Visual Concept Features and Textual Expansion in a Multimodal System for Concept Annotation and Retrieval with Flickr Photos at ImageCLEF2012

J. Benavent[2] , A. Castellanos[1], X. Benavent[2], E. De Ves[2], Ana García-Serrano[1]

[1] Universidad Nacional de Educación a Distancia, UNED
[2] Universitat de València

xaro.benavent@uv.es,{acastellanos,agarcia}@lsi.uned.es

Abstract. This paper presents our submitted experiments in the Concept annotation and Concept Retrieval tasks using Flickr photos at ImageCLEF 2012. This edition we applied new strategies for both the textual and the visual subsystems included in our multimodal retrieval system. The visual subsystem has focus on extending the low-level features vector with concept features. These concept features have been calculated by means of a logistic regression model. The textual subsystem has focus on expanding the query information using external resources. Our best concept retrieval run, a multimodal one, is at the ninth position with a MnAP of 0.0295, being the second best group of the contest for the multimodal modality. This is also our best run in the global ordered list (where eleven textual runs are also better than it). We have adapted our multimodal retrieval process for the annotation task obtaining non-very good results for this first participation, with a MiAP of 0.1020.

Keywords: Multimedia Retrieval, Flickr Expansion, Concept Features, Low-level features, Logistic regression relevance feedback.

1 Introduction

The UNED-UV is a research group with researchers from two universities in Spain, the Universidad Nacional de Educación a Distancia (UNED) and the Valencia University (UV). The group is working together since ImageCLEF08 edition. Notice that this is our first participation in the Photo Annotation and Retrieval Task using Flickr photos, being our previous participations at the Wikipedia retrieval [6] and at the Medical [4] tasks.

The visual concept detection, annotation, and retrieval task is a multi-label classification challenge. The participants are asked to annotate the presence of one or more concepts at the annotation subtask using visual and/or textual features, and use this information in the retrieval process [2]. We have participated in the annotation and in the retrieval subtask using visual and textual information.

Our multimedia retrieval system very similar to the ones already used at previous ImageCLEF editions [4,5] is composed of three subsystems (Fig. 1): the Textual Based Information Retrieval (TBIR) system, the Content Based Information Retrieval System (CBIR), and the Fusion subsystem. The main three steps are the following: TBIR subsystem acts first as a pre-filter, and then the CBIR system works over this pre-filtered collection by re-ranking it. The final ranked list is the fusion of the both mono-modal lists. This retrieval process is based on the idea that textual retrieval subsystem better captures the meaning of the query. So it is expected that the textual subsystem eliminates images that are similar from a visual point of view but completely different from a semantic point of view.

At this edition, the TBIR system has been improved by expanding the textual information of the query to improve the retrieval. Most of the participating groups at the previous retrieval task try to take advantage of Flickr tag annotation of the images for the retrieval process. In this regard, Ksibi et al [8] use Flickr tags to extract contextual relationships between them. Izawa et al [7] also use Flickr tags. They combine a TF-IDF model over the tags with a visual word co-occurrence approximation. Another approach investigated for Spyromitros-Xious et al. [11] use the concepts, instead the tags, in order to improve the textual-based retrieval. Unlike the papers presented before, we decided to go beyond in the use of textual information about the images (including tags). For that we have carried out an expansion of the original collection, using the information of the images existing on Flickr.

The CBIR system uses low-level features for image retrieval. This low-level information although gives quite good results depending on the visual information of the query, it is not able to reduce the "semantic gap" in a semantic complex query. Our proposal [3] is to generate Concept features extracted from the low-level features to obtain the probability of the presence of each trained concept. We call this new vector, the *expanded low-level Concept vector* that is calculated for each image of the collection and also for the example images of the query to process the retrieval task. A model for each concept is trained using a logistic regression [9]. We use these regression models as multi-label classifiers at the annotation subtask and as a features vector for the retrieval subtask.

Our proposals both for the textual as for the visual systems are more oriented to a retrieval process than to an annotation subtask. Anyway, we have adapted them for the multi-label annotation subtask. Section 2 describes the visual, textual and multimodal approaches for the concept annotation subtask with Flickr photos. Section 3 explains our multimodal retrieval system use for the concept retrieval subtask. After that section 4 shows the submitted runs and the results obtained for annotation and for retrieval. Finally, in section 5 we extract conclusions and outlines possible future research lines.

2 Concept annotation subtask with Flickr photos.

2.1 Annotation approach using visual information.

For the annotation subtask we train a logistic regression model [9] for each of the concepts defined by the concept annotation subtask [2]. Each trained model predicts the probability that a given image belongs to a certain concept. The concept annotation subtask gives to the participants a training set, I_s, for each of the concepts. Being I_s^P the training image set for each concept, we refer to them as the relevant or positive images. And, being I_s^N the set of no relevant images for a given concept referred as non-relevant or negative images. The logistic regression analysis calculates the probability for a given image to belong to a certain concept. Each image of the training set, I_s is represented by a K-dimensional low-level features vector$\{x_1, .., x_i, .., x_k\}$. The relevance probability for a certain concept c_i for a given image I_j will be represented as $P_{c_i}(I_j)$. A logistic regression model can estimate these probabilities. Let us consider for a binary Y, and k explanatory variables$x = (x_1, \dots, x_k)$, the model for π(x) = P(Y=1| X) (probability $Y = 1$) for the x values $logit\,[\pi(x)] = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k$, where logit ($\pi$(x))=ln($\pi$(x) / (1-$\pi$(x)). The model parameters are obtained by maximizing the likelihood estimator (MLE) of the parameter vector β 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 (k < 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 x and we have to combine them in order to rank the database according to the image probability or image score (Si).

The explanatory variables $x = (x_1, \dots, x_k)$ to train the model are the visual low-level features based on color and texture information that are calculated by our group. We have a low-level features vector of 293 components divided by five different visual information families.

- **Color information:** Color information has been extracted by calculating both local and *overall histograms* of the images. Overall histograms have been calculated using 10x3 bins on the HS color system. Meanwhile, *local histograms* have been calculated by dividing the images into four fragments of the same size. A bi-dimensional HS histogram with 12x4 bins is computed for each patch. Therefore, a feature vector of 222 components represents the color information of the image.
- **Texture information:** This information is embodied as the *granulometric distribution function.* A granulometry is defined from the morphological opening of the texture using a convex and compact subset containing the origin as structuring el-

ement [1]. In our case we have used a horizontal and a vertical segment as the structuring elements, being 60 components in total for both structuring elements. And the *Spatial Size Distribution* that is another morphological operation defined in [1] using a horizontal segment as structuring element, being 10 components.

Once we have the 99 trained models, we calculate for each image the probability of belonging to a given concept, $P_{c_i}(I_j)$. This probability is a floating-point value between 0 and 1 that is the confidence score for the annotation run. For calculating the binary score, if the concept probability is greater than 0.5 ($P_{c_i}(I_j) > 0.5$) is assumed that the concept is present at the image and then it is marked as 1, otherwise is marked as 0 meaning the absence of the concept.

2.2 Annotation approach using visual and textual information.

Based on visual annotation, presented above, we propose a multimodal annotation by an IR-based approach. Our proposal uses a two-step process. In the first step, the visual annotation approach generates a visual-based results list. Then, in the second step, the textual system refines this visual annotated list as follows:

- The textual system only checks the annotated concept as present in an image according to the visual system (set to 1 at the binary annotation).
- The textual system retrieves the concepts, which are most likely in the image, ranked by score, using the textual information of the image as a query against the information associated to the concepts.
- If the textual system identifies the concept as present, the concept is fixed as present and the confidence score is calculated as the product of both textual and visual confidence scores.
- But, if the textual system does not identify the concept as present, the concept is fixed as not present, regardless of the criteria of the visual annotation.

This proposal entailed a problem, there was not enough information associated with both the images and the concepts. Due to this lack of information, it was decided to expand textual information of the collection by external sources. The expansion was posed both for images information and concepts information.

In order to expand the information associated to the images, Flickr was used to provide an adequate textual description for each image. Two different expansion processes are posed:

- **Expansion using Flickr Description of Image:** To all of the images on the collection, we have retrieved the Flickr Description and we have aggregated it to the image description for all the images on the collection.
- **Expansion using Flickr Description of Similar Images:** It was decided to complement user descriptions with the descriptions of other users on similar images. In order to find images that are similar to each image of the collection, we use the tag annotation of the images. For each image, the Flickr API was queried to retrieve

images that share the same tags (all of them or a subset) and aggregate to the image description, the descriptions of the 50 first images retrieved.

We propose three methods for the expansion of the concepts based on two external sources (Flickr and ImageNET[1]):

- **Expansion using user descriptions of the concept on Flickr:** The name of the concept is used as the query for the Flickr API and gets a set of relevant images. Then, the descriptions of these images are aggregated to the concept description.
- **Expansion using user descriptions on Flickr of images annotated with the same concept:** The idea is similar to the expansion presented previously; but instead of querying for images relevant to each concept, we used the images annotated with the given concept. The method is as follows: 1) for each concept the images annotated with it are identified, 2) the descriptions of these images are taken and finally 3) the image description is aggregated to the concept description.
- **Expansion using structured information (ImageNET):** For this approximation, each concept was manually extended by searching them on ImageNET and adding the definition provided by ImageNET to the concept definition.

3 Concept retrieval subtask with Flickr photos.

The system is composed by three subsystems: the Textual Based Image Retrieval (TBIR) Subsystem and the Content Based Image Retrieval (CBIR) Subsystem and the Fusion subsystem (Fig.1.).

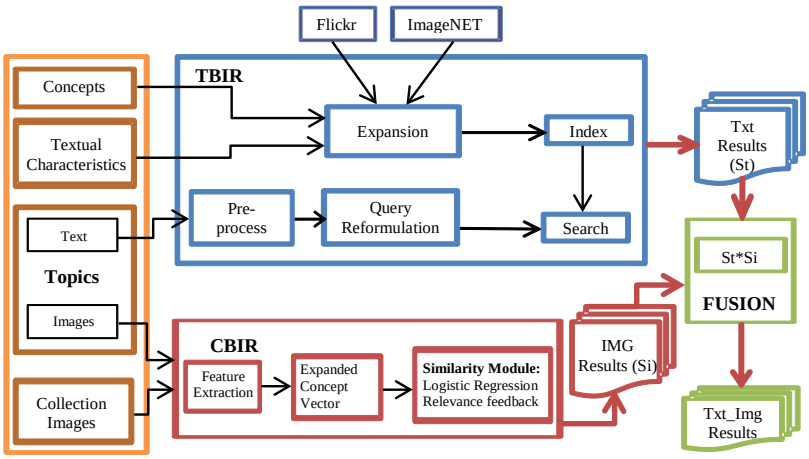

Fig. 1. Retrieval System overview.

[1] http://www.image-net.org/

The TBIR subsystem is responsible of the preprocessing, the expansion, the indexing and, finally, the retrieval process using textual information. The TBIR subsystem recovers only relevant images with a given query and assigns to each image a score (St), based on textual similarity between associated text and query. These relevant images returned by TBIR module are submitted as candidates to the CBIR system as a list, sorted by the score. The TBIR subsystem acts over all images on the collection as a filter. After, the CBIR subsystem assigns another score, Si, to each image based on its work with visual features. In the last step the image list is re-ranked, fusing the scores given by TBIR and CBIR modules by the product of both scores, St*Si.

Each subsystem is described in detail in the two following sections.

3.1 Text Based Information Retrieval Subsystem

This subsystem carries out all the work related with the textual information of the collection (preprocess, query reformulation, collection expansion, indexing, and finally the retrieval). The operation of all these stages is presented below:

- **Preprocess:** Since the indexing and retrieval process are based on terms frequencies, it is important to perform a previous work in favor of normalize and remove noise terms. The preprocessed includes: 1) the special characters, with no statistical meaning, are eliminated; 2) deletion of semantically empty words (i.e. stop-words) in English language, 3) stemming: reduction of word to their base form, using Porter Algorithm and, finally, 4) convert all words to lower case.
- **Query Processed:** The query is processed in two senses. First, meaningless terms or expressions are deleted; more concretely, expressions like *The user is looking for photos showing…* are removed, as it doesn't add any semantic information to the query content. On the other hand, for each query, the concept (or concepts) expected for the results of a given query is identified. This identification is manually done. An example of processing of a query is:
  - Original query:
    The user is looking for photos showing only one or more elderly men, so no other people should be additionally visible
  - Query without meaningless terms:
    one or more elderly men, so no other people should be additionally visible.
  - Concept/s identified:
    Elder Male
- **Collection Expansion:** Textual information associated to the images available in the collection is scarce (both for images and concepts). Due to that our approach requires a significant amount of textual information to work; it became necessary raise an expansion process, using the information available for each image to query external sources. The expansion information, created according the process explained in section 2.2, is aggregated to the collection.
- **Indexing:** For indexing the collection has been used Apache Solr[2]. Solr is an open-source search platform from Apache Lucene[3] project. Through Solr it has been in-

[2] http://lucene.apache.org/solr/

dexed the textual information that the collection had, as well as, the descriptions, generated in the expansion of the collection.

- **Retrieval:** The search process is done by Solr, over Lucene operation. The score function used for calculate the similarity between a given query and the documents is BM25. The results are TREC-format.

3.2 Content Based Information Retrieval Subsystem

The work of the **CBIR subsystem** is based on three main stages: Extraction of the low-level, calculating the Concept features of the images to expand the features vector, and the calculation of the similarity (Si) of each of the images to the image examples given by a query.

1. **Extraction of low-level features:** The first step in the CBIR system is to extract the visual low-level and the Concept features for all the images of the database as well as from the example images given in each question. The low-level features we use are calculated by our group and give color and texture information about the images. These features are the same that we have used for the modality classification task (see section 2.1 for more detailed information).
2. **Calculating the Concept features vector**. The regression models trained for each of the concepts gives for each image on the database and for the example query the probability of the presence of each concept$P_{c_i}(I_j)$. With this probability information for each concept, we extend the low-level features vector to m components, being m the number of concepts trained. Each image I_j on the database is described by the extended vector$F(I_j) = (x_1, \dots x_k, c_1, .., c_m\} \in R^{k+m}$.
3. **Similarity Module:** The similarity module instead of using the classical distance method to calculate the similarity of each of the images of the database to the example images for a given topic uses our own logistic regression relevance algorithm to get the probability of an image belonging to the query set. The sub-models regressions are set to five features inside each features family that are the number of example images given for each topic (see more details of the regression method at section 2.1.). The relevant images are the example images, and the non-relevant images are randomly taken from outside the pre-textual filtered list.

3.3 Fusion subsystem

The **fusion subsystem** is in charge of merging the two score result lists from the TBIR and the CBIR subsystem. In the present work we use the product fusion algorithm (Si*St). The two results lists are fused together to combine the relevance scores of both textual and visually 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.

[3] http://lucene.apache.org/core/

4 Experiments and results

4.1 Concept annotation

In this our first participation on the concept annotation subtask, we have participated with three visual and two multimodal runs (see table 1 for detailed information of the submitted runs).

Our main objective for the visual runs is to test the behavior of our logistic regression model as a classifier for the annotation task, and to adjust the parameters of the regression model. As explained in section 2.1., one of the important parameter is the set of relevant images. The number of images per concept range significantly from 30 to 200 images [2]. We have manually selected the relevant images for runs UNED_UV_02 and UNED_UV_03, and for run UNED_UV_01 all given images are taken up to 100 images. The number of positive plus negative images, k has to be greater than the number of regression parameters to be estimated (see section 2.2). We have fixed for all submitted runs the same number of regression models: a regression for each low-level feature sub-family, being eight regression models varying between 9, 30 or 48 low-level components. It means that we would need between 30 and 50 relevant images. We have fixed the number of relevant images for run UNED_UV_02 and UNED_UV_03 to 30 images.

The other input that the regression model needs is the set of non-relevant images. The number of non-relevant images should be the double of the number of relevant images. The other fact is how to choose the non-relevant images for each concept. At this edition of the Photo annotation Flickr subtask, the concepts have been categorized in family groups [2]. We have used this information to select the number of the non-relevant images. Therefore, at runs UNED_UV_01 and UNED_UV_02 the non-relevant images are selected from a subset of the images outside the family. Meanwhile, at run 3 are selected from subset of images from the same family that not belong to the training concept.

Table 1. Detailed information of the submitted experiments for the concept annotation task.

		Visual information			**Textual information**	
		Relevant images		**Non-relevant images**		
Run	**Modality**	**#number**	**Selection method**	**Selection method**	**Visual baseline**	**Textual algorithm**
UNED_UV_01	Visual	All up to 100	If >100 the nearest to the centroid	Outside the family concept.		
UNED_UV_02	Visual	30	Manually selected	Outside the family concept		
UNED_UV_03	Visual	30	Manually selected	Inside the family concept		
UNED_UV_04	Multimodal				Run2	Textual filter
UNED_UV_05	Multimodal				Run3	Textual filter

The two multimodal runs, UNED_UV_04 and UNED_UV_05 use a different visual run baseline, (run UNED_UV_02 and UNED_UV_03 respectively), and then the textual algorithm described at section 2.2 acts as a filter for the visual run. The two expansion approaches used are the one performed using the *Flickr Description of Similar Images* for the image descriptions and the second expansion using the user descriptions on Flickr of images annotated with the same concept for the concept descriptions. This two expansion approaches are those that provide more information.

Table 2 shows our submitted runs results measured by means of The Interpolated Mean Average Precision (MiAP), Geometric Interpolated Mean Average Precision (GMiAP) and the photo based micro-F1 measure (F-ex). Our best result by MiAP, run UNED_UV_01, is at position 55 from the global result list (80 runs).

In the configurations tested for the visual runs, our results ordered by MiAP from best to worst are run UNED_UV_01, UNED_UV_02 and UNED_UV_03 respectively. It can signify that as more relevant images we have better is the regression model performance. Both runs UNED_UV_01 and UNED_UV_02 outperform run UNED_UV_03 meaning that it is better to select the non-relevant images outside the categorized group.

Concerning these multimodal results, it is clear that the combination of visual and textual annotation proposed does not provide the expected performance. Both multimodal runs (UNED_UV_04 and UNED_UV_05) do not outperform the visual baselines for any of the evaluation measures (MiAP, GMiAP and F-ex).Anyway we think that these not very good results are because the inaccuracy of the information associated with the concepts, that is obtained in the expansion process. Need also to be studied if the filter effect of the textual information over the visual one is too restrictive or not.

It is needed to point out that all the results ordered by the F-ex values are on the opposite way than ordered by MiAP. This fact would have to be analyzed in detail query by query.

Table 2. Results for the submitted concept annotation experiments.

Run	Mode	MiAP	GMiAP	F-ex
UNED_UV_01_CLASS_IMG_NOTADJUST	Visual	0.1020	0.0512	0.1081
UNED_UV_02_CLASS_IMG_RELEVANTSEL_NONREL_OUTSIDE	Visual	0.0932	0.0475	0.1227
UNED_UV_03_CLASS_IMG_RELEVANTSEL_NONREL_INSIDE	Visual	0.0873	0.0441	0.1360
UNED_UV_04_CLASS_Img_base2_TextualFilter	Multimodal	0.0756	0.0376	0.0849
UNED_UV_05_CLASS_Img_base3_TextualFilter	Multimodal	0.0758	0.0383	0.0864

4.2 Concept Retrieval using Flickr photos

We have submitted two textual and eight multimodal runs. Table 3 shows the detailed information for the submitted runs. For the textual baseline, run UNED_UV_01, the content of the topic/query is previously preprocessed and is used to query over the image description in Flickr and also against the description obtained by the expansion using user descriptions on Flickr of images annotated with the same concept (see

section 2.2.). For the UNED_UV_02 run the query process is similar to the previous one, but in addition to use the content of the topic to query over the descriptions, the concept expected for the results of a given query is also used. Given that the concept expected for the queries is not provided, we have identified it in a manually way for each query. For the concepts no expansion information has been used.

Table 3. – Detailed information of the submitted concept retrieval experiments.

		TBIR	CBIR	
Run	Modality	Baseline	Concept model	Features Vector
UNED_UV_01_TXT_EN	Textual			
UNED_UV_02_TXT_EN	Textual			
UNED_UV_03_TXTIMG	Multimodal	UNED_UV_01	Base2	[LF]*[CF]
UNED_UV_04_ TXTIMG	Multimodal	UNED_UV_01	Base2	[LF … CF]
UNED_UV_05_ TXTIMG	Multimodal	UNED_UV_02	Base2	[LF]*[CF]
UNED_UV_06_ TXTIMG	Multimodal	UNED_UV_02	Base2	[LF…CF]
UNED_UV_07_ TXTIMG	Multimodal	UNED_UV_01	Base3	[LF]*[CF]
UNED_UV_08_ TXTIMG	Multimodal	UNED_UV_01	Base3	[LF…CF]
UNED_UV_09_ TXTIMG	Multimodal	UNED_UV_02	Base3	[LF]*[CF]
UNED_UV_10 _TXTIMG	Multimodal	UNED_UV_02	Base3	[LF…CF]

The multimodal runs (runs 3 to 10) have been designed to test the behavior of the expanded features vector. The expanded features vector is obtained as explained at section 2.1. by the regressions models trained at the annotation subtask. We have used two of the four regressions models used at the concept annotation subtask: the experiments two and three from table 2 denoted at table 4 as base2 and base3 respectively. The extended vector $F(I_i) = (x_1, \dots x_k, c_1, \dots, c_m\} \in R^{k+m}$ can be calculated as a unique vector with the low-level and the Concept features (denoted as [LF…CF] at table 4), and as two different vectors (denoted as [LF]*[CF]). For the last scheme, two different probabilities are obtained by the low-level features $S_x(I_i)$, and for the Concept features $S_c(I_i)$, combining both probabilities by the product $S(I_i) = S_x(I_i) * S_c(I_i)$. All multimodal runs use the textual pre-filter algorithm, so the visual system only works over this pre-filtered sub-collection. We have presented four multimodal runs with textual baseline (UNED_UV_01) and the other four with the concept extended textual run (UNED_UV_02). The multimodal runs merged both image and textual scores by the product (St*Si).

The evaluation is done according to the following measures: The overall non-interpolated MAP (MnAP), average of the non-interpolated precisions for each concept and the Average Precision at different values AP@10, AP@20 and AP@100. Table 4 shows our submitted run results.

Our best result, the multimodal run UNED_UV_10 (MnAP of 0.0295) is at the 20th position of the overall result list and at the ninth position for the multimodal

runs, being our group, the UNED_UV, the third group for the multimodal runs, and the fourth best group in the overall results for the concept retrieval results subtask.

Looking at textual runs, our best result is obtained with UNED_UV_02_TXT_AUTO_EN (MnAP of 0.0250), which uses the information about the concept for query expansion. This run improves the results of baseline run without concept information (UNED_UV_01_TXT_AUTO_EN with MnAP of 0.0208).

Our two best multimodal runs, UNED_UV_6 (MnAP of 0.0286) and UNED_UV_10 (MnAP of 0.0295) outperforms its corresponding pre-filtered textual baseline (run UNED_UV_2). This shows that the use of the *expanded concept features vector* as an unique vector or as two different vectors do not make any important difference given that run UNED_UV_6 uses only one vector and UNED_UV_09 run uses two different vectors and both obtain a very similar MnAP values. A similar behavior can be also observed for the annotation base regression model used to get the *expanded concept features vector* so that UNED_UV_6 uses base2 and UNED_UV_09 uses base3, and the MnAP values are similar for both runs. This is also observed for the concept annotation results obtained, in which models 2 and 3 have similar results by MiAP (see Table 2).

Table 4. Results for the submitted concept retrieval experiments.

Run	Mode	MnAP	AP@10	AP@20	AP@100
UNED_UV_01_TXT_AUTO_EN	Textual	0.0208	0.0032	0.0021	0.0653
UNED_UV_02_TXT_AUTO_EN	Textual	0.0250	0.0004	0.0019	0.0250
*Best textual (**IMU**)*		*0.0933*	*0.0187*	*0.0338*	*0.1715*
UNED_UV_03_TXTIMG	Multimodal	0.0271	0.0125	0.0203	0.0813
UNED_UV_04_ TXTIMG	Multimodal	0.0271	0.0131	0.0199	0.0837
UNED_UV_05_ TXTIMG	Multimodal	0.0260	0.0121	0.0224	0.0807
UNED_UV_06_ TXTIMG	Multimodal	0.0286	0.0116	0.0223	0.0819
UNED_UV_07_ TXTIMG	Multimodal	0.0275	0.0112	0.0203	0.0859
UNED_UV_08_ TXTIMG	Multimodal	0.0275	0.0122	0.0198	0.0854
UNED_UV_09_ TXTIMG	Multimodal	0.0270	0.0104	0.0217	0.0822
UNED_UV_10 _TXTIMG	Multimodal	0.0295	0.0125	0.0206	0.0848
*Best Multimodal (**MLKD**)*		*0.0702*	*0.0214*	*0.0342*	*0.1495*

5 Concluding Remarks and Future Work

Our best result is obtained at the concept retrieval subtask in in the multimodal modality. This multimodal run is the UNED_UV_10 (MnAP of 0.0295), at the 20th position of the overall result list and at the ninth position at the multimodal runs list. For the different textual approaches presented, we can conclude that the expansion using information about the concepts outperform the standard retrieval process; even by a simple expansion approach as the presented here. In this regard, we will continue exploring this research line with some remarks. First, a better definition of each con-

cept is desirable in order to improve a better representation of the concept and then a better retrieval process. In this work, we use a simple TF-IDF-based, but more sophisticated approach could be proposed as divergence-based techniques. Second, to address the lack of detailed descriptions of the concepts, we have presented an expansion based on external sources. Although the results shows that these technique improve the baseline results, it has been shown that these expansion introduces a significant amount of noise information. This noise information gets low precision values at the first results.

For the multimodal approaches presented for the concept retrieval subtask, our combination of the textual pre-filtered list as input to the visual system outperform the textual baseline, as it has already been tested in other ImageClef collections, Wikipedia [6] and Medical [4]. Focusing on the visual system, the expanded Concept vector outperforms the use of the low-level features vector in Flickr photo collection and in the Medical collection [5]. Therefore, we will continue working in adjusting the best configuration for the regression models so that no definitive conclusions for the best configuration can be extracted for the present work.

The results obtained at the Concept Annotation subtask have not been as good as the ones obtained at the retrieval subtask. In this first participation our best result is at 56^{th} position from 80 runs. This is due to the fact that both our textual and our visual approaches are retrieval approaches adapted for a classification task. Nevertheless, the regression model system proposed as a multilabel classifier for the annotation concept subtask will deeply be studied. The multimodal approaches do not outperform the visual baseline so that they will also be redefined. We think the textual filter has been too strict, and a relax combination of both confidence scores, textual and visual will get better multimodal annotation results.

Acknowledgments. This work has been partially supported for Regional Government of Madrid under Research Network MA2VIRMR (S2009/TIC-1542), for Spanish Government by projects BUSCAMEDIA (CEN-20091026), HOLOPEDIA (TIN 2010-21128-C02) and MCYT TEC2009-12980.

6 References

1. Ayala, G.; Domingo, J. Spatial Size Distributions. Applications to Shape and Texture Analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2001. Vol. 23, N. 12, pages 1430-1442.
2. Bart Thomee, Adrian Popescu, Overview of the ImageCLEF 2012 Flickr Photo Annotation and Retrieval Task, CLEF 2012 working notes, Rome, Italy, 2012.
3. Benavent, J., Benavent, X., de Ves, E. Recuperación de Información visual utilizando descriptores conceptuales. In Conference Proceedings of the Conferencia Española de Recuperación de Información, CERI 2012, Valencia.
4. Castellanos, A. Benavent, X., Benavent, J. Garcia-Serrano, A.: UNED-UV at Medical Retrieval Task of ImageCLEF 2011. In *Working Notes of CLEF 2011*.

5. Castellanos, A., Benavent, J., Benavent, X., García-Serrano A., de Ves, E.: Using Visual Concept Features in a Multimodal Retrieval System for the Medical collection at ImageCLEF2012, CLEF 2012 working notes, Rome, Italy.
6. Granados, R. Benavent, J. Benavent, X. de Ves, E. Garcia-Serrano, A.: Multimodal Information Approaches for the Wikipedia Collection at ImageCLEF. In *2011 Working Notes*.
7. Izawa, R. Motohashi, N. Takagi, T.: Annotation and Retrieval System Using Confabulation Model for ImageCLEF2011 Photo Annotation. In: *Working Notes of CLEF* 2011.
8. Ksibi, A. Ammar, A.B. Amar, C.B.: REGIMvid at ImageCLEF2011: Integrating Contextual Information to Enhance Photo Annotation and Concept-based Retrieval. In: *Working Notes of CLEF 2011*.
9. Leon T., Zuccarello P., Ayala G., de Ves E., Domingo J.: Applying logistic regression to relevance feedback in image retrieval systems, *Pattern Recognition, V40*, p.p. 2621, 2007.
10. Nowak, S., Nagel, K., Liebetrau, J.: The CLEF 2011 photo annotation and concept-based retrieval tasks. In: *Working Notes of CLEF 2011*.
11. Spyromitros-Xious, E. Sechidis, K. Tsoumakas, G. Vlahavas, I.: MLKD's Participation at the CLEF 2011 Photo Annotation and Concept-Based Retrieval Tasks. In: *Working Notes of CLEF 2011*.