MULTIMODAL SEARCH FOR GRAPHIC DESIGNERS*

Sandra Skaff[1], David Rouquet[2], Emmanuel Dellandrea[3], Achille Falaise[2]
Valérie Bellynck[2], Hervé Blanchon[2], Christian Boitet[2], Didier Schwab[2]
Liming Chen[3], Alexandre Saidi[3], Gabriela Csurka[1] and Luca Marchesotti[1]
[1]*Xerox Research Centre Europe, Grenoble, France*
[2]*Laboratoire d'Informatique de Grenoble, Grenoble, France*
[3]*Ecole Centrale de Lyon, Écully, France*

Keywords: Interface, Visualization.

Abstract: This paper describes OMNIA, a system and interface for searching in multimodal image collections. OMNIA includes a set of tools which allow the user to retrieve assets using different features. The tools are based on extracting different types of asset features, which are content, aesthetic, and emotion. Visual-based features are used to retrieve assets using each of these tools. In addition, text-based features can be used to retrieve image assets based on content. Different datasets are used in OMNIA and retrieved assets are displayed in such a way which facilitates user navigation. It is shown how OMNIA can be used for simple, efficient, and intuitive asset search in the context of graphic design applications.

1 INTRODUCTION

In the OMNIA project, we devise different steps necessary to build a system for digital asset retrieval and management to address the needs in marketing and public relation applications. In enterprise marketing, content creation is needed. Commercial applications such as Corbis, Getty images and Reuters have started to employ semi-automatic tools for asset management. There has also been work on applying existing information visualization techniques to browsing image type assets (Yang et al., 2006). However, the fact remains that digital assets are still difficult to mine, manage, and use. The efforts of the OMNIA project are focused on developing a prototype for digital asset management using cutting-edge technologies for content-based image retrieval (Datta et al., 2008a). In particular, one of our aims is to combine text annotation techniques with digital asset analysis for content creators.

The system proposed consists of different tools for retrieval based on different types of asset analysis. On one hand, digital assets are analyzed over a range of features, namely, content, aesthetic, and emotion. On another hand, the paper also proposes combining text-based annotation techniques with such analysis. In designing a search engine, it is important to provide users with an interface which allows effective search, user-friendly interaction, and efficient visualization. These are characteristics of the OMNIA interface.

A graphic design scenario is used to demonstrate the efficacy of OMNIA in retrieving relevant images. The designer has a brochure template which needs to be illustrated with images. These are selected using OMNIA to query by content first, before browsing through and visualizing the images in one or two dimensions by aesthetic and emotion features. Additional functionalities which are not covered by the given scenario are also shown.

The remainder of the article is structured as follows. Section 2 is an overview of the state of the art in image retrieval systems. Section 3 describes the datasets incorporated into OMNIA. Section 4 details the multi-modal search tools as well as their quantitative evaluations where applicable. Section 5 givens an overview of the OMNIA system, which includes the tools and interface. The utility of OMNIA for fast and user-friendly search as well as intuitive visualization of search results is shown in the context of graphic design in Section 6. Finally, Section 7 summarizes the paper.

2 STATE OF THE ART

Most image retrieval systems are based on textual search such as Google images, or on manual annota-

*http://www.omnia-project.com

Skaff S., Rouquet D., Dellandrea E., Falaise A., Bellynck V., Blanchon H., Boitet C., Schwab D., Chen L., Saidi A., Csurka G. and Marchesotti L..
MULTIMODAL SEARCH FOR GRAPHIC DESIGNERS.
DOI: 10.5220/0003322401640176
In *Proceedings of the International Conference on Imaging Theory and Applications and International Conference on Information Visualization Theory and Applications* (IVAPP-2011), pages 164-176
ISBN: 978-989-8425-46-1

tions search such as Getty images and Flickr in spite of the fact that content-based image search systems were proposed by researchers several decades ago. The most common method for comparing two images in content-based image retrieval, typically a query image and a database image, is using an image similarity measure based on the distance between image signatures.

Early systems mainly used global image descriptors based on color (Swain and Ballard, 1991; Tsujimura and Bannai, 1996; Kasutani, 2007) or on color combined with texture and shape (Flickner et al., 1995). More recent systems extract local features from image patches or segmented image regions and use techniques based on feature matching (Chen and Wang, 2002), build inverted files (Squire et al., 1999), Bag-Of-Visual words (Csurka et al., 2004) or Fisher Vectors (Perronnin et al., 2010).

These new representations paved the way for large advances in both content-based image retrieval (Laaksonen et al., 2002; Sahbi et al., 2007; Jegou et al., 2008; Chum et al., 2009) and in auto-annotation (Jeon et al., 2003; Monay and Gatica-Perez, 2003; Barnard et al., 2004; Li et al., 2006; Zhang et al., 2006; Guillaumin et al., 2009) applications. As images are often accompanied by text and metadata, much research work focuses also on information fusion and multi-modal retrieval systems. Comparisons between different retrieval systems such as visual-based and text-based have been performed for different types of images such as photo and medical in the ImageClef Evaluation Forum (http://www.imageclef.org). In (Müller et al., 2010) there are numerous articles presenting results of several years of competition. The proposed systems range from simple, early and late fusion strategies to more complex cross-media similarity measures (Ah-Pine et al., 2010).

On another hand, new methods were proposed which incorporate aesthetics and emotion to annotate images (Datta et al., 2006; Jacobsen et al., 2006; Datta et al., 2008b; Fedorovskaya et al., 2008; Loui et al., 2008; Davis and Lazebnik, 2008) using learning techniques. Aesthetic and emotion concepts are highly subjective and difficult to learn as shown in these methods. Unlike these approaches, our system does not tag an image with a specific concept but assigns a score to an image for each of the concepts. In this respect, our system provides the flexibility to search by different combinations of these concepts without the need for high accuracy in aesthetic or emotional categorization, as compared with most systems cited above. In addition, our system does not need to combine scores pertaining to different aspects, content, aesthetic, emotion, into a single value for an image such as in (Loui et al., 2008).

Another advantage of our user interface is that it facilitates not only a creative asset navigation, but also visual content creation by allowing a combined visualization of the working draft with different selected images. Such a visualization is shown in the user scenario in Section 6. The system proposed in this paper can incorporate any of the methods mentioned above, including ones based on keyword search. Finally, our system as described in Section 4, can annotate images based on analysis of multilingual text as well.

3 THE DATASETS

The digital assets used in OMNIA are obtained from three databases: MIRFLICKR, BELGANews, and Color Palettes.

Mirflickr. The Mirflickr collection consists of 25,000 photos obtained from Flickr with creative commons license. For more detail, the reader is referred to (http://press.liacs.nl/mirflickr).

BelgaNews. The Belga news collection consists of 500,000 images and English companion texts (about 50 words) obtained from Belga, a Belgian press agency. For more detail, the reader is referred to (http://www.belga.be).

Color Palettes. This database consists of 25,000 color palettes extracted from the Mirflickr images. These palettes are extracted using K-means clustering. The mean of each cluster is represented by one color of the palette. In our work, we assume that $K = 5$, thus extracting palettes with five colors from each image. The color palettes are annotated with the same content the images they are extracted from are annotated with.

4 THE TOOLS

The OMNIA framework consists of a set of tools which allow the user to search and organize digital assets according to their preference in three main dimensions. These tools are based on three types of analysis: content, aesthetic and emotion. While the emotion and aesthetic analysis are solely visual-based, the content analysis performed is either visual-based or text-based.

4.1 Semantic Content Analysis

The first tool our system makes available to the user, is a search by content. In other words, the user can type in "flower", and visualize all the flower images. This search by content is based on tagging the images with one of two techniques. The first one is visual-based, whereby the content of the image is inferred from visual features of the image. The second one is text-based, whereby the content of the image is inferred from text surrounding the image. The two techniques are described below.

4.1.1 Visual-based

One of the most popular approaches to image classification to date has been to describe images with a Bag-Of-Visual-words (BOV) histograms and to classify them using non-linear Support Vector Machines (SVMs) (Csurka et al., 2004). While several variants and extensions were proposed to the original approach, the main schema of the whole system remains often the same. The mains steps of the classification algorithm, referred to as the Generic Visual Categorization (GVC) algorithm, are:

- Detecting patches by interest point detectors, low level image segmentation, or sampling patches on regular grid at single or multiple scales.
- Extracting low level features on these patches.
- Building a visual dictionary using Kmeans, Mean Shift, Gaussian Mixture Models (GMMs) or Random Forest.
- Obtaining a high-level image representation using BOV, Fisher Vectors or image GMMs.
- Employing learning and classification techniques such as Naive Bayes, SVMs or sparse logistic regression.

In our exemplary implementation, we used the training data and the 53 concepts organized into an ontological hierarchy of the ImageClef09 Large Scale Detection and Photo Annotation Task (http://www.imageclef.org/2009/PhotoAnnotation). Therefore, we reproduced the system described in (Ah-Pine et al., 2009) as follows. We sample patches on regular grid at multiple scales and compute local color statistics (COL) and orientation histograms (ORH). We then build a GMM based visual vocabulary on COL and one on ORH features. We use the Fisher Vectors proposed in (Perronnin and Dance, 2007) as high level features and train a non-linear classifier per concept using sparse logistic regression. Finally, the COL and ORH scores obtained for each image from the GVC algorithm described above are averaged and converted to probabilities before post-processing them to verify the ontological constraints.

While the original GVC (Csurka et al., 2004) or any other extension of it as in (Zhang et al., 2007; Tahir et al., 2009) can be employed in the OMNIA system, we chose to use the algorithm of (Ah-Pine et al., 2009) for the following reason. This algorithm had the best scores according to the Hierarchical Measure (HM) which considers the relationships between concepts and the agreement of annotators on concepts (Nowak and Lukashevich, 2010). We consider that this measure is more suitable to evaluate auto-annotation as it scores simultaneously all concepts. Furthermore, the algorithm (van de Sande et al., 2009) which scored the best on Equal Error Rate (EER) and the Area under Curve (AUC) measures at the challenge is more complex than the above system.

Our system used the same 5,000 images to train the classifier, but it annotated all 25,000 images of the Mirflickr collection. The latter includes the 5,000 images but with different names.

4.1.2 Multilingual Text Analysis

In order to process both image companion text and user free text queries, content extraction based on multilingual text is required. A survey of such techniques can be found in (Rouquet et al., 2010). We describe the approach incorporated into the proposed system briefly in what follows as the reader can find more details in (Falaise et al., 2010). The aim is to build formal descriptors or queries to be used in the system. Multilingual content extraction does not imply translation. It has been shown in (Daoud, 2006) that annotating words or chunks with interlingual lexemes is a valid approach to initiate content extraction. We thus skip syntactical analysis, which is an expensive and low quality process, and obtain language independent data early in our flow, allowing further processing to be language independent. We use the lightweight ontology for image classification as the formal knowledge representation that determines relevant information to extract. This ontology is considered as a domain parameter for the content extractor.

The general architecture may be summarized as follows:

- Texts, including companions and queries, are first lemmatised with a language-dependent piece of software. Ambiguities are preserved in a Q-graph structure.
- Then, the lemmatised texts are annotated with interlingual (ideally unambiguous) lexemes, namely Universal Words (UW). This adds a lot of ambi-

guities to the structure, as an actual lemma may refer to several semantically different lexemes.

- The possible meanings for lemmas are then weighted in the Q-graph through a disambiguation process.
- Finally, relevant conceptual information is extracted using an alignment between a domain ontology and the interlingual lexemes.

In the case of OMNIA, conceptual information extracted from companion texts is stored in a database, while conceptual information extracted from user queries are transformed into formal queries for the database (such as SQL and SPARQL).

The implementation follows a Service Oriented Architecture. Each part of the process corresponds to a service. A service supervisor has been built as an interface with the OMNIA prototype to deal with heterogeneity and address normalization issues (e.g. line-breaks, encoding, identification, cookies, page forwarding, etc.). This architecture is able to process multiple tasks concurrently, allowing to deal with user queries in real time while processing companion texts in the background.

The Universal Network Language (UNL) (Boitet et al., 2009; Uchida Hiroshi et al., 2009) is a pivot language that represents the meaning of a sentence with a semantic abstract structure (a hyper-graph) of an equivalent English sentence. We use UNL vocabulary, UniversalWords (UW), as interlingual lexemes to annotate chunks of texts. Thus, further treatment is not language dependent. To add a new language to the system, we only need to link it to the set of UW; otherwise we would have to link it with all other languages in the system. A UW consists of:

1. a *headword*, if possible derived from English, which is a label for the concepts it represents in its original language;
2. a *list of restrictions*, which aims to precisely specify the concept the UW refers to. Restrictions are semantic relations with other UW. The most used is the "icl" relation which points to a more general UW.

Examples of such UW are *book(icl>do, agt>human, obj>thing)* and *book(icl>thing)*. They ideally refer unambiguously to a concept, shared among several languages, and form a pre-ontological structure. We mainly use the 207k UW built by the U++ Consortium (Jesus Cardeosa et al., 2009) from the synsets of the Princeton WordNet, which are linked to natural languages via bilingual dictionaries.

The Q-Systems (Colmerauer, 1970) are the formalism we use to represent ambiguities which can occur in a text interpretation. They represent texts in an adequate graph structure decorated with bracketed expressions (trees) and, moreover, allow processing on this structure via graph rewriting rules (a set of such rewriting rules is a so called Q-System). An example of this formalism is given in Figure 1. It presents successively : the code representing a Q-graph, a rewriting rule and a graphical view of the Q-graph obtained after the application of a set of rules that add UW to the edges.

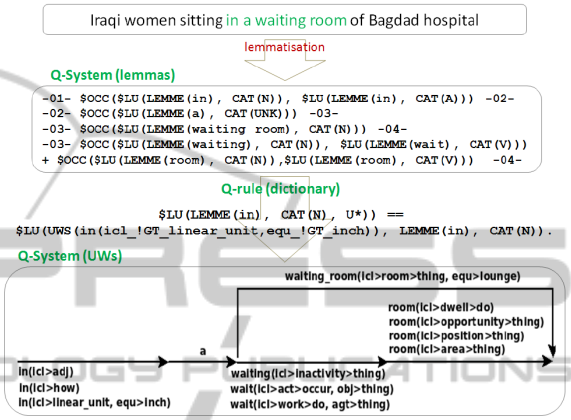


Figure 1: Creation and execution of a Q-System.

The annotation process is composed by the following steps:

1. split of the text in fragments if too long;
2. lemmatisation with a specialized software;
3. transcription in Q-graphs;
4. creation and execution of local bilingual dictionaries (source language - UW) as Q-systems;
5. disambiguation to score the likelihood of each UW.

When texts are represented as Q-graphs decorated with weighted UW, we can start content extraction.

The content extraction process is generic in two aspects :

- It is language independent, as it processes an interlingual representation of the texts.
- The content to be extracted can be specified using a domain ontology as a parameter.

In the OMNIA project, the lightweight ontology for image classification (http://kaiko.getalp.org/kaiko/ontology/OMNIA/OMNIA_current.owl) contains 732 concepts; examples are "animals", "politics", "religion", "army", "sports", "monuments", "transports", "games", and "entertainment". Note that the term concept here is used to refer to content. As the content extractor processes only UW annotations, it is necessary to link ontology elements

to the UW lexicon (Rouquet and Nguyen, 2009). This process is automatic so any OWL ontology (http://www.w3.org/2004/OWL/) can be used to improve performances on specific data collections. Content extraction (relations are not considered yet) is achieved through a three step process:

1. **Concept Matching.** Each UW in the Q-Graph, which matches a concept according to the UW-concept map, is labeled with this concept.
2. **Confidence Calculation.** Each concept label is given a confidence score, in accordance with the disambiguation score of the UW carrying the concept.
3. **Score Propagation.** Since we need autonomous results, we have to perform all ontology-based calculations before releasing them. The confidence scores are propagated in the ontology concept hierarchy according to a fuzzy model. UW restrictions are used if no concept has been found during annotation.

Quantitative Results. Experiments are run on a sub-corpus of 1046 English companion texts from the BelgaNews dataset with the 732 concepts OMNIA ontology. Concepts are retrieved from 77% of texts. The remaining texts are very short: less than ten words with often only dates or names. For example, for the image and companion text shown in Figure 2, we show the concepts extracted and their weights in Table 1.

AWA05 - 20020924 - BAGHDAD, IRAQ : Iraqi women sit under a portrait of Iraqi President Saddam Hussein in a waiting room in Baghdad's al-Mansur hospital 24 September 2002. Saddam Hussein is doggedly pursuing the development of weapons of mass destruction and will do his best to hide them from UN inspectors, the British government claimed in a 55-page dossier made public just hours before a special House of Commons debate on Iraq. Iraqi Culture Minister Hamad Yussef Hammadi called the British allegations "baseless." EPA PHOTO AFPI AWAD AWAD

Figure 2: Image document and companion text example.

We ran a preliminary survey to evaluate the precision of the concept extraction on a sample of 30 images and accompanying texts. Weights were not taken into account in the survey. An extracted concept is considered to match the image and accompanying text under one of two criteria:

1. *Visual relevance*, which considers a concept as correct if it is represented by an element of the image; for instance, it is regarded that there is a match for the concept "sport" if an image contains a minister of sports even if he/she is not actually performing any sport.
2. *Textual relevance*, which considers a concept as correct if it is included in the text, as parts of the text may involve concepts that are not portrayed by the image, such as contextual information, previous events, and so on.

Table 1: The concepts and their corresponding weights extracted from Figure 2.

CONCEPT	WEIGHT
BUILDING	0.098
HOSPITAL	0.005
HOUSE	0.043
MINISTER	0.016
OTHER_BUILDING	0.005
PEOPLE	0.142
PERSON	0.038
POLITICS	0.032
PRESIDENT	0.016
RESIDENTIAL_BUILDING	0.043
WOMAN	0.005

While no concept was found for seven of the image and text samples, 124 concepts were extracted from 23 samples. Of the extracted concepts, 99 matched the sample according to the visual relevance criterion, while 110 matched the sample according to the textual relevance criterion. 14 concepts were incorrect. Therefore the overall precision score is 0.798 according to the visual relevance and 0.887 according to the textual relevance.

4.2 Aesthetic Analysis

A second search tool which can be used in OMNIA is based on aesthetic image analysis. Given one type of images as retrieved in a content search, a user can then organize (cluster using K-means, for example) these images by low level aesthetic feature of his/her choice. The system includes the features listed below:

- Brightness refers to the luminance of an image. It is the average of the brightness values of all the pixels in an image.
- Contrast refers to the efficient use of the dynamic range.
- Saturation refers to the vividness of colored objects in an image.
- The blur of an image is a form of bandwidth reduction typically caused by relative motion between the camera and the original scene or by an optical system which is out of focus.
- Hue refers to the first characteristic of a color that the eye detects.

- Image dimension refers to the number of pixels in an image.
- Color, red, green, or blue, refers to the overall color of an image.

As we will see in Section 6, these features have the role of effectively organizing and representing visually the content of a search space, hence making the search easier and more user-friendly. While, grouping by colors or hue is not new, our system allows for a larger choice of such features and also for combining them. For example, the clustering or grouping of images can be performed in a multidimensional space.

4.3 Emotion Analysis

The third tool which can be used for search in OMNIA is based on the emotion properties of an image. This information is particularly interesting when retrieving images and selecting them according to the effect that they can produce on the viewer, which is typically the case for a graphic designer searching for images as described in Section 6.

Identifying the emotion communicated by an image is still an open problem and even though it is gaining more interest in the research community, contributions remain relatively rare (Wang and Wang, 2005; Wang and He, 2008). To summarize, the three main issues which need to be addressed are the following: emotion representation, image features used to represent emotions and classification schemes designed to handle the distinctive characteristics of emotions. As in any computer vision problems, the main difficulty remains in bridging the gap between low-level features extracted from images and high level semantic concepts, which are emotions in this case.

Several models for the representation of emotions have been considered in the literature (Dunker et al., 2009), and the two main approaches are the discrete one and the dimensional one. The first model consists in considering adjectives or nouns to specify the emotions, such as “happiness”, “sadness”, “fear”, “anger”, “disgust” and “surprise”. The second model describes emotions according to one or more dimensions representing a special emotion characteristic, such as valence, arousal or control. Here we propose to use the second model since it is more suitable to image search for graphic design. Indeed, once images have been retrieved for a given query, a refinement can be performed according to their valence and/or their arousal properties.

Image feature extraction is a key issue for concept recognition in images, and particularly emotions. Features should therefore be designed to carry sufficient information for such recognition to be possible. In this work we have chosen to make use of standard features characterizing the color, texture and shape properties of images. These features include moments of colors, histograms of colors, color correlograms, coocurrence matrices, tamura features (Tamura et al., 1978) and histograms of line orientations. Moreover, two higher level semantic features linked to the affective properties of images are considered: a measure of the image color harmony related to the valence, and a measure of the dynamism related to the arousal (Dellandréa et al., 2010).

The color harmony feature is computed based on Itten’s color theory (Itten, 1961) which states that visual harmony can be obtained by combining hues and saturations so that an effect of stability on the human eye can be produced. This harmony can be represented through the Itten sphere where, in case of harmony, color positions are connected through regular polygons. Thus, to compute the the harmony feature, dominant colors are identified in images and plotted on the color sphere. Then, the polygons linking these colors are characterized by values as follows. A value close to one corresponds to a regular polygon whose center is close to the sphere center which characterizes a harmonious image, and a value close to zero corresponds to an irregular polygon characterizing a non harmonious image. The dynamism feature is computed from the line properties in images. Oblique lines communicate dynamism and action whereas horizontal or vertical lines communicate calmness and relaxation (Columbo et al., 1999). This can be combined with colors in order to produce complex effects suggesting particular feelings or emotions to the viewer. Thus, the dynamism feature is taken to be the ratio of the number of oblique lines in an image to the total number of lines.

An approach for classification of images according to their affective properties based on the theory of evidence (Smets, 1990) has been proposed in (Dellandréa et al., 2010) as an attempt to deal with the ambiguous and subjective nature of emotions. However, this approach requires a discrete representation of emotions, such as “anger”, “sadness”, and “happiness”. As the idea in this paper is to aid in a creative user, namely a graphic designer, in image search, we do not necessarily need to tag images with specific emotions, but to display them in a way which facilitates the search. Therefore a dimensional approach would be more appropriate. We perform the prediction of the emotion properties of images in terms of valence and arousal given two traditional SVMs: *SVM_valence* and *SVM_arousal*. Thus, each of these SVMs is fed with the features described earlier and outputs one value directly representing the valence

through *SVM_valence* and a second value representing the arousal through *SVM_arousal*. For example, a negative value given by *SVM_valence* indicates that the image communicates a negative feeling, whereas a positive value indicates that the image communicates a positive feeling. Moreover, a negative value given by *SVM_arousal* indicates that the image communicates a passive feeling, whereas a positive value indicates that the image communicates an active feeling. In both cases, values close to 0 correspond to neutral emotions.

Quantitative Results. This approach has been evaluated on a subset of the Mirflickr dataset containing 2000 images. In order to obtain the ground truth related to the emotions communicated by these images, we organized an annotation campaign within the OMNIA project. A total of 20 people have been asked to annotate a subset of this dataset by assigning a score for valence and a score for arousal ranging from -10 (very negative/passive) to +10 (very positive/active) to each image. On average, each image was annotated seven times. 80% of the annotated data have been used for training the SVMs (*SVM_valence* and *SVM_arousal*) and the remaining 20% for testing them. Several configurations of these SVMs have been experimented with and the best results were attained using a Gaussian kernel as follows. For *SVM_valence* which is predicting the level of valence, a classification accuracy of 73.3% has been obtained with recall and precision rates of 79.6% and 70.0% respectively for the positive class, and 67.4% and 77.5% respectively for the negative class. For *SVM_arousal* which is predicting the level of arousal, a classification accuracy of 71.9% has been obtained with recall and precision rates of 81.0% and 69.1% respectively for the active class, and 62.5% and 76.1% respectively for the passive class. We consider these results to be good given the fact that emotion recognition is a difficult task even for humans. The results can also be improved and one direction we envisage is enriching the feature set used to represent images by computing higher-level semantic features based on the Gestalt theory of visual perception (Desolneux et al., 2004).

5 THE OMNIA SYSTEM

The searching tools as described in the previous section are the core of the OMNIA system. These tools are accessible through the user interface which was designed to facilitate user interaction with the system. This is done through a Web Interface that also provides different ways of visualizing the results as we will show through examples in Section 6.

Through this interface, the user can provide three different query modes: free-text, image, or a color palette. For the free-text mode, the user can type in a content name and retrieve images or color palettes tagged with that content. In addition, the user can select an image and retrieve images or color palettes of the same color combination as the query image. The user can retrieve the same assets in the case of selecting a color palette.

The system offers different types of visualization of results depending on user needs and queries. Following a content search by free-text, a user can arrange image results in one or two dimensions depending on the number of features he/she selects using the aesthetic and/or emotion tools. Further details through examples are given in Section 6.

6 APPLICATIONS IN GRAPHIC DESIGN

In this section we demonstrate the use of the OMNIA system through showing its different functionalities which can be used in the application of graphic design. We start by going through a scenario which is centered around a graphic designer editing a typical design product such as a brochure. We then show additional functionalities of OMNIA which were not covered in the scenario illustrated.

6.1 Graphic Design Scenario

We assume that the graphic designer needs to finalize a given brochure by illustrating it with appropriate images. The empty brochure as well as the illustrated one are shown in Figure 3.

In this case, images need to have a pertinent content as well as a “look and feel” which goes well with the rest of the brochure. Typically, a designer would mine existing image repositories to select appropriate images. Examples of search engines which may be used are Getty images, Google images, Corbis and/or a proprietary dataset. We show how the OMNIA system can be used to perform the image search to select images iteratively and illustrate the brochure shown in Figure 3.

Firstly, the search space is presented unordered to the user. Then the user creates a content search space through a textual query such as “summer”. The images are presented in a compact visualization allowing up to 350 thumbnails on the same page as shown in Figure 4.

(a)

(b)

Figure 3: (a) The brochure which needs to be illustrated. (b) The brochure after illustration with images obtained using OMNIA.

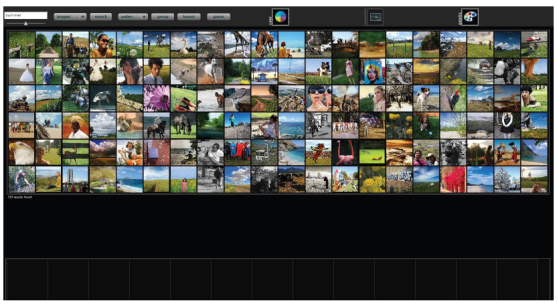


Figure 4: Images retrieved with the query "summer".

To refine the query, the user can select one or several of the aesthetic and emotion features to visualize the grouped images in the space. For example, if the designer selects "contrast" and the red color, the system clusters the images in two dimensions into a predefined number of clusters. Our current implementation of the interface includes three clusters in each dimension and for each corresponding axis. The ranges of values of the axes are taken to be those of the features computed for the images considered. Figure 5 shows an example where the clusters are ordered by the red color on the vertical axis and by the level of contrast on the horizontal axis to enforce a visual coherence. Alternatively, for a quicker grouping and better visualization, especially in the case of more than two features, the system can incorporate the functionality of splitting the space into NxM dimensions and placing the images accordingly.

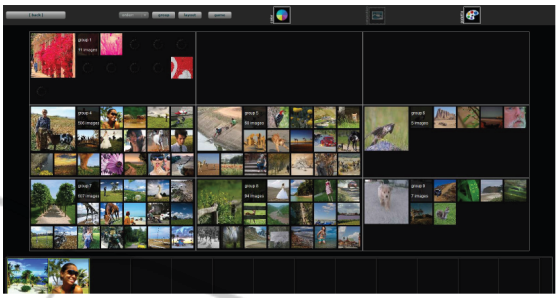


Figure 5: The "summer" images grouped and ordered by "contrast" on the horizontal axis and their red color on the vertical axis.

Then the user can click on one of these groups and hence restrain the search space by the images of that group. Figure 6 shows the images visualized in the case the ones with the highest levels of red color and contrast which are in the top left cluster of Figure 5. Note that the displaying of images in Figure 6 does not correspond to a visual coherence.

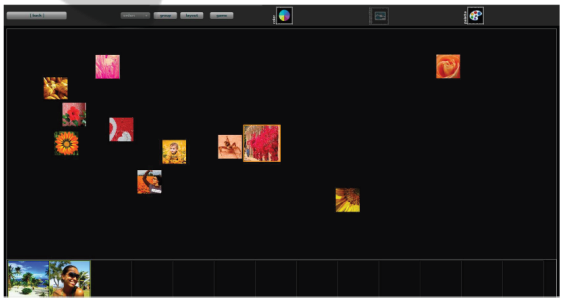


Figure 6: The "summer" images of the top left cluster of Figure 5.

When the user is interested in a specific image, he can visualize it in a higher resolution with a set of extra information as illustrated in Figure 7. This information can be the metadata which includes the image ID, aesthetic feature values such as average brightness or blur, and the color palette, which represents the main colors in the image. The system additionally shows the nearest neighbor images obtained based on the closeness in distance of the color palettes extracted from the images (as described in Section 3). More detail on computing distances between color palettes is given in Section 6.2. An Expectation-Maximization (EM) algorithm which estimates the parameters of a GMM over the pixel rep-

resentations has also been investigated (Csurka et al., 2010). Note that other features such as Fisher Vector representations of images can also be used to obtain the nearest neighbor images. The selected image is

Figure 7: The selected image visualized in high resolution with pertinent information and its nearest neighbors.

automatically saved in the light-box which is shown at the bottom of the interface and which contains previously selected images saved at earlier stages. The images in the light-box are potential candidates for completing the brochure, but the light-box can also contain images which the user selected for later usage.

For graphic designers, the "color feel" of the image is often important. He/she might be also interested in finding a set of colors which fits well with the image. The OMNIA system hence allows the user to visualize the nearest neighbors of the image color palette as shown in Figure 8.

Figure 8: The selected image visualized with the its nearest neighbor color palettes.

Then after selecting a palette, he/she can further visualize the images which have their color palettes similar to the one selected (Figure 9). Note that similarly the selected palette is placed in the light-box and hence can be further used for either querying or for designing a new brochure.

An important feature of the OMNIA interface is that it allows the user to visualize at any time the selected images in the light-box concurrently with the draft of the brochure. He/she can then easily place

Figure 9: The selected color palettes visualized with the images which have similar color palettes.

images on the previewed document and resize them to verifying compatibility and obtain a feeling of the whole brochure as illustrated in Figure 10.

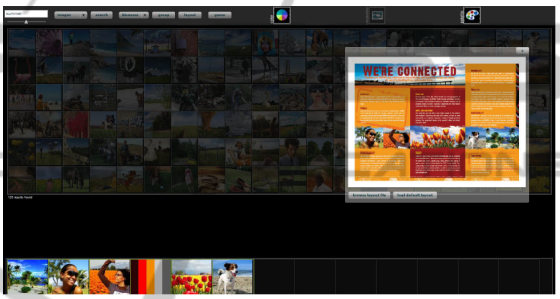


Figure 10: The working brochure previewed with the images selected in the light-box.

6.2 Additional Functionalities

There are additional functionalities of the OMNIA interface which are not covered in the scenario illustrated in Section 6.1.

Search by Text-based Signatures. The content-based search performed in Section 6.1 used images annotated by their visual features. However, in most databases, images are accompanied with text. Designers can query in this case using one or several keywords. The system then searches a structured body of text such as a sentence, a caption, or an article. The OMNIA system builds text signatures for all the images using the Universal Words (UW) as described in Section 4.1.2 and hence allowing to query those images in a language other than that of the original text. In contrast to typical tag- or text-based searches, texts are indexed using a predefined image ontology and for each concept in the ontology a concept score is computed (see details in Section 4.1.2). We show in Figure 11 images retrieved by the concept scores for "sports".

Similar to Section 6.1, we can again group these images by aesthetic features to further restrain the

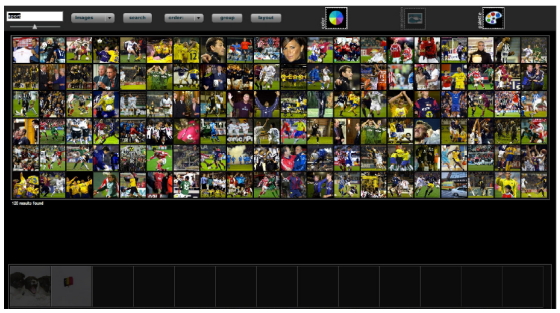


Figure 11: Images retrieved with the query "sport" using concept scores computed on the text accompanying the images.

search space. In addition, when an image is selected, the system also displays its nearest neighbors (Figure 12).

Figure 12: The selected image visualized in high resolution with its metadata and its nearest neighbors. The accompanying text is also visualized.

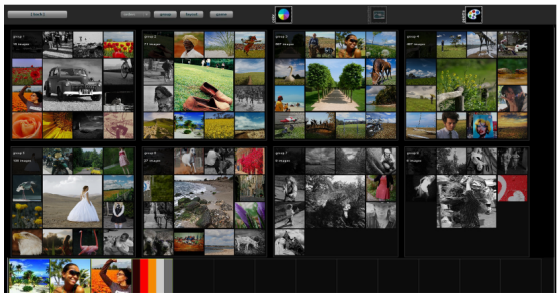


Figure 13: "Summer" images grouped by their emotion valence score.

Finally, note that visual and textual scores can be easily combined with score averaging after appropriate normalization which ensures that all the scores are between 0 and 1.

Ordering or Grouping by Image Emotion. As mentioned earlier, emotion is highly subjective and it is difficult to categorize images according to a few predefined categories such as "happy", "angry" or

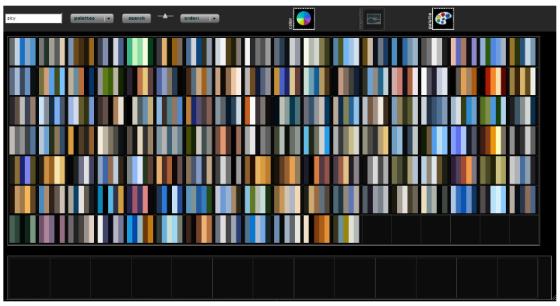
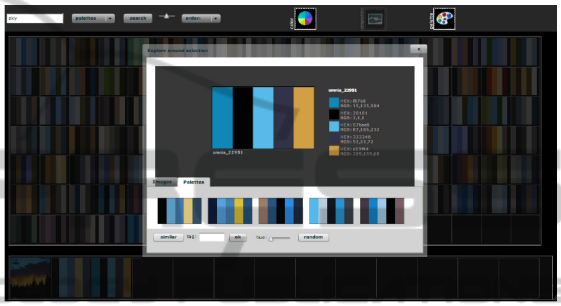
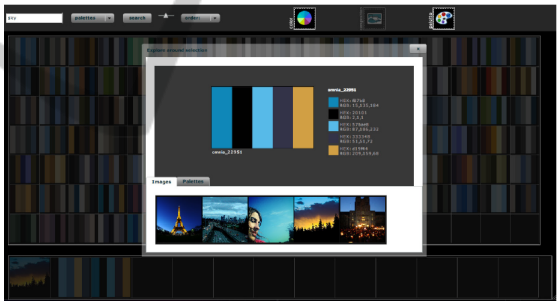


Figure 14: Retrieved "sky" color palettes.



(a)



(b)

Figure 15: Selecting a palette and visualizing its nearest neighbor (a) palettes and (b) images.

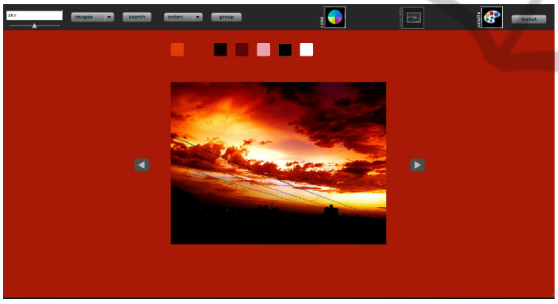
"boring". Therefore a model which allows for representing emotions in two dimensions, *valence* and *arousal*, is used (see Section 4.3). Hence we obtain a a valence and an arousal score for each image. These scores are used in a similar way as the values of the aesthetic features, and hence images are reordered and clustered according to these scores. Figure 13 shows the "summer" images of Figure 4 grouped and ordered by their valence score. The clusters are ordered from the highest, or most positive, scores towards their lowest, or most negative, scores.

Browsing by Color Palettes. The OMNIA interface allows users to browse through color palettes. The idea is that given a query, instead of visualizing the images the system visualizes palettes labeled with

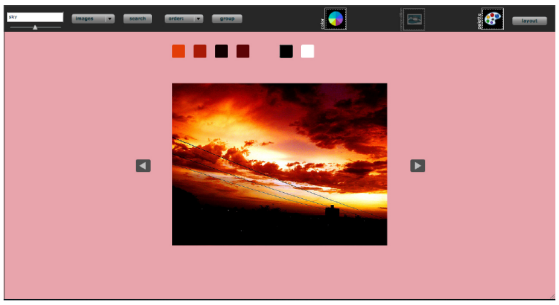
the concept given by the query. Figure 14 shows such an example where "sky" color palettes were retrieved.

As in the case of images, similar techniques, including ordering and grouping by aesthetic and emotion values, can be applied to palettes. When a palette is selected we can visualize pertinent information, which includes the RGB and hexadecimal representations of its colors, as well as its nearest neighbors as shown in Figure 15(a). We can also visualize images which have the most similar palettes as the selected one as shown in Figure 15(b). Similarity is assessed based on the closeness in Euclidean distance between the Lab color representations of the colors in the two palettes.

Selecting a Background Color for an Image. Finally, a graphic designer can set the different colors of a selected palette as backgrounds of a particular image. This functionality aids the designer in selecting a background color which fits a selected image as shown in Figure 16.



(a)



(b)

Figure 16: A sky image with the (a) second and (b) fifth color of the palette set as a background.

7 SUMMARY

This paper described OMNIA, a system and interface for searching in multimodal image collections. The different tools OMNIA includes are based on extracting different types of asset features, which are content, aesthetic, and emotion. It was shown how visual-based features are used to classify assets by content and by emotion. In addition, these features can be used to rank images by different aesthetic concepts. It was then shown how text-based features are used to classify images by content. In addition, text-based features can be combined with visual-based features to achieve better classification. The paper also described the OMNIA system which includes the tools and interface. Finally, the utility of OMNIA in simple, efficient and intuitive search is demonstrated through different applications in graphic design. In the future, we plan on conducting user studies to show the impact of OMNIA on the everyday life of graphic designers.

ACKNOWLEDGEMENTS

This work was supported by the Agence Nationale de la Recherche through the OMNIA project (ANR-07-MDCO- 009-02).

REFERENCES

Ah-Pine, J., Clinchant, S., Csurka, G., and Liu, Y. (2009). XRCE's participation to ImageCLEF 2009. In *Proc. of the Working Notes of the 2009 CLEF Workshop*, Crete, Greece.

Ah-Pine, J., Clinchant, S., Csurka, G., Perronnin, F., and Renders, J.-M. (2010). Leveraging image, text and cross-media similarities for diversity-focused multimedia retrieval. In *The Information Retrieval Series*. Springer.

Barnard, K., Duygulu, P., de Freitas, N., Forsyth, D., Blei, D., and Jordan, M. (2004). Matching words and pictures. *Journal of Machine Learning Research*, 3:1107–1135.

Boitet, C., Boguslavskij, I., and Cardeosa, J. (2009). An evaluation of UNL usability for high quality multilingualization and projections for a future UNL++ language. In *Proc. of the International Conference on Computational Linguistics and Intelligent Text Processing*, pages 361–373.

Chen, Y. and Wang, J. (2002). A region-based fuzzy feature matching approach to content based image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24:1252–1267.

Chum, O., Perdoch, M., and Matas, J. (2009). Geometric min-hashing: Finding a thick needle in a haystack. In *Proc. of the IEEE Conference on Comptuer Vision and Pattern Recognition*.

Colmerauer, A. (1970). Les systmes-q ou un formalisme pour analyser et synthtiser des phrases sur ordinateur. *dpartement d'informatique de l'Universit de Montral, publication interne*, 43.

Columbo, C., Bimbo, A. D., and Pala, P. (1999). Semantics in visual information retrieval. *IEEE Multimedia*, 6(3):38–53.

Csurka, G., Dance, C., Fan, L., Willamowski, J., and Bray, C. (2004). Visual categorization with bags of keypoints. In *Proc. of the ECCV Workshop on Statistical Learning for Computer Vision*.

Csurka, G., Skaff, S., Marchesotti, L., and Saunders, C. (2010). Learning moods and emotions from color combinations. In *Proc. of the Indian Conference on Computer Vision, Graphics, and Image Processing*.

Daoud, D. (2006). *Il faut et on peut construire des systmes de commerce lectronique interface en langue naturelle restreints (et multilingues) en utilisant des mthodes orientes vers les sous-langages et le contenu*. PhD thesis, Université Joseph Fourier.

Datta, R., Joshi, D., Li, J., and Wang, J. (2008a). Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys*, 40(2):1–60.

Datta, R., Joshi, D., Li, J., and Wang, J. Z. (2006). Studying aesthetics in photographic images using a computational approach. In *Proc. of the European Conference on Computer Vision*, volume 3, pages 288–301.

Datta, R., Li, J., and Wang, J. Z. (2008b). Algorithmic inferencing of aesthetics and emotion in natural images. In *Proc. of the IEEE International Conference on Image Processing*, San Diego, CA.

Davis, B. and Lazebnik, S. (2008). Analysis of human attractiveness using manifold kernel regression. In *Proc. of the IEEE International Conference on Image Processing*.

Dellandréa, E., Liu, N., and Chen, L. (2010). Classification of affective semantics in images based on discrete and dimensional models of emotions. *Proc. of the International Workshop on Content-Based Multimedia Indexing*, pages 1–6.

Desolneux, A., Moisan, L., and Morel, J.-M. (2004). *Seeing, Thinking and Knowing*, chapter Gestalt Theory and Computer Vision, pages 71–101. A. Carsetti ed., Kluwer Academic Publishers.

Dunker, P., Nowak, S., Begau, A., and Lanz, C. (2009). Content-based mood classification for photos and music. *Proc. of the ACM International Conference on Multimedia Information Retrieval*, pages 97–104.

Falaise, A., Rouquet, D., Schwab, D., Blanchon, H., and Boitet, C. (2010). Ontology driven content extraction using interlingual annotation of texts in the OMNIA project. In *Proc. of the International Workshop On Cross Lingual Information Access*, Peking, China.

Fedorovskaya, E., Neustaedter, C., and Wei, H. (2008). Image harmony for consumer images. In *Proc. of the IEEE International Conference on Image Processing*.

Flickner, M., Sawhney, H., Niblack, W., Ashley, J., Huang, Q., Dom, B., Gorkani, M., Hafner, J., Lee, D., Petkovic, D., Steele, D., and Yanker, P. (1995). Query by image and video content: the QBIC system. *IEEE Computer*, 28:23–32.

Guillaumin, M., Mensink, T., Verbeek, J., and Schmid, C. (2009). Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. In *Proc. of the IEEE International Conference on Computer Vision*.

Itten, J. (1961). The art of color. *Otto Maier Verlab, Ravensburg, Germany*.

Jacobsen, T., Schubotz, R. I., Hfel, L., and v. Cramon, D. Y. (2006). Brain correlates of aesthetic judgment of beauty. *NeuroImage*, 29:276–285.

Jegou, H., Douze, M., and Schmid, C. (2008). Hamming embedding and weak geometric consistency for large scale image search. In *Proc. of the European Conference on Computer Vision*.

Jeon, J., Lavrenko, V., and Manmatha, R. (2003). Automatic image annotation and retrieval using cross-media relevance models. In *Proc. of the Annual ACM SIGIR conference on Research and development in informaion retrieval*.

Jesus Cardeosa et al. (2009). The U++ consortium (accessed on september 2009). http://www.unl.fi.upm.es/consorcio/index.php.

Kasutani, E. (2007). Image retrieval apparatus and image retrieving method, US Patent application.

Laaksonen, J., Koskela, M., and Oja, E. (2002). Picsom self-organizing image retrieval with mpeg-7 content descriptions. *IEEE Transactions on Neural Networks*, 13:841–853.

Li, X., Chen, L., Zhang, L., Lin, F., and Ma, W. (2006). Image annotation by large-scale content based image retrieval. In *Proc. of the ACM International Conference on Multimedia*.

Loui, A., Wood, M. D., Scalise, A., and Birkelund, J. (2008). Multidimensional image value assessment and rating for automated albuming and retrieval. In *Proc. of the IEEE International Conference on Image Processing*.

Monay, F. and Gatica-Perez, D. (2003). On image auto-annotation with latent space models. In *Proc. of the International Conference On Multimedia.*

Müller, H., Clough, P., Deselaers, T., and Caputo, B. (2010). Imageclef- experimental evaluation in visual information retrieval. In *The Information Retrieval Series*. Springer.

Nowak, S. and Lukashevich, H. (2010). Multilabel classification evaluation using ontology information. In *Proc. of the International Conference on Multimedia Information Retrieval*, pages 35–44.

Perronnin, F. and Dance, C. (2007). Fisher kernels on visual vocabularies for image categorization. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*.

Perronnin, F., Liu, Y., Sanchez, J., and Poirier, H. (2010). Large-scale image retrieval with compressed fisher vectors. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*.

Rouquet, D. and Nguyen, H. (2009). Multilingusation d'une ontologie par des corespondances avec un lexique pivot. In *TOTh09*, Annecy, France.

Rouquet, D., Trojahn, C., Scwab, D., and Srasset, G. (2010). Building correspondences between ontologies and lexical resources. In *to be published*.

Sahbi, H., Audibert, J., and Keriven, R. (2007). Graph-cut transducers for relevance feedback in content based image retrieval. In *Proc. of the IEEE International Conference on Computer Vision*.

Smets, P. (1990). The combination of evidence in the transferable belief model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(5):447–458.

Squire, D., W.Müller, H.Müller, and Raki, J. (1999). Content-based query of image databases, inspirations from text retrieval: Inverted files, frequency-based weights and relevance feedback. In *Proc. of the Scandinavian Conference on Image Analysis*.

Swain, M. and Ballard, D. (1991). Color indexing. *International Journal of Computer Vision*, 7:11–32.

Tahir, M., Kittler, J., Mikolajczyk, K., Yan, F., van de Sande, K., and Gevers, T. (2009). Visual category recognition using spectral regression and kernel discriminant analysis. In *Proc. of the IEEE International Conference on Computer Vision Workshop on Subspace Methods*.

Tamura, H., Mori, S., and Yamawaki, T. (1978). Textural features corresponding to visual perception. *IEEE Transaction on Systems, Man, and Cybernetics*, 8(6):460.

Tsujimura, K. and Bannai, Y. (1996). Image searching method and apparatus thereof using color information of an input image, US Patent application.

Uchida Hiroshi et al. (2009). The UNDL foundation (accessed on september 2009). http://www.undl.org/.

van de Sande, K. E. A., Gevers, T., , and Smeulders, A. W. M. (2009). The university of amsterdam's concept detection system at imageclef 2009. In *Proc. of the Working Notes for the CLEF 2009 Workshop, Corfu, Greece*.

Wang, S. and Wang, X. (2005). Emotion semantics image retrieval: a brief overview. *Proc. of the International Conference on Affective Computing and Intelligent Interaction*, pages 490–497.

Wang, W. and He, Q. (2008). A survey on emotional semantic image retrieval. *Proc. of the IEEE International Conference on Image Processing*, pages 117–120.

Yang, J., Fan, J., Hubball, D., and Gao, Y. (2006). Semantic image browser: Bridging information visualization with automated intelligent image analysis. In *Proc. of the IEEE Symposium on Visual Analytics Science And Technology*.

Zhang, J., Marszalek, M., Lazebnik, S., and Schmid, C. (2007). Local features and kernels for classification of texture and object categories: a comprehensive study. *International Journal of Computer Vision*, 73(2).

Zhang, X. W. L., Jing, F., and Ma, W. (2006). Annosearch: Image auto-annotation by search. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*.