Multimedia Information Retrieval Based on Late Semantic Fusion Approaches: Experiments on a Wikipedia Image Collection

Xaro Benavent, Ana Garcia-Serrano, Ruben Granados, Joan Benavent, and Esther de Ves

***Abstract*—Main goal of this work is to show the improvement of using a textual pre-filtering combined with an image re-ranking in a Multimedia Information Retrieval task. The defined three step-based retrieval processes and a well-selected combination of visual and textual techniques help the developed Multimedia Information Retrieval System to overcome the *semantic gap* in a given query. In the paper, five different late semantic fusion approaches are discussed and experimented in a realistic scenario for multimedia retrieval like the one provided by the publicly available ImageCLEF Wikipedia Collection.**

***Index Terms*—Content-based information retrieval, multimedia information fusion, multimedia retrieval, textual-based information retrieval.**

I. Introduction

As a result of the different information sources present in a multimedia resource (video, image, audio and text), multimedia fusion has become in a very interesting field of research in recent times for Information Retrieval (IR) and search in Multimedia Databases or on the Web. In the particular case of image retrieval, both textual and visual features are usually provided: annotations or metadata as textual information, and low level features (color, texture, etc.) as visual information.

The idea behind multimedia fusion is to exploit the individual advantages of each mode, and use the different sources as complementary information to accomplish a particular search task. In an image retrieval task, multimedia fusion tries to help in solving the semantic gap problem while obtaining accurate results.

Main proposal of this paper is to present several late semantic fusion experiments that combine textual pre-filtering with visual re-ranking in order to solve the semantic gap in a Multimedia Information Retrieval (MIR) setting. The experiments have been carried out using the Wikipedia collection at ImageCLEF 2011 (http://www.imageclef.org) that contains almost 240 thousand socially annotated images in their Wikipedia articles (also provided), 50 multimedia topics and the related relevance judgments [23].

In a human search, there is an important gap between the low-level features that search engines use and the human perception (semantic gap) [22]. In queries with a great "semantic meaning" such us the ImageCLEF 2011 Wikipedia retrieval task topics [23], the TBIR (text-based image retrieval) systems can better capture the conceptual meaning of the question than CBIR (content-based image retrieval) systems [5], [13]. We use this assumption by applying a textual pre-filtering approach. Then, the CBIR system better succeeds in this semantically reduced collection avoiding false positives, that is, images visually similar from the low-level visual features but with different concept meaning. Furthermore, CBIR process will be significantly reduced, both in terms of time and computation. These textual pre-filtering techniques have been successfully used in ImageCLEF 2011 International Contest reaching the second position at the global ranking, and briefly reported in [13].

Most of the fusion techniques in image retrieval are based on combining monomodal (textual and visual) results following symmetric schemas. These strategies combine decisions coming from text and visual-based systems by mean of aggregation functions or classical combinations algorithms, which don't take into account the different semantic level of each modality. At best, some of them just use weighted factors to assign different levels of confidence to each mode. Our proposal is an asymmetric multimedia fusion strategy, which exploits the complementarity of each mode. The schema consists in a prefiltering textual step, which semantically reduces the collection for the visual retrieval, followed by a monomodal results fusion phase. Results will show how retrieval performance is improved, while the task is made scalable thanks to the significant reduction of the collection.

The experimentation performed in this paper includes the comparison of five different late fusion algorithms (Product, OWA operators, Enrich, MaxMerge and FilterN), in our case, implemented with and without the proposed pre-filtering. In the present paper current state of the art late fusion algorithms are presented under a common framework, giving a much more deep understanding of the fusion information problem.

Section II will show a brief introduction to multimedia/multimodal fusion as well as the motivation for this work and its main

Manuscript received May 10, 2012; revised February 14, 2013; accepted April 10, 2013. Date of publication June 11, 2013; date of current version November 13, 2013. This work was supported in part by Spanish projects BUSCAMEDIA under CEN-20091026, MA2VICMR under S2009/TIC-1542, and MCYT under TEC2009-12980. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Francesco G. B. De Natale.

X. Benavent, J. Benavent, and E. de Ves are with the Computer Science Department, Universidad de Valencia, Valencia 46022, Spain (e-mail: xaro.benavent@uv.es).

A. Garcia-Serrano and R. Granados are with the ETSI Informática, Universidad Nacional de Educación a Distancia (UNED), Madrid 28040, Spain (e-mail: agarcia@lsi.uned.es).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2013.2267726

objectives. Section III will review the main late fusion algorithms used in image retrieval tasks. The ImageCLEF Wikipedia collection will be described in Section IV. In Section V, the system architecture and its subsystems are presented. Section VI will describe the fusion algorithms implemented and analyzed in this work. A detailed description of the experiments carried out and their results are given in Section VII. Finally, conclusions will be presented in Section VIII.

II. Multimedia Information Retrieval

Multimedia Information Retrieval is usually addressed from a textual point of view in most of the existing commercial tools, using annotations or metadata information associated with images or videos. In this work we deal with both textual and visual information, carrying out both monomodal and multimodal experiments using different multimedia fusion techniques and algorithms.

Multimedia fusion tries to use the different media sources as complementary information to increase the accuracy of the retrieved results [15], in order to help in solving the semantic gap problem, referred to the difficulty in understanding the information that the user perceives from the low level characteristics of the multimedia data. Specifically, in the case of Image Retrieval, the semantic gap is the lack of correspondence between the information from visual features (e.g., histograms) and the interpretation of these data by a user in a certain situation (visually similar images to the query in terms of low level features can be very different in terms of meaning).

The benefits of multimedia fusion come from approaches that improve the results of the monomodal search, balancing the cost and complexity of the implementation and deployment and providing correct and complementary information to the monomodal results. When multimedia approaches are used [2], several aspects have to be taken into account in order to select the most appropriate. The complexity arises from:

1) The asynchrony in available information of resources from different media, as well as different information format that has to be considered.
2) Correlated modalities can appear in extracted low-level features (early fusion) as well as in semantic-level decisions (late fusion). Correlation may be used to reinforce a particular decision when it has been achieved from different sources. Independent modalities could be also very fruitful, as they may provide additional information when obtaining a decision.
3) According to the task to be performed, the confidence in the different media may vary.
4) The fusion process can be highly influenced by the cost and availability of the media process.

In any scenario a balance has to be taken into account when defining the fusion approach to be used. Even if we can enumerate all possible combinations of modalities, some of them are not presented in this work, so in the following we focus only on those related to the textual and visual information of the images.

A second series of important decisions when designing a multimodal fusion algorithm or technique is related to the different existing levels of fusion: early fusion (features), late fusion (decisions) or hybrid.

The early fusion approach is based on the extracted features (visual, text, audio, motion, metadata, etc.) from the different information sources, and their combination at this level. The main advantages of this level of fusion are the possibility of using the correlation between the multiple features, and that only one learning phase in the combined feature vector is required. Difficulties are related to the synchronization between the features, and with the need to represent all features in the same format before fusion.

Fusion at the decision level, or late fusion, consists in combining the individual decisions obtained based on each of the monomodal features. Simplicity, scalability and flexibility are the main advantages of this approach, which has been widely used for combining visual and textual information for image search processes [9]. Drawbacks are related to the failure to use correlation, and with the need for a classifier for each modality.

This work will focus its research on several late fusion algorithms, and in the way they can be used in order to exploit the combination of textual and visual information in the context of image retrieval.

III. Late Fusion in Image Retrieval

In any Image Retrieval task it is well known that text-based search is usually more efficient than visual-based one [5]. However, it is also known that when it is possible to combine textual and visual information in the correct way, taking advantage of each one of the modalities, the combination will be beneficial to multimedia retrieval [4]. Because of the problem of the semantic gap, the obtaining of good results is very difficult for CBIR systems, but "content-based methods can potentially improve retrieval accuracy even when text annotations are present by giving additional insight into the media collections" [17].

Within the task of Image Retrieval, where both visual and textual information are available, late multimedia fusion approaches are based on combining the evidence from both the TBIR and CBIR subsystems. These decisions will be in the form of numerical similarities (scores). Most basic fusion techniques use these scores (denoted hereinafter as St from textual-based retrieval and Si from the visual-based) and merge them by means of aggregation functions. Late fusion algorithms between text and visual modalities are known to perform better than those of early fusion [7].

Some late fusion strategies have been proposed in the literature [26]. For example, combination rules such as *combMAX* (maximum combination), *combSUM* (sum combination) and *combMNZ* (product of maximum and non-zero numbers) were first proposed in [10]. *combMAX* computes the fused score as the maximum value obtained from all of the results lists to be combined. In the case of textual and visual subsystems it will be the maximum between St and Si. *combSUM* computes the combined score as the sum of all the monomodal scores $(\mathrm{St} + \mathrm{Si})$. Finally, *combMNZ* was built to give more importance to the objects retrieved by several subsystems.

Other kind of late fusion algorithm was inspired by voting systems. For example, the *Borda-fuse* [1] idea consists in voting with a linear penalization based on the rank, assigning a point

```
<?xml version="1.0" encoding="UTF-8" ?>
<image id="61364" file="images/7/61364.jpg">
<name>Diving emperor penguin-2.jpg</name>
<textxml:lang="en">
<description>A Emperor Penguin dives into the water somewhere in
Antarctica . Emperor Penguins can dive as deep as 550 meters, with
20 minutes of air in a single
breath.</description><comment/><caption/> </text>
<textxml:lang="de">
<description>EinKaiserpinguin der irgendwo in der Antarktis ins
Wasserspringt.
Kaiserpinguinekönnenbiszu 550 Meter tieftauchen und 20
MinutenunterWasser
bleiben.</description>
<comment>EinKaiserpinguin der irgendwo in der Antarktis ins
Wasserspringt.
Kaiserpinguinekönn)</comment>
<caption article="text/de/1/404028">EinKaiserpinguinspringt ins
Wasser </caption>  </text>
<textxml:lang="fr">
<description>Un manchotempereurplongedansl'eauquelque part en
Antarctique . Les
manchotsempereurspeuventplongerjusqu'àuneprofondeur de 550
mètres et rester plus de 20 minutes dansl'eau sans reprendreleur
respiration.</description>
<comment/><caption/></text>
<comment>({{assessments|com=1|eswiki=1}} {{picture of the
day|year=2008|month=08|day=31}} == Summary == {{Information
|Description= {{de|Ein Kaiserpinguin der irgendwo in der Antarktis
ins Wasserspringt. Kaiserpinguine könn)</comment>
<license>Public Domain</license></image>
```

Fig. 1. Image in the collection and metadata. Source http://imageclef.org/2010/wiki.

to the last ranked object from each modality, two points to the penultimate, and so on, generating the final fused list. Another example is *Condorcet-fuse* [19], which is based on a pair-wise comparison that compares each object with all the rest and assigns a point for each win. Adding up these points will generate the ranking for the fused list.

A widely used combination technique is the so called *image re-ranking*, consisting of an initial step where the textual subsystem retrieves a set of ranked objects, followed by a reorder step of these objects according to the visual score (Si). In this work the reorder step is carried out by the CBIR subsystem, which computes the visual scores (Si) working only over the subset of selected objects by the TBIR subsystem. This idea was first used by our group at the ImageCLEF 2010 edition [3], and we call it *a textual pre-filter* to constrain the whole image collection by cutting out those images with no textual similarity with the queries. The combination of this textual pre-filtering and late fusion algorithms was also used at the same time at the ImageCLEF 2011 edition by the Xerox group who called it Late Semantic Combination [6].

Another important issue to take into account, when fusing together different ranked results lists, is the normalization of the scores from each modality [24]. The performance using normalized scores for fusion depends highly on the score definition of each run [26]. The normalization has also been analyzed for

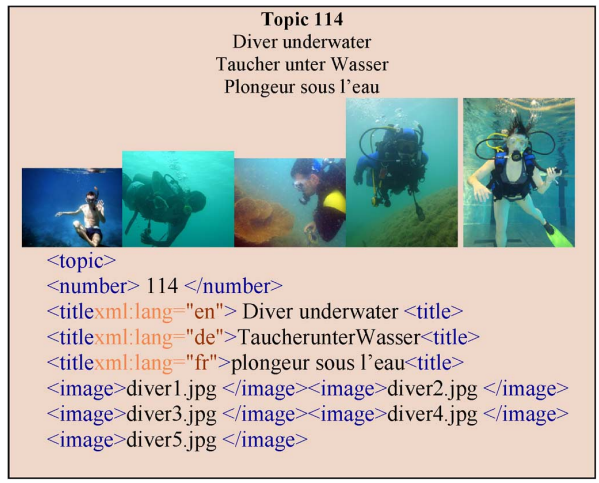

Fig. 2. Topic 114 (Multimodal query). Example taken from http://imageclef.org/2011/wikipedia/.

our scenario, and our proposal is to use an independent normalization for each modality in order to rely on their own confidence, although keeping the same value range for the different modalities.

IV. ImageCLEF Wikipedia Collection

The collection used for the experiments included in this paper[1] [20], were built using the traditional TREC-style methodology[2] to ensure representation, quantity, visual quality (high resolution, clarity and contrast) and semantically rich image annotations.

The Wikipedia image database [21] is annotated with user-generated textual descriptions of variable quality and length, and will be used under copyright. The database consists of 237,434 images and their associated user-supplied annotations. Images are associated with unstructured and noisy textual information in English, German and French [23]. The collection is completed with a set of topics and its corresponding relevance judgments (ground truth).

The Wikipedia task organizers define a topic as the description of multimedia information needs that contain textual and visual hints [21]. Each topic has a textual description of the query in three languages (English, French and German) and four or five image examples, chosen so as to illustrate the visual diversity of the topic. Fig. 2 illustrates a topic from the collection. Participants in ImageCLEF should address a query-based retrieval, in order to obtain relevant images from the collection.

Fig. 1 shows an image example from the collection, along with its metadata. Textual information includes several fields that describe the given image. The name of the image, as found in the Wikipedia Commons repository, is given in the field <name>. When the language of the annotations is identified, <description>, <comment> and <caption> are provided for that language. The caption is the text that accompanies the image in Wikipedia articles, which are provided and linked inside the <caption> element. Another <comment> element, in this case

[1]Publicly available at http://imageclef.org/wikidata since 2011 campaign

[2]http://trec.nist.gov/presentations/TREC2004/04intro.pdf

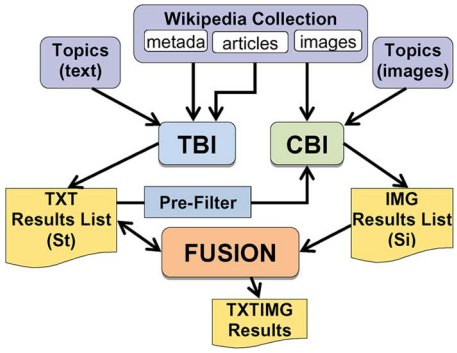

Fig. 3. Environment overview.

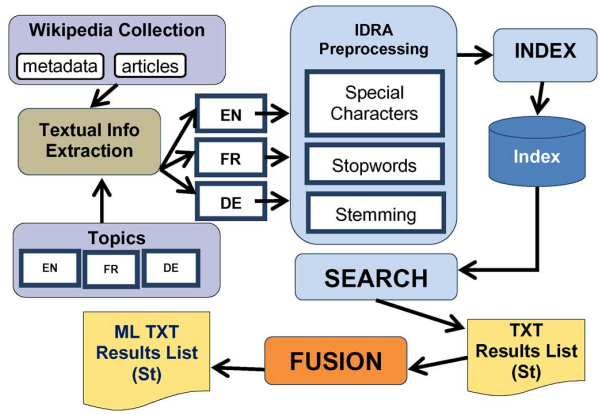

Fig. 4. TBIR sub-system overview.

language independent, is provided within metadata, giving a raw annotation of the image. This text can be in one or more languages, not necessarily English, German or French.

The benefits of the collections provided by ImageCLEF are also based on the difficulty in creating realistic and effective queries that fit with user's information needs as well as with evaluation purposes. TREC (http://trec.nist.gov) and ImageCLEF (among others) mostly use a pooling approach to assess relevance. ImageCLEF topics [14] are created mainly by focusing on (a) formulating information needs in different media such as image and text and (b) trying to fit some of the following goals: correspondence with a specific user model (a searcher in a given context), correspondence to real needs of IR systems, diverse results, and solvable with the given database. A set of 50 topics was developed in order to respond to diverse multimedia information needs at the ImageCLEF 2011 Wikipedia image retrieval task edition.

A set of relevant judgments for every topic is also given for the accuracy of the evaluation results of the experiments. The relevant judgments for these topics were created by assuming binary relevance (relevant vs. non relevant) and by assessing only the images in the pools created by the retrieved images contained in the runs submitted by the participants with a pool depth of 100.

V. Description of the Developed Environment

A. Architecture Description

To carry out the experiments, a three-subsystem architecture was developed (Fig. 3): TBIR (Text-Based Image Retrieval), CBIR (Content-Based Image Retrieval), and Fusion subsystem. A simple interface was also developed to perform these experiments.

Both the textual (TBIR) and the visual subsystem (CBIR) obtain a ranked list of images based on a similarity scores (St and Si) for a given query. Firstly, TBIR uses the textual information from the annotations (metadata and articles) to obtain these scores (St). This textual pre-filtered list is then used by the CBIR sub-system. It extracts the visual information from the given example images of the topic and generates a similarity score (Si). The fusion sub-system is in charge of merging these two lists of results, taking into account the scores and rankings, in order to obtain the final result list.

The TBIR subsystem is based on the well-known Vector Space Model[3] and TF-IDF measurement[4]. The IDRA tool [11] is used for preprocessing the textual information associated with the images in the collection. *Lucene*[5] is used to index a basic configuration. To configure, implement and compare textual results, the IDRA tool and the *trec_eval* tool[6] are also used.

The CBIR subsystem uses the CEDD[7] low-level features together with its own logistic regression relevance feedback algorithm to get the score (Si) for each image.

B. Text-Based Information Retrieval (TBIR) Sub-System

This sub-system (Fig. 4) is in charge of retrieving relevant images for a given query taking into account the textual information available in the collection. Different steps are required in order to accomplish this task: information extraction, textual preprocessing, indexation and retrieval. A text-based ranked results list of images will be obtained, containing the relevance or score (St) of the retrieved images for the concrete query.

Textual Information Extraction: Two different textual information sources can be used in the collection: the metadata and the articles files. The metadata XML tags extracted in the experiments presented are: the <name> and the general <comment> for all languages, and <description>, <comment> and <caption> for each particular language (English, French and Dutch). The <caption> tag may include a link to the article/s from Wikipedia in which the images appear. Based on an analysis of the Wikipedia articles structure, only <title> and <categories> fields from these articles will be extracted and included as part of the image textual description. Using all the textual information available in Wikipedia articles will introduce noise in the TBIR system, since these may be related with a more general concept that the one in a particular image. The <title> field is selected because it contains the name of the article and it can be considered as useful general information. The <categories> fields are intended to group together pages on similar subjects,

[3]http://en.wikipedia.org/wiki/Vector_space_model

[4]http://en.wikipedia.org/wiki/Tf-idf

[5]http://en.wikipedia.org/wiki/Lucene

[6]http://trec.nist.gov/trec_eval/

[7]http://www.imageclef.org/wikidata

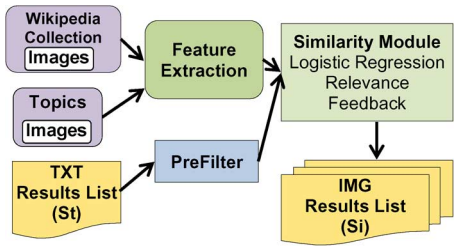

Fig. 5. CBIR sub-system overview.

so we thought that its textual information will be somehow related with the image. Several experiments confirm this idea, obtaining a relative improvement of 10–20 % for all the three languages in the collection when adding this textual information from Wikipedia articles where images appear in.

The final output of this component is the selected textual information describing the images, coming from both the metadata and the articles, or the text corresponding to the topics. In both cases it will be separated by language: English French or Dutch.

Textual Preprocessing: This component (included in the IDRA tool) processes the selected text in three steps: 1) characters with no statistical meaning, like punctuation marks or accents, are eliminated; 2) exclusion of semantic empty words (*stopwords*) from specifics lists for each language; and 3) stemming or derived words to their stem. Experiments evaluating the benefits of applying stemming to the extracted text were developed, and the obtained results recommend its use for the three languages (not only English). The extracted textual information associated with the topics is also preprocessed.

Indexation: Once the extracted textual information is preprocessed, it is indexed using *Lucene*, following a basic implementation that uses the *WhiteSpaceAnalyzer*, which just separates tokens and does not apply any other linguistic preprocess to the text. Monolingual indexations are carried out for each of the languages (English, French and Dutch). A fourth multilingual index is also created by indexation together the text from all the three languages.

Search: Preprocessed topic texts are launched against the index, obtaining the textual (TXT) results list with the retrieved images ranked by their similarity score (St). Queries could be launched separated by language as monolingual queries, or all together as multilingual ones. Depending on this, we will obtain a monolingual or a multilingual result list of images. In the case of monolingual experiments, a second type of multilingual (ML) result list is obtained by merging the three monolingual results (from English, French and Dutch), following a late fusion algorithm (MaxMerge). This algorithm will be also used in another multimedia fusion experiment, so it will be described in depth later in the corresponding section.

C. Content-Based Information Retrieval (CBIR) Sub-System

The CBIR sub-system (Fig. 5) is in charge of retrieving a list of relevant images taking into account the image examples given by the topic. At the ImageCLEFwiki2011 five example images are given for each topic (see Fig. 2). The two main steps of the CBIR sub-system are: the feature extraction and the similarity module. The CBIR sub-system ranks an image result list based on the image score (Si) for each given query.

Feature Extraction: The visual low-level features for all the images in the database for the example images for each topic are extracted using the CEDD [4] given by the ImageCLEF2011 organization. The CEDD descriptors, which include more than one feature in a compact histogram (color and texture information), belong to the family of Compact Composite Descriptors. The structure of CEDD consists of 6 texture areas. In particular, each texture area is separated into 24 sub-regions, with each sub-region describing a color. The CEDD color information comes from 2 fuzzy systems that map the colors of the image in a 24-color custom palette. To extract texture information, the CEDD uses a fuzzy version of the five digital filters proposed by the MPEG-7 EHD. The histogram is normalized within the interval [0,1] and for binary representation in a three bits per bin quantization. The most important attribute of CEDDs is the achievement of very good results with better performance than the similarly-sized MPEG-7 descriptors [4]. The evaluation of the CEDD descriptors performance has been tested at different image databases (WANG's, MPEG-7 CCD, UCID43, img(Rummager) and Nister database) [4].

The CBIR subsystem could also work with any other kind of low-level features at the state of the art [8], such as Tamura, Gabor, global color histogram, MPEG global features. These global features have a similar performance to the CEDDs features [4]. Also, local features as SIFT [18] within the bag-of-words model could also be used. These approaches are more oriented to object recognition, and are complicated models from a large set of training data. They require an enormous amount of training data and lead to tremendous computing times to create these models. These reasons make them almost impracticable for general CBIR systems, although CBIR community is very interested in them.

Similarity module: The similarity module uses our own logistic regression relevance feedback algorithm [16] to calculate the Similarity (Si) of each of the images of the collection to the query.

The algorithm calculates the probability of an image belonging to a set of those images sought by the query, and models the *logit* of this probability as the output of a generalized linear model whose inputs are the visual low-level image features. The algorithm needs examples and counter-examples (positive and negative images). The positive images are the example images of the topic given by the organization (five at the Wikipedia2011 edition). As ImageCLEF does not provide any set of non-relevant images, the M counter-examples are obtained by applying a procedure, which chooses J random images from the whole database. The Euclidean distance ranks these J images, and the latest M images are taken as negative examples. For the pre-filtering approaches, the M counter-examples set is composed by those images filter out by the textual approach.

We will explain the way the logistic regression relevance feedback algorithm works. Let us consider the (random) variable Y giving the user evaluation where $Y = 1$ means that the image is positively evaluated and $Y = 0$ means a negative

evaluation. Each image in the database has been previously described by using low-level features in such a way that the $j-th$ image has the k-dimensional feature vector x_j associated. Our data will consist of (x_j, y_j), with $j = 1....n$, where n is the total number of images, x_j is the feature vector and y_j the image evaluation ($1 = \text{positive}$ and $0 = \text{negative}$). The image feature vector x is known for any image and we try to predict the associated value of Y. In this work we have used a logistic regression where $\mathrm{P(Y = 1|x)}$ i.e. the probability that $\mathrm{Y} = 1$ (the image is positively evaluated) given that the feature vector x is related to the systematic part of the model (a linear combination of the feature vector) by means of the *logit* function. For a binary response variable Y and p explanatory variables $x_1....x_p$ the model for $\pi(\mathrm{x}) = \mathrm{P(Y = 1|x)}$ at values $\mathrm{x} = (\mathrm{x_1} \ldots \mathrm{x_p})$ of predictors is $\text{logit}\,[\pi(\mathrm{x})] = \alpha + \beta_1 \mathrm{x_1} + \ldots + \beta_\mathrm{p} \mathrm{x_p}$. Where $\text{logit}\,[\pi(\mathrm{x})] = \ln(\pi(\mathrm{x})/(1 - \pi(\mathrm{x})))$. The model parameters are obtained by maximizing the likelihood function given by:

$$l(\beta) \prod^{n} \pi(x_i)^{y_i} \left[1 - \pi(x_i)\right]^{1-y_i} \tag{1}$$

The maximum likelihood estimator (MLE) of the parameter vector β is calculated by using an iterative method.

We have a major difficulty when having to adjust an overall regression model in which we take the whole set of variables into account because the number of selected images (the number of positive plus negative images, k) is typically smaller than the number of characteristics ($\mathrm{k} < \mathrm{p}$). In this case the adjusted regression model has as many parameters as the amount of data and many relevant variables could be not considered. In order to solve this problem our proposal is to adjust different smaller regression models: each model considers only a subset of variables consisting of semantically related characteristics of the image. Consequently each sub-model will associate a different relevance probability to a given image and we face the question of how to combine them in order to rank the database according to the user's preferences.

This problem has been solved by means of an ordered averaged weighted operator (OWA) [25]. The general procedure is described in detail in the following:

INPUTS:

- Let I_i be the set of images in the DB, $i \in \{1, ..N\}$.
- Let $\mathrm{ZP^p} = (\mathrm{x_1^p}, \mathrm{x_2^p}, \ldots, \mathrm{x_k^p}, \mathrm{y_p}))$ be the set of descriptor vectors in $\mathrm{R^k}$ of the chosen positive images (of size L) plus the evaluation $(\mathrm{y_p} \leftarrow 1)$.
- Let $\mathrm{ZN^n} = (\mathrm{x_1^n}, \mathrm{x_2^n}, \ldots, \mathrm{x_k^n}, \mathrm{y_n})$ be the set of descriptor vectors in $\mathrm{R^k}$ of chosen negative images (of size K J plus the evaluation $((\mathrm{y_n} \leftarrow 1)$.
- $\mathrm{g_o}$ the number of group of characteristics, $\mathrm{g_o} \leq \mathrm{k}$.

PARAMETER ESTIMATION by MLE:

- Initialization:

$$ini \leftarrow 1$$
$$fi \leftarrow \left[\frac{k}{g0}\right]$$

for $r = 1, ldots, \mathrm{g_o}$ **do**

- Build the matrix of data $\boldsymbol{W}_{Mx[k/g\mathrm{o}]} = (\boldsymbol{WP}; \boldsymbol{WN})$ with positive and negative examples $(\mathrm{M} = \mathrm{L} + \mathrm{K})$ for each group of characteristics:

$$\boldsymbol{WP} = \left(x_{ini}^p, \ldots x_{fi}^p, y_p\right)$$
$$\boldsymbol{WN} = \left(x_{ini}^n, \ldots, x_{fi}^n, y_n\right)$$

- Estimate the parameters $\overline{\boldsymbol{\mu}} = (\alpha, \beta_1, \ldots, \beta_{(k/g\mathrm{o})})$ for data $\mathbf{W}$ by MLE estimator that implies compute the maximum of this expression:

$$l(\mu) = \prod_{1}^{M} \pi(x_i)^{y_i} \left[1 - \pi(x_i)\right]^{1-y_i}$$
$$\text{where} \quad \ln \frac{\pi(x)}{1 - \pi(x)} = 1 + \alpha + \beta_1 + \ldots + \beta_{\frac{g0}{k}}$$

- Predict the probability $\pi_r(I_i)$ for each image in the database with the estimated parameters $\overline{\mu}$.
- Update:

$$ini \leftarrow fi + 1$$
$$fi \leftarrow r \times \left[\frac{k}{g_0}\right]$$

end

OUTPUT:

- An **OWA** operator is used to fusion the probability vector $\boldsymbol{\pi}(I_i) = (\pi_1(I_i), \ldots, \pi g\mathrm{o}(I_i))$ in an unique score:

$$\pi_{fi}(I_i) = OWA\left(\boldsymbol{\pi}(I_i)\right)$$

D. Multimodal Late Fusion Sub-System

The five different late fusion algorithms evaluated in this paper will fuse together two ranked results lists of relevant images (with respect to a topic or query) into a final one. They can be classified into two classes:

1) Fusion based on the generation of a new relevance score from initial scores (St or Si) of the retrieved images for each query in the different results lists taking into account (textual, visual or multimodal retrieval), and
2) Fusion based only in the position of the results, acting as a merging technique and re-ranking.

In the first class, four algorithms are evaluated: the Product, the OWA Operator, the Enrich and the MaxMerge. For the second class the FilterN is presented.

The first set of experiments uses late fusion algorithms without textual pre-filtering for the CBIR sub-system. It means that the CBIR sub-system is based on the pure visual baseline. Meanwhile, the second set uses the pre-filtered or reduced image database of the collection. It also means that the CBIR sub-system uses the textual pre-filtered list to obtain the image scores (it is the so-called *re-ranking*). The filtering step consists of delimiting the full images database to those images which obtain $\mathrm{St} > 0$, that is, the filter eliminates those images with

no relevance according to the TBIR subsystem. That means that only images with some probability of being relevant for a specific query will be taken into account by the CBIR subsystem. One of the main goals of this paper is to prove the initial assumption that the use of this filter is based on the fact that textual subsystem initially captures the conceptual meaning of a topic in a better way than the visual one. Text and pure visual-based experiments corroborate this.

VI. Late Fusion Algorithms

In the following, some details about the five fusion algorithms selected, implemented and experimented are given and organized in two categories, one for those based on the relevance score obtained and other for the ones based on re-ranking. A third subsection it is included in order to show the relevant details about score normalization in the experiments.

A. Late Fusion Based on Relevance Scores

Product ($\mathbf{Si^*St}$): two results lists are fused together to combine the relevance scores of both textual and visual retrieved images (St and Si). Both subsystems will have the same importance for the resulting list: the final relevance of the images will be calculated using the Product. Notice that the Product simulates the filtering when St is 0 (no relevant image for the query), so the image will never appear in the fused list ($\mathrm{St^*Si}$ is 0).

OWA Operators: The Ordered Mathematical Aggregation operator [25] OWA transforms a finite number of inputs into a single output.With the OWA operator no weight is associated with any particular input; instead, the relative magnitude of the input decides which weight corresponds to each input. In our application, the inputs are the textual and image scores (St and Si), and this property is very interesting because we do not know, a priori, which subsystem will provide us the best information.

As OWA operators are bounded by the max (an OR operator), and the min operators (the AND operator), Yager [25] introduced a measure called *orness* to characterize the degree to which the aggregation is like an *or* (max) operation:

$$orness(w) = \frac{1}{n-1}\sum_{i=1}^{n}(n-i)w_i \tag{2}$$

Notice that OWA operators with many of the weights close to their highest values will be *or-like* operators ($orness(W) <= 0.5$), while those operators with most of the weights close to their lowest values will be *and-like* operators ($orness(W) >= 0.5$).

The aggregation weights used for the experiments cover all the range weight values from the minimum to the maximum with the different *Orness* weights used (Table I). For example, an *Orness(0.3)* means that a weight of 0.3 is given to the higher probability value and a weight of 0.7 to the lower probability of the inputs (St,Si).

Enrich: this strategy uses two results lists, a main list (from the textual module) and a support list (from the visual one). If a certain result appears in both lists for the same query, the relevance of this result in the fused list will be increased in the following way:

TABLE I
OWA Aggregation Weights Used for the Experiments

w	f(St,Si)
(0,1)	Min(St,Si)
(0.1,0.9)	Orness01
(0.2,0.8)	Orness02
(0.3,0.7)	Orness03
(0.4,0.6)	Orness04
(0.5,0.5)	Average(St,Si)
(0.6,0.4)	Orness06
(0.7,0.3)	Orness07
(0.8,0.2)	Orness08
(1,0)	Max(St,Si)

$$newRel = mainRel + \frac{supRel}{posRel + 1} \tag{3}$$

where *newRel* is the relevance value in the fused list, *supRel* is the relevance value in the support list (Si), *mainRel* is the relevance value in the main list (St) and *posRel* is the position in the support list. Relevance values will be normalized from 0 to 1. Every result appearing in the support list but not in the main one (for each query) will be added at the end of the mixed list. In this case, the relevance values will be normalized according to the lower value in the main list.

MaxMerge: this algorithm selects from the lists to merge those retrieved images with a higher relevance or score for a specific query, independently of the subsystem (textual or visual) they belongs to.

B. Late Fusion Based on Ranking

FilterN: this algorithm is used to remove from the textual results list those images not appearing in the first N results of the visual list. The idea is to eliminate the images that the visual module is not very sure of; those with a low score Si. This technique will try to clean the textual results based on the visual ones.

C. Normalization of the Scores (St, Si)

As it has been described at related work the normalization is an important issue [24]. We propose that different value modalities be within the same range of between 0 and 1, but the normalization will be independent for each one. The idea is that each module obtains its score between the predefined ranges in order not to normalize to the maximum for each independent query. The objective of this idea is that each module relies on its own confidence.

The score obtained from the CBIR subsystem (Si) is the probability that a given image belongs to a certain set of images. The probability value is between 0 and 1, so that it does not need to be normalized. A Si score would be 1 if the CBIR sub-system were completely sure that the image belongs to the set, so that the probabilities obtained by the CBIR system are close to 1 if they are very similar to the given examples.

To obtain the TBIR similarity score (St) of a document for a specific query, *Lucene* does not exactly calculate the cosine measure (as supposed in the VSM approach) for reasons of usability [2]. With $\mathrm{V}(q)$ and $\mathrm{V}(d)$ representing the weighted vectors of the textual query q and a concrete document d, the original cosine function normalizes the dot product of these vectors $(\mathrm{V}(q) \bullet \mathrm{V}(d))$ by dividing it by the Euclidean norm of these vectors ($|V(q)|$ and $|V(d)|$), normalizing them to a unit vector. But *Lucene* uses a different document length normalization for $V(d)$, which normalizes it to a vector equal to or larger than the unit vector. Instead of calculating the Euclidean norm for each document, which would have a very high computational cost, *Lucene* uses the document length, which is computed when the document is added to the index in accordance with the number of tokens in the document. Also for efficient score computation, query Euclidean norm ($|V(q)|$) is computed when search starts, as it is independent of the document being scored. Normalizing the query vector $V(q)$ provides comparability (to a certain extent) of two or more queries. We have analyzed two kinds of normalization one based on the highest St for each query, and the other on the St obtained by an artificially constructed "perfect" document with the same text as the query.

VII. Experiments Carried Out and Their Results

Several experiments (the so-called runs) were defined (Table II) with the aim of firstly evaluating the improvement of using the textual pre-filtering step combined with the visual re-ranking, and secondly the performance of five late fusion algorithms. The first block experiments (green background) in Table II are the monomodal runs, the second block (white background) are the mixed runs without textual pre-filtering, and the last block (pink background) are the mixed runs with textual pre-filtering.

Textual and visual baselines are initially established, showing how the TBIR and CBIR systems work separately (runs 1, 2) and run3 is CBIR after textual pre-filtering, the so-called image re-ranking. Then, the five described late fusion algorithms are used to fuse these monomodal decisions. The first set of experiments (runs 4 to 18) combine the results lists of the two baseline modalities that are without textual pre-filtering. Meanwhile the second set (runs 19 to 33) uses the image re-ranking (run 3), that is with textual pre-filtering. We will show that the pre-filter step combined with image re-ranking helps to improve the fusion result, confirming the fact that the conceptual meaning of a topic is initially better captured by the textual subsystem and the fusion of the image re-ranking and the textual pre-filtered lists achieves better results than normal late fusion strategies.

The evaluation of the experiments is carried out following the TREC methodology, taking into account the first 1,000 retrieved images for each topic. The results of each of the experiments will be shown in terms of MAP (mean average precision), and precisions at different levels, P@5, P@10 and P@20.

A. Monomodal Results

Table III shows our monomodal results obtained at ImageCLEF 2011 (runs 1–3), along with the best experiment presented to the competition and the average computed over all the participating groups, for both the textual (Txt) and visual (Img) modalities [23].

The first column shows the position obtained for the runs submitted to the same Wikipedia retrieval task [23] and the position for the runs not submitted, which they would have got if submitted. In total, at the Wikipedia retrieval task 2011 edition there were 51 textual, 2 visual and 57 mixed runs submitted.

The textual baseline (run 1) is generated from monolingual retrieval for the three languages using the MaxMerge algorithm to obtain the final multilingual result list of images. Our textual run is in the 14th position of all runs (at 15% of the first total runs), and in the 3rd position for textual modality runs in terms of MAP. Early precisions (P@5, P@10 and P@20) also show how our text-based approach ($\mathrm{P@5} = 0.5600, \mathrm{P@10} = 0.5060$,

TABLE II
Experiments Description

runID	Mode	Details		
		Pre-filter	Algorithm	Fusion based on ranks
run1	Textual		-	
run2	Visual		-	
run3	Visual	✓	-	
run4	Mixed		St*Si	
run5	Mixed		OWA(min)	
run6	Mixed		OWA(Orness01)	
run7	Mixed		OWA(Orness02)	
run8	Mixed		OWA(Orness03)	
run9	Mixed		OWA(Orness04)	
run10	Mixed		OWA(Average)	
run11	Mixed		OWA(Orness06)	
run12	Mixed		OWA(Orness07)	
run13	Mixed		OWA(Orness08)	
run14	Mixed		OWA(Orness09)	
run15	Mixed		OWA(Max)	
run16	Mixed		FilterN	✓
run17	Mixed		Enrich	
run18	Mixed		MaxMerge	
run19	Mixed	✓	St*Si	
run20	Mixed	✓	OWA(min)	
run21	Mixed	✓	OWA(Orness01)	
run22	Mixed	✓	OWA(Orness02)	
run23	Mixed	✓	OWA(Orness03)	
run24	Mixed	✓	OWA(Orness04)	
run25	Mixed	✓	OWA(Average)	
run26	Mixed	✓	OWA(Orness06)	
run27	Mixed	✓	OWA(Orness07)	
run28	Mixed	✓	OWA(Orness08)	
run29	Mixed	✓	OWA(Orness09)	
run30	Mixed	✓	OWA(Max)	
run31	Mixed	✓	FilterN	✓
run32	Mixed	✓	Enrich	
run33	Mixed	✓	MaxMerge	

TABLE III
Monomodal Results in ImageCLEF 2011

Pos	Run	Mode	MAP	P@5	P@10	P@20
14	run1	Txt	0.3044	0.5600	0.5060	0.4040
111	run2	Img	0.0014	0.0060	0.0060	0.0040
107	run3	Img	0.0618	0.0880	0.0880	0.0910
-	Average	Txt	0.2169	0.4598	0.3973	0.3228
11	Best	Txt	0.3141	0.5720	0.5160	0.4270
-	Average	Img	0.0039	0.0340	0.0270	0.0245
109	Best	Img	0.0044	0.0400	0.0340	0.0280

TABLE IV
MULTIMODAL RESULTS WITHOUT TEXTUAL PRE-FILTERING

Pos	Run	Fusion	MAP	P@5	P@10	P@20
14	run1		0.3044	0.5600	0.5060	0.4040
9	**run4**	**St*Si**	**0.3300**	**0.6160**	**0.5480**	**0.4380**
17	run5	OWA(min)	0.2974	**0.5680**	0.4960	0.4010
10	**run6**	**OWA(Orness01)**	**0.3163**	**0.6040**	**0.5160**	**0.4280**
9	**run7**	**OWA(Orness02)**	**0.3249**	**0.6200**	**0.5220**	**0.4330**
13	**run8**	**OWA(Orness03)**	**0.3095**	**0.5720**	**0.5160**	**0.4140**
10	**run9**	**OWA(Orness04)**	**0.3149**	**0.6240**	**0.5460**	**0.4350**
16	run10	OWA(avg)	0.3003	**0.5760**	**0.5140**	**0.4110**
48	run11	OWA(Orness06)	0.2575	**0.6200**	**0.5320**	**0.4130**
17	run12	OWA(Orness07)	0.2906	**0.5640**	**0.5060**	**0.4050**
99	run13	OWA(Orness08)	0.1481	0.4960	0.3920	0.2670
104	run14	OWA(Orness09)	0.1009	0.4040	0.2880	0.1980
40	run15	OWA(max)	0.2689	**0.5600**	0.5000	0.3980
105	run16	FilterN	0.0713	0.3720	0.2840	0.2080
13	**run17**	**Enrich**	**0.3054**	**0.5600**	**0.5080**	**0.4040**
41	run18	MaxMerge	0.2689	**0.5600**	0.5000	0.3980
Average		-	0.2558	0.5176	0.4542	0.3678
1	Best	-	0.3880	0.7080	0.6320	0.5100

TABLE V
MULTIMODAL RESULTS WITH TEXTUAL PRE-FILTERING

Pos	Run	Fusion	MAP	P@5	P@10	P@20
14	run1		0.3044	0.5600	0.5060	0.4040
8	**run19**	**St*Si**	**0.3404**	**0.6280**	**0.5480**	**0.4530**
16	run20	OWA(min)	0.3005	0.0880	**0.5080**	**0.4170**
10	**run21**	**OWA(Orness01)**	**0.3148**	**0.6360**	**0.5180**	**0.4230**
9	**run22**	**OWA(Orness02)**	**0.3281**	**0.5760**	**0.5300**	**0.4350**
9	**run23**	**OWA(Orness03)**	**0.3350**	**0.6160**	**0.5420**	**0.4430**
9	**run24**	**OWA(Orness04)**	**0.3340**	**0.6280**	**0.5380**	**0.4500**
9	**run25**	**OWA(avg)**	**0.3235**	**0.4680**	**0.5380**	**0.4490**
14	run26	OWA(Orness06)	0.3027	**0.6160**	**0.5300**	**0.4470**
38	run27	OWA(Orness07)	0.2735	**0.6160**	**0.5120**	**0.4220**
60	run28	OWA(Orness08)	0.2351	**0.5840**	0.4680	0.3640
79	run29	OWA(Orness09)	0.1994	**0.5720**	0.4180	0.3100
93	run30	OWA(max)	0.1732	0.4680	0.3700	0.2820
38	run31	FilterN	0.2804	**0.5920**	**0.5080**	**0.4140**
17	run32	Enrich	0.2951	0.5120	0.4780	0.3960
93	run33	MaxMerge	0.1732	0.4680	0.3700	0.2820
Average		-	0.2558	0.5176	0.5600	0.4542
1	Best	-	0.3880	0.7080	0.6320	0.5100

$\mathrm{P@20} = 0.4040$) performance is similar to the best experiment presented ($\mathrm{P@10} = 0.5130$, $\mathrm{P@20} = 0.4270$). Early precision measures are considered with special interest, due to the fact that most of the users using a search system will take into account only the images retrieved in the top positions. These results reinforce our assumption that textual approach initially better captures the meaning of the query.

The visual baseline (run 2) is generated using the whole database collection while in the image re-ranking (run 3) only those images filtered by the textual subsystem was used. Table III shows that visual monomodal approaches have lower MAP values than textual monomodal results, being 0.2160 for the average textual monomodal approaches of the runs submitted at Wikipedia2011 while 0.0039 for the visual monomodal approaches. This is due to the fact that Wikipedia topics are very "semantic" questions allowing better performance to the TBIR systems than to the CBIR ones, as it is already known at literature [23]. It is important to highlight that when using the textual pre-filter: the image re-ranking, the CBIR sub-system obtains better results than when using a pure visual baseline (0.0614 MAP against 0.0014 MAP). At image re-ranking, false positives images from the visual low-level point of view have been removed from the collection by the textual pre-filter.

B. Multimodal Results Without Textual Pre-Filtering

Results for mixed or multimodal runs without pre-filtering are shown at Table IV, and those with pre-filtering at Table V. The late fusion experiments which outcome the textual baseline are highlighted in bold.

By comparing Tables IV and V it can be observed that better results are obtained using the textual pre-filtering than when not. This is mainly for two reasons: firstly, those images that are visually similar to the query example images in terms of color or texture, but do not contain semantic related information are filtered out by the textual pre-filter (see Fig. 8 and Fig. 9). Secondly, the regression relevance algorithm of the CBIR sub-system calculates more accurate image scores when working with the reduced collection than when not. The reason of this better CBIR performance is that the textual pre-filter helps to select better the inputs for the relevance algorithm (set of relevant and non-relevant images). While the relevant images are the same for the two fusion techniques, with and without textual pre-filter, the counter-examples images for the relevance algorithm are taken from the non textual pre-filtered set of images meanwhile at the non textual pre-filtering are taken by the whole collection.

It is important to notice that from a collection of 237,434 images, where just 3,440 are relevant, the textual pre-filter significantly reduces the number of images that CBIR will need to manage (5,490 images on average for each topic, which implies a reduction of 97.69% of the whole collection). This will make the work of the CBIR subsystem quicker in terms of time and computing.

The potential drawback of the use of the pre-filtering technique is that the visual and fusion subsystems depend on the good work of the textual subsystem. So that if the textual module does not retrieve certain relevant images in a given query these images are out of the scope. The recall for the textual pre-filtered baseline is 0.8597, which means that almost the 86% of the relevant images in the collection were selected by the textual pre-filter.

C. Multimodal Results With Textual Pre-Filtering

The best semantic late fusion algorithm amongst the different experiments carried out at this scope is the Product of the two modality scores ($\mathrm{St} \cdot \mathrm{Si}$) with a MAP of 0.3404 (see Table V). Scores from both modalities are taken into account with the same importance. This run was submitted at the Wikipedia2011 edition, it being the second best group in the task, in the eighth position of the overall ranking, and it being the first position with a MAP of 0.3880 for the Xerox group.

The FilterN and the Enrich algorithms do not outperform the textual baseline (see Table V). It can be inferred that algorithms that calculate the new score based on the value score modalities get better fusion results (Product, OWA) than those based on rank position information (FilterN and Enrich). The other semantic multimedia late fusion algorithms that outperform the textual baseline in all evaluation measurements are the ones based on the OWA operator (with different values, from the Orness01 to the Orness05). These are the more and-like operators

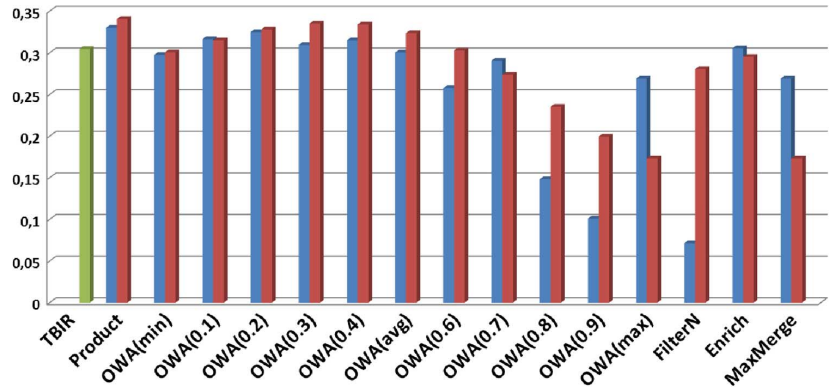

Fig. 6. Late fusion algorithms performance (MAP) with (red right bars) and without text pre-filtering (blue left bars).

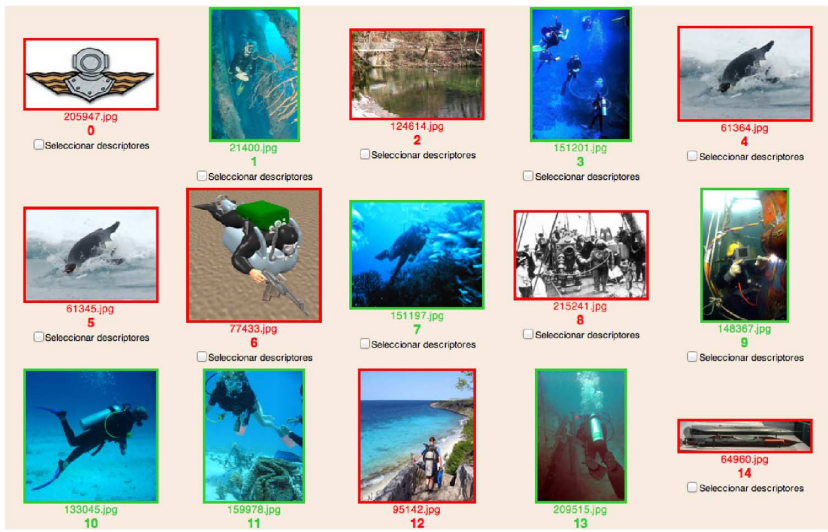

Fig. 7. Top retrieved results by the textual approach (run1) for topic 114 (Fig. 2). Images with a green frame are relevant images and images with a red frame are not relevant.

so that they are more restrictive. It means that for an image to get a high final score it needs to get higher scores for the two modalities information. The OWA operators, that do not outperform the textual baseline, are the minimum, the maximum, and the Orness06 to Orness09. As the OWA operator is less restricted, the performance declines: the minimum is the worst OWA operator because it is the least restrictive thus that it gets the lower score. The MaxMerge is like a maximum OWA so it does not outperform the textual baseline because is too restrictive.

Fig. 6 shows graphically the performance, in terms of MAP, of the different late fusion algorithms when using the textual Prefilter (red bars at the right) and when not (blue bars at the left). The green bar represents the monomodal textual approach. It can be used as a reference baseline to compare which late fusion algorithms outcome the textual baseline.

Table VI summarises the monomodal runs performance and the best multimedia late fusion algorithm: Product. The improvement achieved by the Product Late Fusion approach over the textual baseline is also calculated. Table VI shows that our proposed late semantic fusion Product algorithm overcomes the textual monomodal ones, improving MAP (11%) and precisions at the first retrieved images (11% for P@5, 8% for P@10 and 11% for P@20).

Fig. 7, 8, 9 and 10 show the top fifteen retrieved images for the topic 114 (Fig. 2): by the textual baseline (run1, Fig. 7), by the pure visual baseline (run 2, Fig. 8), by the image re-ranking (run3, Fig. 9), and by the Product multimedia fusion with textual pre-filtering (run19, Fig. 10) respectively. Images with a green frame are relevant images according to the relevance judgements, images with a red frame are not relevant and those with no frame are out of the pool.

For topic 114 (Fig. 2), which textual query information is "diver underwater", the textual baseline (see Fig. 7) retrieves those images with words annotations related to "diver" and/or "underwater". Some of the retrieved images do not refer to "diving underwater" meaning being this fact immediately seen by a human user (e.g. the first top retrieved image at Fig. 7 is an insignia for divers: or the fifth top retrieved image is a diving emperor penguin). The "insignia for divers" image is removed from the top retrieved images by using the Product multimedia late fusion algorithm (see Fig. 10).

Fig. 8 shows that all top images retrieved by the pure visual approach are false positive images from the visual low-level features point of view (e.g. the fifth image retrieved is a fish under the water similar to a diver under water taking into account color and texture information). The textual pre-filter algorithm elimi-

Fig. 8. Top retrieved results by pure visual approach (run2) for topic 114 (Fig. 2). Images with bordered not highlighted were out of the pool in the evaluation process, so they are not relevant.

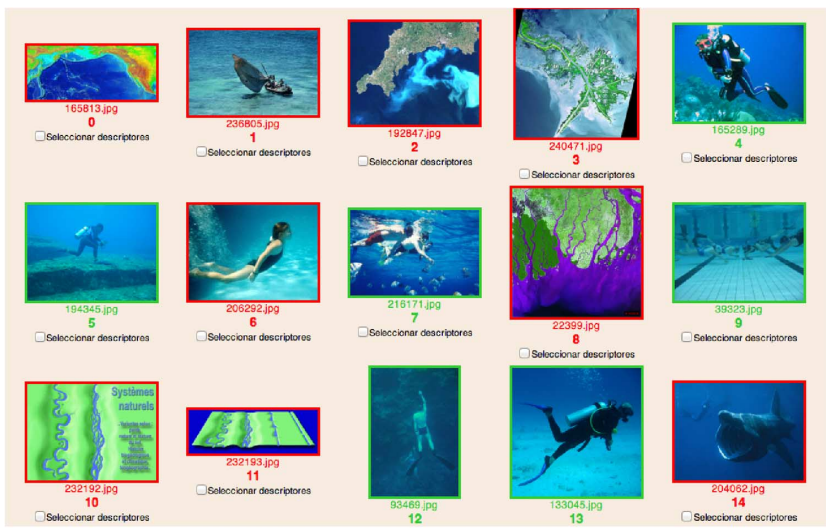

Fig. 9. Top retrieved results by image re-ranking approach (run3) for topic 114 (Fig. 2).

TABLE VI
IMPROVEMENT FOR THE LATE SEMANTIC FUSION PRODUCT ALGORITHM

Run	Mode	MAP	P@5	P@10	P@20
run1	Txt	0.3044	0.5600	0.5060	0.4040
run3	Img	0.0618	0.0880	0.0880	0.0910
run13	TxtImg	0.3404	0.6280	0.5480	0.4530
Improvement		10.57%	10.82%	7.66%	10.81%

nates these false positives images so that they not appear at the image re-ranking approach (Fig. 9).

At Fig. 9 (run 3, image re-ranking), we can see that the image re-ranking captures images similar to the given examples query which mean images with a blue background (colour features), and images with only one big object on the background (texture features). These visual features can be extracted from a diver underwater (see images with a green frame in Fig. 9); or, from other visually but not semantic images as a ship on the sea (second top retrieved image), or a girl diving underwater but without a diving mask (seventh top retrieved image). This semantic information will be inferred with the Product semantic late fusion algorithm.

These improvements point out the good performance of combining the textual pre-filter with the image re-ranking. Firstly, the textual pre-filter helps visual sub-system to get rid of the images that do not contain semantic information, but can be visually similar. Later, the visual sub-system helps the textual one to better score the images, which are visually not related to that which the semantic user, seeks. Results of the Product late fusion algorithm could also be improved if the monomodal, textual and image results were improved. The Product multimedia late fusion algorithm overcomes the textual baseline, improving P@5 from 0.4 to 1.0, P@10 from 0.4 to 0.9 and P@15 from 0.47 to 0.67. This can be checked comparing Fig. 7 and Fig. 10. The visual information makes images with annotated information related to "diving" or "underwater" terms, but that do not refer to

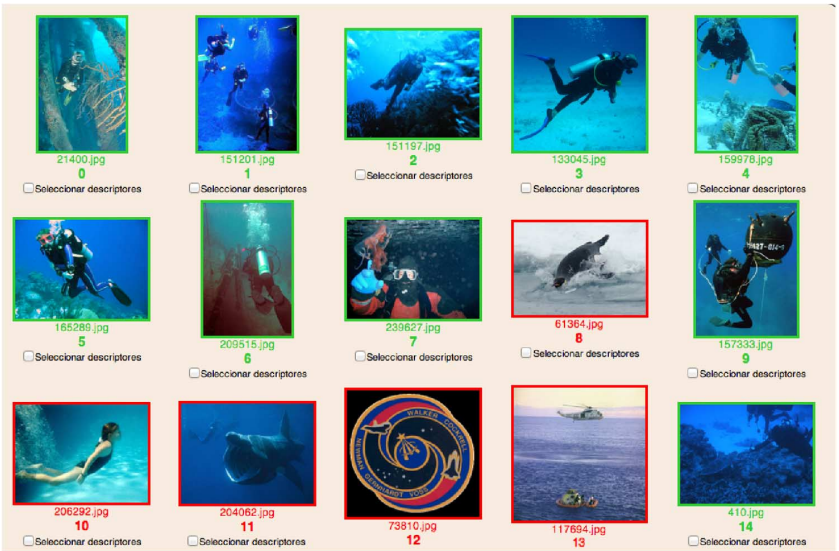

Fig. 10. Top retrieved results by the semantic Product multimedia late fusion approach (run13) for topic 114 (Fig. 2).

the "diving underwater" concept, listed down in the fusion results list. For example, image in the first top retrieved position is an insignia for divers; images in the 5th and 6th top retrieved positions are a diving emperor penguin (see Fig. 7). In a good performance, these non-relevant images are listed down from the top 15 as occurs in the Product multimedia late fusion algorithm results (see Fig. 10).

VIII. Conclusions

In the present paper a detailed description and analysis of textual pre-filtering techniques are given. These textual pre-filtering techniques reduce in a suitable way the size of the multimedia database improving the final fused retrieval results. Experiments show that the combination of textual pre-filtering and image re-ranking lists in a late fusion algorithm outperforms those without pre-filtering. It seems that textual information better captures the semantic meaning of a topic and that the image re-ranking (S_i) fused with the textual score (S_t) helps to overcome the semantic gap. Notice that all this performance improvement is carried out while significantly reducing the complexity of the CBIR process, in terms of both time and computation.

With respect to the late fusion algorithms analyzed, better results are obtained with those that work only with the value scores than others, which rely on the ranked positions. The best performance has been obtained with the Product algorithm that means that both modality scores are taken into account with the same importance. OWA restricted operators (and-like ones) also achieve well results according to the performed retrieval experiments.

Acknowledgment

The UNED-UV is a research group made up of researchers from two Universities in Spain, the Universidad Nacional de Educación a Distancia (UNED) and the Valencia University (UV), working together since 2008 [3], [11]–[13].

References

[1] J. A. Aslam and M. Montague, "Models for metasearch," in *Proc. 24th Annu. Int. ACM SIGIR Conf. Res. Develop. Inform. Retrieval*, New Orleans, LA, USA, 2001, pp. 276–284.

[2] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanballi, "Multimedia Fusion for Multimedia Analysis: A Survey," *Multimedia Syst.*, vol. 16, pp. 345–379, 2010.

[3] J. Benavent, X. Benavent, E. de Ves, R. Granados, and A. García-Serrano, "Experiences at ImageCLEF 2010 using CBIR and TBIR mixing information approaches," in *Proc. CLEF 2010*, Padua, Italy, 978-88-904810-2-4, Notebook papers.

[4] S. A. Chatzichristofis, K. Zagoris, Y. S. Boutalis, and N. Papamarkos, "Accurate image retrieval based on compact composite descriptors and relevance feedback information," *Int. J. Pattern Recog. Artif. Intell.*, vol. 24, no. 2, pp. 207–244, Feb. 2010, World Scientific.

[5] S. Clinchant, G. Csurka, and J. Ah-Pine, "Semantic combination of textual and visual information in multimedia retrieval," in *Proc. 1st ACM Int. Conf. Multimedia Retrieval*, New York, NY, USA, 2011.

[6] G. Csurka, S. Clinchant, and A. Popescu, "XRCE and CEA LIST's Participation at Wikipedia Retrieval of ImageCLEF 2011," in *CLEF 2011 Working Notes*, V. Petras, P. Forner, and P. Clough, Eds., Amsterdam, The Netherlands, Sep. 2011.

[7] A. Depeursinge and H. Müller, "Fusion Techniques for Combining Textual and Visual Information Retrieval," in *ImageCLEF: Experimental Evaluation in Visual Information Retrieval*. Berlin, Germany: Springer-Verlag, 2010, ch. 6, pp. 95–114.

[8] T. Deselaers, D. Keysers, and H. Ney, "Features for image retrieval: An experimental comparison," *Inf. Retrieval*, vol. 11, pp. 77–107, Apr. 2008.

[9] H. Escalante, C. Hernadez, L. Sucar, and M. Montes, "Late fusion of heterogeneous methods for multimedia image retrieval," in *Proc. 1st ACM Int. Conf. Multimedia Inf. Retrieval*, 2008, pp. 172–179.

[10] E. A. Fox and J. A. Shaw, "Combination of multiple searches," in *Proc. 2nd Text Retrieval Conf.*, 1993, pp. 243–252.

[11] A. García-Serrano, X. Benavent, R. Granados, E. de Ves, and J. Miguel Goñi, "Multimedia Retrieval by Means of Merge of Results from Textual and Content Based Retrieval Subsystems," in *Multilingual Information Access Evaluation II. Multimedia Experiments: 10th Workshop of the Cross-Language Evaluation Forum, CLEF 2009, Corfu, Greece, September 30 - October 2, 2009, Revised Selected Papers*. Berlin, Germany: Springer-Verlag, 2010, pp. 142–149.

[12] A. García-Serrano, X. Benavent, R. Granados, and J. M. Goñi-Menoyo, "Some results using different approaches to merge visual and text-based features in CLEF'08 photo collection," in *Evaluating Systems for Multilingual and Multimodal Information Access: 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark, September 17-19, 2008, Revised Selected Papers*. Berlin, Germany: Springer-Verlag, 2009, pp. 568–571.

[13] R. Granados, J. Benavent, X. Benavent, E. de Ves, and A. Garcia-Serrano, "Multimodal Information Approaches for the Wikipedia Collection at ImageCLEF 2011," in *Proc. CLEF 2011 Labs Workshop, Notebook Papers*, Amsterdam, The Netherlands, 2011.
[14] M. Grubinger, "Analysis and Evaluation of Visual Information Systems Performance," Ph.D. thesis, School Comput. Sci. Math., Faculty Health, Engi., Sci., Victoria Univ., Melbourne, Australia, 2007.
[15] J. Kludas, E. Bruno, and S. Marchand-Maillet, "Information fusion in multimedia information retrieval," in *AMR Int. Workshop Retrieval, User Semantics*, 2007.
[16] T. Leon, P. Zuccarello, G. Ayala, E. de Ves, and J. Domingo, "Applying logistic regression to relevance feedback in image retrieval systems," *Pattern Recog.*, vol. 40, pp. 2621–2632, Jan. 2007.
[17] M. S. Lew, N. Sebe, C. Djeraba, and R. Jain, "Content-based multimedia information retrieval: State of the art and challenges," *ACM Trans. Multimedia Comp., Commun., Appl.*, vol. 2, no. 1, pp. 1–19, Feb. 2006.
[18] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International J. Comput. Vision*, vol. 60, no. 2, pp. 91–110, 2004.
[19] M. Montague and J. A. Aslam, "Condorcet fusion for improved retrieval," in *Proc. 11th Int. Conf. Inf. Knowledge Manage.*, McLean, VA, USA, 2002, pp. 538–548.
[20] "ImageCLEF: Experimental Evaluation in Visual Information Retrieval," in *The Information Retrieval Series*, H. Müller, P. Clough, T. Deselaers, and B. Caputo, Eds. New York, NY, USA: Springer-Verlag, 2010, vol. 32.
[21] A. Popescu, T. Tsikrika, and J. Kludas, "Overview of the wikipedia retrieval task at ImageCLEF 2010," in *Proc. CLEF 2010 Labs Workshop, Notebook Papers*, Padua, Italy, 2010 [Online]. Available: clef2010.org/resources/proceedings/clef2010labs_submission_124.pdf
[22] Y. Rui, S. Huang, M. Ortega, and S. Mehrotra, "Relevance feedback: A power tool for interactive content-based image retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 8, no. 5, Sep. 1998.
[23] T. Tsikrika, A. Popescu, and J. Kludas, "Overview of the wikipedia image retrieval task at ImageCLEF 2011," in *Proc. CLEF 2011 Labs Workshop, Notebook Papers*, Amsterdam, The Netherlands, 2011.
[24] S. Wu, F. Crestani, and Y. Bi, "Evaluating score normalization methods in data fusion," in *Proc. 3rd Asia Conf. Inform. Retrieval Technol.*, 2006, pp. 642–648.
[25] R. Yager, "On ordered weighted averaging aggregation operators in multi criteria decisionmaking," *IEEE Trans. Syst., Man, Cybern.*, vol. 18, no. 1, pp. 183–190, Jan./Feb. 1988.
[26] X. Zhou, A. Depeursinge, and H. Müller, "Information fusion for combining visual and textual image retrieval," in *Proc. 20th Int. Conf. Pattern Recog.*, 2010, pp. 1590–1593.

Xaro Benavent was born in Valencia (Spain). She received the M.S. degree in Computer Science from the Polytechnic University of Valencia in 1994, and Ph.D. in Computer Science from the University of Valencia in 2001. Since 1996 she has been with the Department of Computer Science from the University of Valencia, as an Associate Professor. Her current interests are in the areas of image database retrieval and multimodal fusion algorithms.

Ana Garcia-Serrano was born in Madrid (Spain). She received the M.S. degree in Computer Science from the Technical University of Madrid (UPM) in 1983, and Ph.D. in Computer Science from the Technical University of Madrid (UPM) in 1987. Since 2007 she has been within the Department of Computer Science at the Universidad Nacional de Educación a Distancia (UNED) as Associate Professor. Previously she spends 25 years teaching and researching at UPM. Her current interests are in the areas of Linguistic Engineering and Multimedia Information.

Ruben Granados was born in Madrid (Spain). He received the B.E degree in Computer Science in 2006 and the M.S. degree in Artificial Intelligence Research in 2008 from the Technical University of Madrid (UPM). He is currently pursuing the PhD. degree at UNED. His research interests include multimedia information retrieval and multimedia fusion.

Joan Benavent was born in Valencia, (Spain). He received the M.S. degree in Computer Science from the UNED in 2006, and a M.S. degree in Physics in 2010 from de Universidad de Valencia (UV). Now he is a PhD candidate at the UV.

Esther de Ves was born in Almansa (Spain). She received the M.S. degree in Physics and the Ph.D. in Computer Science from the University of Valencia in 1993 and 1999, respectively. Since 1994 she has been within the Department of Computer Science from the University of Valencia, as an Associate Professor. Her current interests are in the areas of medical image analysis and multimedia databases retrieval.