 1999.

[Csurka 04] G. Csurka, C. Dance, L. Fan, J. Willamowski & C. Bray. *Visual Categorization with Bags of Keypoints*. In ECCV Workshop on Statistical Learning for Computer Vision, 2004.

[Daoud 06] Daoud Daoud. *Il faut et on peut construire des systèmes de commerce électronique à interface en langue naturelle restreints (et multilingues) en utilisant des méthodes orientées vers les sous-langages et le contenu*. PhD thesis, UJF, September 2006.

[Dempster 68] A. P. Dempster. *A generalization of Bayesian inference*. Journal of the Royal Statistical Society, Series B, vol. 30, pages 205–247, 1968.

[Dunker 09] P. Dunker, S. Nowak, A. Begau & C. Lanz. *Content-based mood classification for photos and music*. ACM MIR, pages 97–104, 2009.

[Everingham 06] M. Everingham, A. Zisserman, C. Williams & L. Van Gool. *The PASCAL Visual Object Classes Challenge 2006*, 2006.

[Grishman 97] Ralph Grishman. *Information Extraction: Techniques and Challenges*. In International Summer School on Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology, pages 10–27. Springer-Verlag, 1997.

[Hajlaoui 07] Najeh Hajlaoui & Christian Boitet. *Portage linguistique d'applications de gestion de contenu*. In TOTh07, Annecy, 2007.

[ImagCLEF ] ImagCLEF. *http://ir.shef.ac.uk/imageclef/*.

[Jesus Cardeñosa et al. 09] Jesus Cardeñosa et al. *The U++ Consortium (accès en septembre 2009)*. http://www.unl.fi.upm.es/consorcio/index.php, September 2009.

[Kuroda 02] K. Kuroda & M. Hagiwara. *An image retrieval system by impression words and specific object names IRIS*. Neurocomputing, vol. 43, pages 259–276, 2002.

[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. University of Florida, Gainesville, FL, 2008.

[Li 07] C.-T. Li & M.-K. Shan. *Emotion-based impressionism slideshow with automatic music accompaniment.* ACM Multimedia, pages 839–842, 2007.

[Nguyen 07] H.T. Nguyen, C. Boitet & G. Sérasset. *PIVAX, an online contributive lexical data base for heterogeneous MT systems using a lexical pivot.* In SNLP, Bangkok, Thailand, 2007.

[Nguyen 09] Hong-Thai Nguyen. *EMEU_w, a simple interface to test the Q-Systems (accès en septembre 2009).* http://sway.imag.fr/unldeco/SystemsQ.po?localhost=/home/nguyenht/SYS-Q/MONITEUR/, September 2009.

[Perronnin 06] F. Perronnin, C. Dance, G. Csurka & M. Bressan. *Adapted Vocabularies for Generic Visual Categorization.* In ECCV, 2006.

[Picard 97] R.W. Picard. *Affective Computing.* MIT Press, Cambridge, 1997.

[Popescu 07] Adrian Popescu. *Image Retrieval Using a Multilingual Ontology.* Pittsburg PA, USA, June 2007.

[Rouquet 09] David Rouquet & Hong-Thai Nguyen. *Multilinguïsation d'une ontologie par des corespondances avec un lexique pivot.* In TOTh09, volume à paraitre, Annecy, May 2009.

[Schwab 05] Didier Schwab. *Approche hybride - lexicale et thématique - pour la modélisation, la détection et l'exploitation des fonctions lexicales en vue de l'analyse sémantique de texte.* PhD thesis, Université Montpellier 2, July 2005.

[Serrasset 09] Gilles Serrasset & Mathieu Mangeot. *The Jibiki project on LIGforge (accès en septembre 2009).* http://ligforge.imag.fr/projects/jibiki/, September 2009.

[Smets 90] P. Smets. *The combination of evidence in the transferable belief model.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 12, no. 5, pages 447–458, 1990.

[Sérasset 05] Gilles Sérasset. *Multilingual Legal Terminology on the Jibiki Platform: The LexALP Project.* In SNLP05, Thailand, 2005.

[Terence Parr et al. 09] Terence Parr et al. *ANTLR Parser Generator (accès en septembre 2009).* http://www.antlr.org/, September 2009.

[Uchida Hiroshi et al. 09] Uchida Hiroshi et al. *The UNDL Foundation (accès en septembre 2009).* http://www.undl.org/, September 2009.

[Wang 05] S. Wang & X. Wang. *Emotion semantics image retrieval: a brief overview.* ACII, pages 490–497, 2005.

[Wang 08] W. Wang & Q. He. *A survey on emotional semantic image retrieval.* ICIP, pages 117–120, 2008.

[Weining 06] W. Weining, Y. Yinglin & J. Shengming. *Image retrieval by emotional semantics: A study of emotional space and feature extraction.* IEEE ICSMC, vol. 4, pages 3534–3539, 2006.

[Wu 05] Q. Wu, C. Zhou & C. Wang. *Content-based Affective Image classification and retrieval using support vector machines.* ACII, pages 239–257, 2005.

[Zeng 09] Z. Zeng. *A survey of affect recognition methods: audio, visual and spontaneous expressions.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 31, no. 1, pages 39–58, 2009.