MLKD's Participation at the CLEF 2011 Photo Annotation and Concept-Based Retrieval Tasks

Eleftherios Spyromitros-Xioufis, Konstantinos Sechidis,
Grigorios Tsoumakas, and Ioannis Vlahavas

Dept of Informatics
Aristotle University of Thessaloniki
Thessaloniki 54124, Greece
{espyromi,sechidis,greg,vlahavas}@csd.auth.gr

Abstract. We participated both in the photo annotation and concept-based retrieval tasks of CLEF 2011. For the annotation task we developed visual, textual and multi-modal approaches using multi-label learning algorithms from the Mulan open source library. For the visual model we employed the ColorDescriptor software to extract visual features from the images using 7 descriptors and 2 detectors. For each combination of descriptor and detector a multi-label model is built using the Binary Relevance approach coupled with Random Forests as the base classifier. For the textual models we used the boolean bag-of-words representation, and applied stemming, stop words removal, and feature selection using the chi-squared-max method. The multi-label learning algorithm that yielded the best results in this case was Ensemble of Classifier Chains using Random Forests as base classifier. Our multi-modal approach was based on a hierarchical late-fusion scheme. For the concept based retrieval task we developed two different approaches. The first one is based on the concept relevance scores produced by the system we developed for the annotation task. It is a manual approach, because for each topic we manually selected the relevant topics and manually set the strength of their contribution to the final ranking produced by a general formula that combines topic relevance scores. The second approach is based solely on the sample images provided for each query and is therefore fully automated. In this approach only the textual information was used in a query-by-example framework.

1 Introduction

ImageCLEF is the cross-language image retrieval track run annually since 2003 as part of the Cross Language Evaluation Forum (CLEF)[1]. This paper documents the participation of the Machine Learning and Knowledge Discovery (MLKD) group of the Department of Informatics of the Aristotle University of Thessaloniki at the photo annotation task (also called visual concept detection and annotation task) of ImageCLEF 2011.

[1] http://www.clef-campaign.org/

This year, the photo annotation task consisted of two subtasks. An annotation task, similar to that of ImageCLEF 2010, and a new concept-based retrieval task. Data for both tasks come from the MIRFLICKR-1M image dataset [1], which apart from the image files contains Flickr user tags and Exchangeable Image File Format (Exif) information. More information about the exact setup of the data can be found in [4].

In the annotation task, participants are asked to annotate a test set of 10,000 images with 99 visual concepts. An annotated training set of 8,000 images is provided. This multi-label learning task [8] can be solved in three different ways according to the type of information used for learning: 1) visual (the image files), 2) textual (Flickr user tags), 3) multi-modal (visual and textual information). We developed visual, textual and multi-modal approaches for this task using multi-label learning algorithms from the Mulan open source library [9]. In this task, the relative performance of our textual models was quite good, but that of our visual models was bad (our group does not have expertise on computer vision), leading to an average multi-modal (and overall) performance.

In the concept-based retrieval task, participants were given 40 topics consisting of logical connections between the 99 concepts of the photo annotation task, such as "*find all images that depict a small group of persons in a landscape scenery showing trees and a river on a sunny day*", along with 2 to 5 examples images of each topic from the training set of the annotation task. Participants were asked to submit (up to) the 1,000 most relevant photos for each topic in ranked order from a set of 200,000 unannotated images. This task can be solved by manual construction of the query out of the narrative of the topics, followed by automatic retrieval of images, or by a fully automated process. We developed a manual approach that exploits the multi-label models trained in the annotation task and a fully automated query-by-example approach based on the tags of the example images. In this task, both our manual and automated approaches ranked 1st in all evaluation measures by a large margin.

The rest of this paper is organized as follows. Sections 2 and 3 describe our approaches to the annotation task and concept-based retrieval task respectively. Section 4 presents the results of our runs for both tasks. Section 5 concludes our work and poses future research directions.

2 Annotation Task

This section presents the visual, textual and multi-modal approaches that we developed for the automatic photo annotation task. There were two (eventually three) evaluation measures to consider for this task: a) mean interpolated average precision (MIAP), b) example-based F-measure (F-ex), c) semantic R-precision (SR-Precision). In order to optimize a learning approach based on each of the initial two evaluation measures and type of information, six models should be built. However, there were only five runs allowed for this task. We therefore decided to perform model selection based on the widely-used mean average precision (MAP) measure for all types of information. In particular, MAP was estimated

using an internal 3-fold cross-validation on the 8,000 training images. Our multi-modal approach was submitted in three different variations to reach the total number of five submissions.

2.1 Automatic Annotation with Visual Information

We here describe the approach that we followed in order to learn multi-label models using the visual information of the images. The flowchart of this approach is shown in Fig. 1.

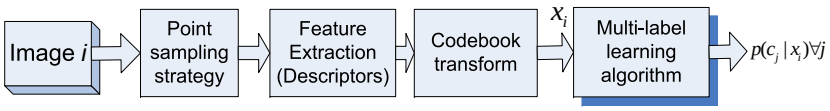

Fig. 1. Automatic annotation using visual information.

As our group does not have expertise in computer vision, we largely followed the color-descriptor extraction approach described in [6,7] and used the accompanying software tool[2] for extracting visual features from the images.

Harris-Laplace and Dense Sampling were used as point detection strategies. Furthermore seven different descriptors were used: SIFT, HSV-SIFT, HueSIFT, OpponentSIFT, C-SIFT, rgSIFT and RGB-SIFT. For each one of the 14 combinations of point detection strategy and descriptor, a different codebook was created in order to obtain a fixed length representation for all images. This is also known as the bag-of-words approach. The k-means clustering algorithm was applied to 250,000 randomly sampled points from the training set, with the codebook size (k) fixed to 4096 words. Finally, we employed hard assignment of points to clusters.

Using these 4,096-dimensional vector representations along with the ground truth annotations given for the training images we built 14 multi-label training datasets. After experimenting with various multi-label learning algorithms we found that the simple *Binary Relevance* (BR) approach coupled with *Random Forests* as the base classifier (number of trees = 150, number of features = 40) yielded the best results.

In order to deal with the imbalance in the number of positive and negative examples of each label we used instance weighting. The weight of the examples of the minority class was set to $(min+maj)/min$ and the weight of the examples of the majority class was set to $(min+maj)/maj$, where min is the number of examples of the minority class and maj the number of examples of the majority class. We also experimented with sub-sampling, but the results were worse than instance weighting.

[2] Available from http://www.colordescriptors.com

Our approach concludes with a late fusion scheme that averages the output of the 14 different multi-label models that we built.

2.2 Automatic Annotation with Flickr User Tags

We here describe the approach that we followed in order to learn multi-label models using the tags assigned to images by Flickr users. The flowchart of this approach is shown in Fig. 2.

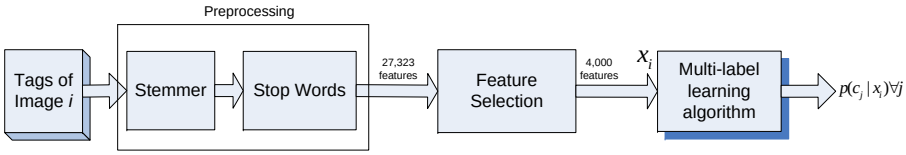

Fig. 2. Automatic annotation using Flickr user tags.

An initial vocabulary was constructed by taking the union of the tag sets of all images in the training set. We then applied stemming to this vocabulary and removed stop words. This led to a vocabulary of approximately 27000 stems. The use of stemming improved the results, despite that some of the tags were not in the English language and that we used an English stemmer. We further applied feature selection in order to remove irrelevant or redundant features and improve efficiency. In particular, we used the χ^2_{max} criterion [3] to score the stems and selected the top 4000 stems, after experimenting with a variety of sizes (500, 1000, 2000, 3000, 4000, 5000, 6000 and 7000).

The multi-label learning algorithm that was found to yield the best results in this case was *Ensemble of Classifier Chains* (ECC) [5] using Random Forests as base classifier. ECC was run with 15 classifier chains and Random Forests with 10 decision trees, while all other parameters were left to their default value. The approach that we followed to deal with class imbalance in the case of visual information (see the previous subsection), was followed in this case too.

2.3 Automatic Annotation with a Multi-Modal Approach

Our multi-modal approach is based on a late fusion scheme that combines the output of the 14 visual models and the single textual model. The combination is not an average of these 15 models, because in that case the visual models would dominate the final scores. Instead, we follow a hierarchical combination scheme. We separately average the 7 visual models of each point estimator and then combine the output of the textual model, the Harris-Laplace average and the Dense Sampling average, as depicted in Fig. 3. The motivation for this scheme was the three different views of the images that existed in the data (Harris-Laplace, Dense Sampling, user tags) as explained in the following two paragraphs.

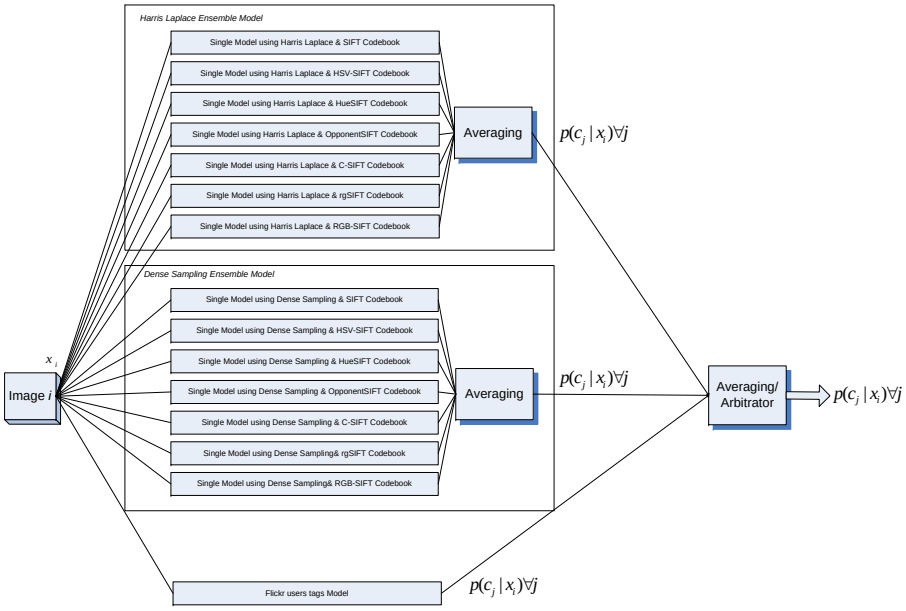

Fig. 3. Automatic annotation with a multi-modal approach

We can discern two main categories of concepts in photo annotation: *objects* and *scenes*. For objects, Harris-Laplace performs better because it ignores the homogeneous areas, while for scenes, Dense Sampling performs better [6]. For example, two of the concepts where Dense Sampling achieves much higher *Average Precision* (AP) from Harris-Laplace are *Night* and *Macro*, which are abstract, while the inverse is happening in concepts *Fish* and *Ship*, which correspond to things (organisms, objects) of particular shape.

Furthermore, we observe that the visual approach performs better in concepts, such as *Sky*, which for some reason (e.g. lack of user interest for retrieval by this concept) do not get tagged. On the other hand the textual approach performs much better when it has to predict concepts, such as *Horse*, *Insect*, *Dog* and *Baby* that typically get tagged by users. Table 1 shows the average precision for 10 concepts, half of which suit much better the textual models and half the visual models.

Two variations of this scheme were developed, differing in how the output of the three different views is combined. The first one, named *Multi-Modal-Avg*, used an averaging operator, similarly to the one used at the lower levels of the hierarchy. The second one, named *Multi-Modal-MaxAP*, used an arbitrator function to select the best one out of the three outputs for each concept, according to internal evaluation results in terms of average precision. Our third multi-modal submission, named *Multi-Modal-MaxAP-RGBSIFT*, was a preliminary version of *Multi-Modal-MaxAP*, where only the RGBSIFT descriptor was used.

Table 1. Average precision for 10 concepts, half of which suit much better the textual models and half the visual models.

Concept	Textual	Visual	Concept	Textual	Visual
Airplane	**0.6942**	0.0946	Trees	0.3004	**0.5501**
Horse	**0.5477**	0.0541	Clouds	0.4744	**0.6949**
Bird	**0.5260**	0.1275	Sky	0.6021	**0.7945**
Insect	**0.5087**	0.1241	Overexposed	0.0183	**0.1937**
Dog	**0.6190**	0.2406	Big_Group	0.1510	**0.3245**

2.4 Thresholding

The multi-label learners used in this work provide us with a confidence score for each concept. This is fine for an evaluation with MIAP and SR-Precision, but does not suffice for an evaluation with example-based F-measure, which requires a bipartition of the concepts into relevant and irrelevant ones. This is a typical issue in multi-label learning, which is dealt with a *thresholding* process [2].

We used the thresholding method described in [5], which applies a common threshold across all concepts and provides a close approximation of the label cardinality (LC) of the training set to the predictions made on the test set. The threshold is calculated using the following formula:

$$t = argmin_{\{t \in 0.00, 0.05, \ldots, 1.00\}} |LC(D_{train}) - LC(H_t(D_{test}))| \quad (1)$$

where D_{train} is the training set and H_t is a classifier which has made predictions on a test set D_{test} under threshold t.

3 Concept-Based Retrieval Task

We developed two different approaches for the concept-based retrieval task. The first one is based on the concept relevance scores produced by the system we developed for the annotation task. It is a *manual* approach, because for each topic we manually selected the relevant topics and manually set the strength of their contribution to the final ranking produced by a general formula that combines topic relevance scores. The second one is based solely on the sample images provided for each query and is therefore fully *automated*.

3.1 Manual Approach

Let $I = 1, \ldots, 200,000$ be the collection of images, $Q = 1, \ldots, 40$ the set of topics and $C = 1, \ldots, 99$ the set of concepts. We first apply our automated image annotation system to each image $i \in I$ and obtain a corresponding 99-dimensional vector $S_i = [s_i^1, s_i^2, ..., s_i^{99}]$ with the relevance scores of this image to each one of the 99 concepts. For efficiency reasons, we used simplified versions of

our visual approach, taking into account only models produced with the RGB-SIFT descriptor, which has been found in the past to provide better results compared to other single color descriptors [7].

Then, based on the description of each of the 40 queries, we manually select a number of concepts that we consider related to the query, either positively or negatively. Formally, for topic $q \in Q$ let $P_q \subseteq C$ denote the set of concepts that are positively related to q and $N_q \subseteq C$ the set of concepts that are negatively related to q, $P_q \cap N_q = \emptyset$. For each concept c in $P_q \cup N_q$, we further define a real valued parameter $m_q^c \geq 1$ denoting the strength of relevance of concept c to q. The larger this value, the stronger the influence of concept c to the final relevance score. For each topic q and image i, the scores of the relevant concepts are combined using (2).

$$S_{q,i} = \prod_{c \in P_q} (s_i^c)^{m_q^c} \prod_{c \in N_q} (1 - s_i^c)^{m_q^c} \tag{2}$$

Finally, for each topic, we arrange the images in descending order according to the overall relevance score and we retrieve a fixed number of images (in our submissions we retrieved 250 and 1,000 images).

Note that for each topic, the selection of related concepts and the setting of values for the m_q^c parameters was done using a trial-and-error approach involving careful visual examination of the top 10 retrieved images, as well as more relaxed visual examination of the top 100 retrieved images. Two examples of topics and corresponding combination of scores follow.

Topic 5: rider on horse. *Here we like to find photos of riders on a horse. So no sculptures or paintings are relevant. The rider and horse can be also only in parts on the photo. It is important that the person is riding a horse and not standing next to it.* Based on the description of this topic and experimentation, we concluded that concepts 75 (Horse) and 8 (Sports) are positively related (rider on horse), while concept 63 (Visual_Arts) is negatively related (no sculptures or paintings). We therefore set $P_5 = \{8, 75\}$, $N_5 = \{63\}$. All concepts were set to equal strength for this topic: $m_{8,5} = m_{63,5} = m_{75,5} = 1$.

Topic 24: funny baby. *We like to find photos of babies looking funny. The baby should be in the main focus of the photo and be the reason why the photo looks funny. Photos presenting funny things that are not related to the baby are not relevant.* Based on the description of this topic and experimentation, we concluded that concepts 86 (Baby), 92 (Funny) and 32 (Portrait) are positively related. We therefore set $P_{24} = \{32, 86, 92\}$, $N_{24} = \emptyset$. Based on experimentation the concept Funny was given twice the strength of the other concepts, we set $m_{32,24} = m_{86,24} = 1$ and $m_{92,24} = 2$.

For some topics, instead of explicitly using the score of a group of interrelated concepts we considered introducing a virtual concept with score equal to the maximum of this group of concepts. This slight adaptation of the general rule of (2), enhances its representation capabilities. The following example clarifies this adaptation.

Topic 32: underexposed photos of animals. *We like to find photos of animals that are underexposed. Photos with normal illumination are not relevant. The animal(s) should be more or less in the main focus of the image.* Based on the description of this topic and experimentation, we concluded that concepts 44 (Animals), 34 (Underexposed), 72 (Dog), 73 (Cat), 74 (Bird), 75 (Horse), 76 (Fish) and 77 (Insect) are positively related, while concept 35 (Neutral_Illumination) is negatively related. The six last specific animal concepts were grouped into a virtual concept, say concept 1001, with score, the maximum of the scores of these six concepts. We then set $P_{32} = \{34, 44, 1001\}$, $N_{32} = \{35\}$ and $m_{34,32} = m_{44,32} = m_{1001,32} = m_{35,32} = 1$.

Figure 4 shows the top 10 retrieved images for topics 5, 24 and 32, along with the Precision@10 for these topics.

3.2 Automated Approach

Apart from the narrative description, each topic of the concept-based retrieval task was accompanied by a set of 2 to 5 images from the training set which could be considered relevant for the topic. Using these examples images as queries we developed a *Query by Example* approach to find the most relevant images in the retrieval set. The representation followed the bag-of-words model and was based on the Flickr user tags assigned to each image.

To generate the feature vectors, we applied the same method as the one used for the annotation task. Thus, each image was represented as a 4000-dimensional feature vector where each feature corresponds to a tag from the training set which was selected by the feature selection method. A value of 1/0 denotes the presence/absence of the tag in the tags accompanying an image.

To measure the similarity between the vectors representing two images we used the *Jaccard* similarity coefficient which is defined as the total number of attributes where two vectors A and B both have a value of 1 divided by the the total number of attributes where either A or B have a value of 1.

Since more than one images where given as examples for each topic, we added their feature vectors in order to form a single query vector. This approach was found to work well in comparison to other approaches, such as taking only one of the example images as query or measuring the similarity between a retrieval image and each example image separately and then returning the images from the retrieval set with the largest similarity score to any of the queries. We attribute this to the fact that by adding the feature vectors, a better representation of the topic of interest was created which could not be possible if only one image (with possibly noisy or very few tags) was considered.

As in the manual approach, we submitted two runs, one returning the 250 and one the 1000 most similar images from the retrieval set (in descending similarity order).

Figure 5 shows the top 10 retrieved images, along with the Precision@10 for the following topics:

– **Topic 10: single person playing a musical instrument.** *We like to find pictures (no paintings) of a person playing a musical instrument. The person*

can be on stage, off stage, inside or outside, sitting or standing, but should be alone on the photo. It is enough if not the whole person or instrument is shown as long as the person and the instrument are clearly recognizable.

- **Topic 12: snowy winter landscaper.** *We like to find pictures (photos or drawings) of white winter landscapes with trees. The landscape should not contain human-made objects e.g. houses, cars and persons. Only snow on the top of a mountain is not relevant, the landscape has to be fully covered in (at least light) snow.*
- **Topic 30: cute toys arranged to a still-life.** *We like to find photos of toys arranged to a still-life. These toys should look cute in the arrangement. Simple photos of a collection of toys e.g. in a shop are not relevant.*

We see that the 10 retrieved images for topic 30 are better than those of topics 12 and 10. This can be explained by noticing that topic 12 is a difficult one, while the tags of the example images for topic 10 are not very descriptive/informative.

4 Results

We here briefly present our results, as well as our relative performance compared to other groups and submissions. Results for all groups, as well as more details on the data setup and evaluation measures can be found in [4].

4.1 Annotation Task

The official results of our runs are illustrated in Table 2. We notice that in terms of MIAP, the textual model is slightly better than the visual, while for the other two measures, the visual model is much better than the textual. Among the multi-modal variations, we notice that averaging works better than arbitrating, and as expected using all descriptors is better than using just the RGB-SIFT one. In addition, we notice that the multi-modal approach significantly improves over the MIAP of the visual and textual approaches, while it slightly decreases/increases the performance of the visual model in the two example-based measures. This may partly be due to the fact that we performed model selection based on MAP.

Table 2. Official results of the MLKD team in the annotation task.

Run Name	MIAP	F-measure	SR-Precision
Textual	0.3256	0.5061	0.6527
Visual	0.3114	**0.5595**	0.6981
Multi-Modal-Avg	**0.4016**	0.5588	**0.6982**
Multi-Modal-MaxAP-RGBSIFT	0.3489	0.5094	0.6687
Multi-Modal-MaxAP	0.3589	0.5165	0.6709

Table 3 shows the rank of our best result compared to the best results of other groups and compared to all submissions. We did quite good in terms of textual information, but quite bad in terms of visual information, leading to an overall average performance. Lack of computer vision expertise in our group may be a reason for not being able to get results out of the visual information. Among the three evaluation measures, we notice that overall we did better in terms of MIAP, slightly worse in terms of F-measure, and even worse in terms of SR-Precision. The fact that model selection was performed based on MAP definitely played a role for this result.

Table 3. Rank of our best result compared to the best results of other teams and compared to all submissions in the annotation task.

Approach	Team Rank			Submission Rank		
	MIAP	F-Measure	SR-Prec	MIAP	F-Measure	SR-Prec
Visual	9th/15	5th/15	9th/15	25th/46	12th/46	17th/46
Textual	3rd/7	2nd/7	3rd/7	3rd/8	2nd/8	4th/8
Multi-modal	5th/10	5th/10	7th/10	9th/25	7th/25	15th/25
All	5th/18	7th/18	10th/18	9th/79	19th/79	31st/79

4.2 Concept-Based Retrieval Task

In this task, participating systems were evaluated using the following measures: Mean Average Precision (MAP), Precision@10, Precision@20, Precision@100 and R-Precision.

The official results of our runs are illustrated in Table 4. We first notice that the first 5 runs, which retrieved 1000 images, lead to better results in terms of MAP and R-Precision compared to the last 5 runs, which retrieved 250 images. Obviously, in terms of Precision@10, Precision@20 and Precision@100, the results are equal. Among the manual runs, we notice that the visual models perform quite bad. We hypothesize that a lot of concepts that favor textual rather than visual models, as discussed in Sect. 2, appear in most of the topics. The textual and multi-modal models perform best, with the Multi-Modal-Avg model having the best result in 3 out of the 5 measures.

The automated approach performs slightly better than the visual model of the manual approach, but still much worse than the textual and multi-modal manual approaches. As expected, the knowledge that is provided by a human can clearly lead to better results compared to a fully automated process. However, this is not true across all topics, as can be seen in Table 5, which compares the results of the best automated and manual approach for each individual topic. We can see there that the automated approach performs better on 9 topics, while the manual on 31.

Table 4. Official results of the MLKD team in the concept-based retrieval task.

Run Name	MAP	P@10	P@20	P@100	R-Prec
Manual-Visual-RGBSIFT-1000	0.0361	0.1525	0.1375	0.1080	0.0883
Automated-Textual-1000	0.0849	0.3000	0.2800	0.2188	0.1530
Manual-Textual-1000	0.1546	0.4100	**0.3838**	0.3102	0.2366
Manual-Multi-Modal-Avg-RGBSIFT-1000	**0.1640**	0.3900	0.3700	**0.3180**	**0.2467**
Manual-Multi-Modal-MaxAP-RGBSIFT-1000	0.1533	**0.4175**	0.3725	0.2980	0.2332
Manual-Visual-RGBSIFT-250	0.0295	0.1525	0.1375	0.1080	0.0863
Automated-Textual-250	0.0708	0.3000	0.2800	0.2188	0.1486
Manual-Textual-250	0.1328	0.4100	**0.3838**	0.3102	0.2298
Manual-Multi-Modal-Avg-RGBSIFT-250	0.1346	0.3900	0.3700	**0.3180**	0.2397
Manual-Multi-Modal-MaxAP-RGBSIFT-250	0.1312	**0.4175**	0.3725	0.2980	0.2260

Table 5. Comparison of AP for each topic between automated and manual approach

Topic	Automated	Manual	Topic	Automated	Manual
1	**0.235**	0.2201	21	**0.0799**	0.0312
2	0.0294	**0.1518**	22	0	**0.1018**
3	**0.0893**	0.0613	23	0.0405	**0.0617**
4	0.257	**0.3701**	24	0.0231	**0.1226**
5	0.0011	**0.5478**	25	0.009	**0.1691**
6	0.12	**0.3574**	26	0.0027	**0.0056**
7	0.0142	**0.2164**	27	0.0477	**0.1311**
8	0.0864	**0.0879**	28	0.0123	**0.1315**
9	0.0001	**0.1143**	29	0.0232	**0.118**
10	0.1618	**0.2528**	30	**0.135**	0.0378
11	0.1393	**0.3133**	31	0.0794	**0.1535**
12	0.0519	**0.0734**	32	0.0221	**0.1135**
13	0.0275	**0.1516**	33	0.0343	**0.434**
14	0.0087	**0.0968**	34	**0.4464**	0.4341
15	0.0455	**0.3327**	35	0.3065	**0.3685**
16	0.0711	**0.0715**	36	0.0001	**0.1426**
17	0.0349	**0.0401**	37	**0.2232**	0.0207
18	0.0011	**0.0044**	38	**0.2431**	0.1477
19	0.0379	**0.0691**	39	**0.0226**	0.0153
20	**0.1837**	0.1168	40	0.0508	**0.1703**
			MAP	0.0849	**0.1640**

Table 6 shows the rank of our best result compared to the best results of other groups and compared to all submissions. Both our manual and our automated approach ranked 1st in all evaluation measures.

Table 6. Rank of our best result compared to the best results of other teams and compared to all submissions in the annotation task.

Configurations	Team Rank					Submission Rank				
	MAP	P@10	P@20	P@100	R-Prec	MAP	P@10	P@20	P@100	R-Prec
Automated	1st/4	1st/4	1st/4	1st/4	1st/4	1st/16	1st/16	1st/16	1st/16	1st/16
Manual	1st/3	1st/3	1st/3	1st/3	1st/3	1st/15	1st/15	1st/15	1st/15	1st/15
All	1st/4	1st/4	1st/4	1st/4	1st/4	1st/31	1st/31	1st/31	1st/31	1st/31

5 Conclusions and Future Work

Our participation to the very interesting photo annotation and concept-based retrieval tasks of CLEF 2011, led to a couple of interesting conclusions. First of all, we found out that we need the collaboration of a computer vision/image processing group to achieve better results. In terms of multi-label learning algorithms, we noticed that binary approaches worked quite well, especially when coupled with the strong Random Forests algorithm and class imbalance issues are taken into account. We also reached to the conclusion, that we should have performed model selection separately for each evaluation measure. We therefore suggest that in future versions of the annotation task, the allowed number of submissions should be equal to the number of evaluation measures multiplied by the number of information types, so that there is space in the official results for models with all kinds of information.

There is a lot of room for improvements in the future, both in the annotation and the very interesting concept-based retrieval task. In terms of textual information, we intend to investigate the translation of non-English tags. We would also like to investigate other hierarchical late fusion schemes, such as an additional averaging step for the two different visual modalities (Harris-Laplace, Dense Sampling) and more advanced arbitration techniques. Other thresholding approaches for obtaining bipartitions is another interesting direction for future study.

Acknowledgments

We would like to acknowledge the student travel support from EU FP7 under grant agreement no 216444 (PetaMedia Network of Excellence).

References

1. Huiskes, M.J., Thomee, B., Lew, M.S.: New trends and ideas in visual concept detection: The mir flickr retrieval evaluation initiative. In: MIR '10: Proceedings of the 2010 ACM International Conference on Multimedia Information Retrieval. pp. 527–536. ACM, New York, NY, USA (2010)

2. Ioannou, M., Sakkas, G., Tsoumakas, G., Vlahavas, I.: Obtaining bipartitions from score vectors for multi-label classification. Tools with Artificial Intelligence, IEEE International Conference on 1, 409–416 (2010)
3. Lewis, D.D., Yang, Y., Rose, T.G., Li, F.: Rcv1: A new benchmark collection for text categorization research. J. Mach. Learn. Res. 5, 361–397 (2004)
4. Nowak, S., Nagel, K., Liebetrau, J.: The clef 2011 photo annotation and concept-based retrieval tasks. In: Working Notes of CLEF 2011 (2011)
5. Read, J., Pfahringer, B., Holmes, G., Frank, E.: Classifier chains for multi-label classification. In: Proc. 20th European Conference on Machine Learning (ECML 2009). pp. 254–269 (2009)
6. van de Sande, K.E.A., Gevers, T.: University of Amsterdam at the Visual Concept Detection and Annotation Tasks, The Information Retrieval Series, vol. 32: ImageCLEF, chap. 18, pp. 343–358. Springer (2010)
7. van de Sande, K.E.A., Gevers, T., Snoek, C.G.M.: Evaluating color descriptors for object and scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 32(9), 1582–1596 (2010)
8. Tsoumakas, G., Katakis, I., Vlahavas, I.: Mining multi-label data. In: Maimon, O., Rokach, L. (eds.) Data Mining and Knowledge Discovery Handbook, chap. 34, pp. 667–685. Springer, 2nd edn. (2010)
9. Tsoumakas, G., Spyromitros-Xioufis, E., Vilcek, J., Vlahavas, I.: Mulan: A java library for multi-label learning. Journal of Machine Learning Research (JMLR) 12, 2411–2414 (July 12 2011)

Fig. 4. Retrieved images for topics 5, 24 and 32 using manual retrieval. Images come from the MIRFLICKR-1M image dataset [1].

Fig. 5. Retrieved images for topics 10, 12 and 30 using automated retrieval. Images come from the MIRFLICKR-1M image dataset [1].