CERTH's participation at the photo annotation task of ImageCLEF 2012

Eleni Mantziou, Georgios Petkos, Symeon Papadopoulos, Christos Sagonas, and Yiannis Kompatsiaris

Information Technologies Institute
Centre for Research and Technology Hellas, Thessaloniki, Greece
{lmantziou, gpetkos, papadop, sagonas, ikom}@iti.gr

Abstract. This paper describes the approaches and experimental settings of the five runs submitted by CERTH at the photo annotation task of ImageCLEF 2012. Two different approaches were used, the first using the Laplacian Eigenmaps of an image similarity graph for learning, and the second using a "same class" learning model. Four runs were submitted using the first, and one using the second approach. A multitude of textual and visual features were employed, making use of different aggregation (BoW, VLAD) and post-processing schemes (WordNet, pLSA). The best performance scores in the test set was achieved by Run 3 (first approach using all features), which amounted to 0.321 in terms of MiAP and 0.2547 in terms of GMiAP (7th out of 18 compteting teams), and Run 5 which led to an F-ex score of 0.495 (6th out of 18 teams).

1 Introduction

This document describes the participation of CERTH at the photo annotation task of the 2012 ImageCLEF competition [1]. CERTH submitted five runs using two different approaches. The first approach, to be described in subsection 2.1, computes the similarity between test images and train images, constructs an image similarity graph, and trains concept detectors by using the graph Laplacian Eigenmaps (LE) [7] as features. This is done for each modality and the final result is obtained by performing late fusion using a linear classifier. The second approach, to be detailed in subsection 2.2, utilizes the concept of a "same class" model that takes as input the set of distances (as many as the number of used features) between the image to be annotated and a reference item that represents a target concept, and predicts whether the image belongs to the target concept. Section 3 outlines each of the submitted runs and presents the obtained test results. Section 4 presents some general remarks and conclusions.

2 Overview of methods

2.1 Concept detection using image similarity graphs

The first approach used by CERTH is based on the construction of a similarity graph between the images. This graph is used to obtain a low-dimensional feature

representation: we use the first eigenvectors of the graph Laplacian as features. These features correspond well to semantically coherent groups of images, and are thus used to train concept classifiers.

The idea of utilizing the implicit relational structure that can be derived by computing similarities between the images of a collection has been proposed before. In [8], an extended similarity measure is proposed that takes into account the local neighbourhood structure of images, i.e, the content and label information (if available) of images that are similar to the input image. The aforementioned measure is used in combination with two well-known semi-supervised learning methods [16] and is shown to improve their performance both in synthetic experiments and in benchmark video annotation task. Our work is mostly related to [6] that introduces the concept of "social dimensions", i.e. the top-k eigenvectors of a graph Laplacian, as an alternative to tackling the relational classification problem, [10], i.e. the classification of a graph node by taking into account information from neighbouring nodes. Here, we adopt a similar representation for graph structure features.

Method overview: Given a set of K target concepts $\mathcal{Y} = \{Y_1....Y_K\}$ and an annotated set $\mathcal{L} = \{(\mathbf{x_i}, \mathbf{y_i})\}_{i=1}^{l}$ of training samples, where $\mathbf{x_i} \in \Re^D$ stands for the feature vector extracted from content item i and $\mathbf{y_i} \in \{0,1\}^K$ for the corresponding concept indicator vector, a transductive learning algorithm attempts to predict concepts associated with a set of unknown items $\mathcal{U} = \{x_j\}_{j=l+1}^{l+u}$, by processing together sets $\mathcal{L}$ and $\mathcal{U}$. Based on the features of the input items, a graph G = (V,E) is constructed that represents the similarities between all pairs of items. The nodes of the graph include the items of both sets of media items ($\mathcal{L}$ and $\mathcal{U}$), i.e. $V = V_L \cup V_U$ with $|V| = n$. There are different options for constructing such a graph. We adopt the kNN graph in which an edge is inserted between items i and j as long as one of them belongs to the set of top-k most similar items of the other. Similarity can be computed by means of different schemes, e.g. inner product or heat kernel (Equation 1).

$$w_{ij} = \exp\left(-\frac{|x_i - x_j|^2}{t}\right) \tag{1}$$

The basic variants of this scheme are *symmetric* and *asymmetric*, depending on whether both items to be linked need to belong to the set of top-k most similar items of each other or not. Having constructed the similarity graph between the input items, our approach proceeds with mapping the graph nodes to feature vectors that represent the associations of nodes with latent groups of nodes forming densely connected clusters. To extract such features, we first construct the normalized graph Laplacian:

$$\tilde{L} = D^{-1/2} L D^{-1/2} = I - D^{-1/2} A D^{-1/2} \tag{2}$$

where D and A are the degree and adjacency matrix of the graph respectively, and L = D-A is the graph Laplacian. Computing the eigenvectors of $\tilde{L}$ corresponding to the C_D smallest non-zero eigenvalues of the matrix results in a set

of n vectors with C_D dimensions, which are then stacked to form the input matrix $S \in \Re^{nxC_D}$, each row of which is denoted as $S_i \in \Re^{C_D}$ and constitutes the graph structure feature vector for media item i. These features are also known as Laplacian Eigenmaps (LE) [7].

Training and performance tuning: We approximately optimise the values from the top-k most similar items, the LEs and the c parameter from the SVM linear classifier in order to construct sets of parameters per concept for each initial feature. In practice, we choose six different top-k [100, 200, 500, 1000, 1500, 2000] nearest neighbours values and for each one we compute LE vectors for six different values [10, 50, 100, 200, 400, 500] for C_D using spectral clustering. For the parameter c we investigate the performance of the SVM classifier by doing cross validation and to decide which of the five different values [0.1, 1, 5, 50] yields the best performance. In most cases the best classification was achieved for $c = 5$ and, thus, we set this as the default value. This procedure was done for every feature and every concept in order to choose the best parameter set (top-k, C_D) for each concept-feature configuration. A late fusion step would then output an overall prediction score. This simple late fusion technique is implemented by simple LE vector concatenation and an optional feature normalization step after the final step was evaluated, but led to marginally lower performance, thus it was not used for preparing the final submission. In the final step, a linear classifier is trained using the structure feature vectors of the labelled items as input. In our implementation, we opted for the use of SVM. Apart from classification performance considerations, it is important for retrieval applications that the classifier produces real-valued prediction scores for unlabelled items, so that they can be ranked per concept.

2.2 Concept detection using a same class model

A very large variety of features can be extracted from an image. For detecting different concepts, the use of specific features or modalities may be more appropriate than others. That is, it could be that for some concept, similarity according to some feature or modality between an image and some set of images that belong to a specific concept is a very strong indicator that the image belongs to that concept; whereas for other concepts similarity according to some other set of features may be more indicative. The second approach attempts to deal with this issue; i.e. to learn in an automatic manner which modalities should be used for the detection of specific concepts. It uses what is termed the "same class" model. A "same-class" model takes as input the set of pairwise dissimilarities between two images according to the set of features and modalities that are used and predicts if these two images belong to the same class.

There are two options for training and predicting with such a model. In the first, all images that belong to the target concept are used. Pairs of samples from these images are generated in order to come up with the positive examples of the classifier. Additionally, a set of images that do not belong to the target concept are selected and pairs of images consisting of an image that does and an image that does not include the concept are generated in order to come up with the

negative examples. For a new image, the pairwise distances between it and the set of reference images that belong to that class would be computed and fed into the classifier that would output a set of scores, each of which is a prediction if the new image belongs to the same class as each image in the reference set for that concept. A final fusion step would then output an overall prediction score. This approach is depicted in Figure 1.

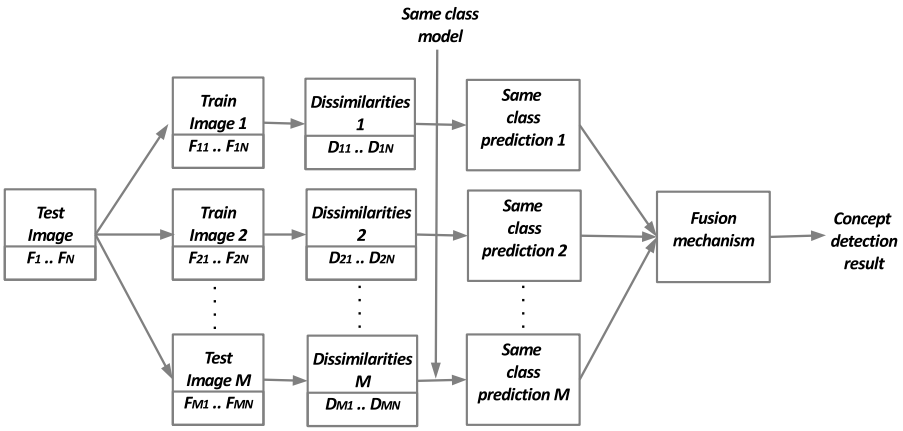

Fig. 1. A vector of dissimilarities for the set of used features is computed between the image to be annotated and the each of example the images that belong to the concept. These vectors are fed to the "same class" model and the predictions are fused to obtain a final prediction for the membership of the test image to the particular concept.

The first option essentially represents each concept by the set of images that belong to that concept and requires a final fusion step. The other option is to represent each concept using a mock average image, e.g. for each feature the average value for the images that belong to that class is computed and the set of all average values is used to build a prototype feature set for the items that belong to that class. The rest of the procedure is similar as in the first scenario: a set of positive examples is generated by computing the multimodal distances between images that belong to the target concept and the prototype representation of the concept. A set of negative examples is generated in a similar manner. When a new image is being annotated, the vector of distances between it and the prototype image is computed and fed into the classifier. Contrary to the previous case, there is no need for a final fusion step, as the classifier provides a single "same-class" prediction. This approach is depicted in Figure 2.

Compared to the first option, the second is more crude, in the sense that information from individual images that may be important for concept detection may be lost during averaging. On the other hand, the second option is computationally more efficient and does not require a final fusion step. Pursuing the

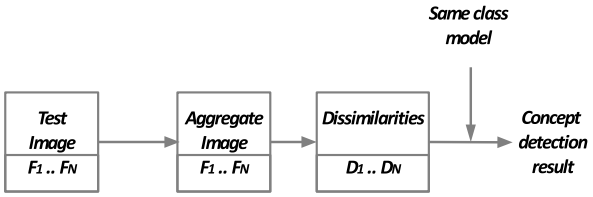

Fig. 2. A vector of dissimilarities for the set of used features is computed between the image to be annotated and the aggregate image that represents the concept. The vector of dissimilarities is fed to the same class model in order to obtain a final prediction for the membership of the test image to the particular concept.

second option, the hope is that the important parts of features will be so prevalent for each concept, that the averaging procedure will manage to maintain them in this generic prototype representation. In practice, the first option did not perform well in preliminary tests and the second option was eventually used.

The same class approach has been applied before for dealing with multimodal problems in a clustering task [9]. In that work, it is recognized that when attempting to cluster items that may be represented by multiple modalities or features, different clustering results that correspond to different conceptual organizations of the data may result by putting emphasis on different modalities (i.e. by following different fusion strategies). Instead of looking for appropriate fusion strategies, it was deemed interesting to allow an example clustering to guide the clustering procedure. The example clustering was used to obtain the "same-class" model, which in turn was used to group together items that had similar "same-class" relationships to the rest of the dataset.

3 Description of Runs and Results

3.1 Runs Description

This section describes the experimental settings of our submissions. Runs 1, 2, 3 and 5 are based on the first approach and Run 4 is based on the second. Runs 1, 2, 3, and 5 were performed on two quad core machines (Intel Quad Core i7-950 @3.07Ghz, 12G RAM and Intel Quad Core Q6600 @2.4Ghz, 8G RAM) and coded in Matlab. Run 4 was performed on a dual core machine (Intel Dual Core Q900 @2.5Ghz, 4G RAM) and coded in Java. Run 1 uses textual, Run 2 visual, while Runs 3-5 make use of both visual and textual features.

Run 1 (Approach 1, textual only): The total Mean interpolated Average Precision (MiAP) in this run was 0.2913 in training and 0.2311 in testing. In this run we used seven textual descriptors (Table 1). The TOP-TAGS feature was created using the 5000 most frequent tags. The TAGS-BOW textual feature

was extracted using the 5000-dimensional bag of words (BoW) representation following the approach of [13]. We took the union of raw tags of all images in the training set and applied stemming and stop word removal. This led to a vocabulary of approximately 32000 stems. Then, we applied feature selection to select the most important features using the χ^2_{max} criterion and finally selected the top 5000 features. The next three features were extracted using WordNet. The TAGS-WNET-TOP500 uses BoW representation using a codebook of 500 words. In order to define the codebook the full set of tags accompanying the ImageCLEF images was pre-processed by removing stop words and words not recognized by WordNet [3]. Then, the 500 most frequent tags were selected to compose the codebook. Finally, every image in the dataset was expressed as the occurrence count histogram of the codebook words in its set of tags, resulting in 500-dimensional feature vectors. As above, the TAGS-WNET-TOP5712 feature was extracted by selecting 5712 distinct tags instead of 500 to compose the codebook. The resulting feature vectors were 5712-dimensional. The last feature (TAGS-WNET-KRN-TOP500) was extracted using WordNet-based kernel similarities to enhance the semantic information enclosed by the BoW representation [11], by measuring the semantic relatedness of every word in the codebook with all other members of the codebook. Subsequently, the resulting matrix was multiplied with the original 500-dimensional BoW representation, to generate a new feature space with 500 dimensions. The last three features were extracted by applying probabilistic Latent Semantic Analysis (pLSA), a technique that considers a single document as a mixture of topics and learns the conditional distribution of features (words) given that some topic is present in the document [4]. According to this, the PLSA-TOP10000TAGS was extracted by applying pLSA on top 10000 tags feature vectors using 100 latent topics and the PLSA-TOPTAGS by applying pLSA on the top 10000 tags feature vectors using 100 latent topics respectively.

Run 2 (Approach 1, visual only): We achieved a MiAP of 0.3118 in training and 0.2628 in testing. In both training and testing, visual features were found to yield higher scores than textual. We used Dense and Harris Laplace sampling to extract keypoints. For local feature aggregation, hard assignment was used only in the TOPSURF+BOW descriptor, while Vector of Locally Aggregating Descriptors (VL) [5] was used for the rest. Two of the used visual features include the GIST and TOPSURF+BOW descriptors made available by the ImageCLEF organizers. The SURF features were extracted from all training images and codebooks of sizes k = 64, 128 and 256 were learned using the k-means algorithm (code provided by the authors of [12]). This process led to three sets of SURF+VL features with dimensionalities 64x64 (4096), 64x128 (8192) and 64x256 (16384). The final vectors where power (a=0.5) and L2 normalized. The SIFT(D)+VL features, were computed in the same way as SURF+VL using codebooks of k=64 visual words, with dimensionalities 64x128 (8192) computed on a dense multi-scale grid. The HUESIFT(D)+VL feature where computed in the same way as SURF+VL using codebooks of k=64 visual words, with dimensionalities 64*165 (10560) computed on a dense multi-scale grid. The RG-

BSIFT(D)+VL, OPPONENTSIFT(D)+VL, RGSIFT(D)+VL, CSIFT(D)+VL and HSVSIFT(D)+VL where computed in the same way as SURF+VL using codebooks of k=64 visual words, with dimensionalities 64x384 (24576) computed on a dense multi-scale grid. The SIFT(H)+VL, RGBSIFT(H)+VL, RGSIFT(H) +VL and HUESIFT(H)+VL were computed in the same way as SURF+VL using codebooks of k=64 visual words, with dimensionalities 64x128 (8192, SIFT) and 64x384 (24576) where regions found with Harris Laplace keypoint detector. In the end, we used the GIST-PLSA by applying pLSA on the GIST feature vectors using 100 latent topics. In total, we combined 17 different visual features.

Run 3 (Approach 1, multimodal): In this run, MiAP was 0.3894 in training and 0.3210 in testing. This was the best MiAP performance achieved by CERTH. Figure 3 illustrates the MiAP for each concept for this run. In this run all aforementioned features and also the hybrid feature which combines the GIST and TOP-TAGS descriptors by applying pLSA were used. More specifically, the pLSA model was applied independently in both the GIST and TOP-TAGS features resulting in the 100-dimensional GIST-PLSA and PLSA-TOPTAGS. Motivated by the fact that both feature spaces refer to latent semantic spaces and express probabilities (i.e., the degree to which a certain topic exists in the image), we assume that the topics obtained from both modalities are homogeneous and can be indiscriminately considered as the words of a common Topic Word Vocabulary. Based on this assumption we applied a second level pLSA model that operates on the feature space generated by concatenating the GIST-PLSA and PLSA-TOPTAGS (i.e. 100 + 100 = 200-dimensions). In total we combined 25 visual and textual features.

Run 4 (Approach 2, multimodal): In this run, MiAP was 0.3014 in the training set and 0.2887 in the test set. Figure 4 illustrates the MiAP for each concept for this run. SVM was used in order to learn the same class model. The following set of features were used: textual using only the tags (no stemming and stop word removal was applied) and a bag of words representation, SURF using a bag of words representation, SURF using a VLAD aggregation scheme with 2048 dimensions and GIST. For each concept, a separate same class model was used. The positive examples for each model were obtained by selecting all items that belong to the concept and computing the set of distances between them and the prototype of the concept. The negative examples were obtained by randomly sampling a number of images that do not belong to the concept. The number of negative examples was equal to the number of positive examples.

Run 5 (Approach 1, multimodal): In the final run, MiAP was 0.3769 and 0.3012 in the training and test set respectively. MiAP was not better than Run 3, to which the features were similar, but we managed to achieve higher F-measure (0.495) in the test set. In this run all features of Run 3 were used except the ones that were pre-processed with pLSA and extracted using WordNet (Table 1, features 1, 3-19).

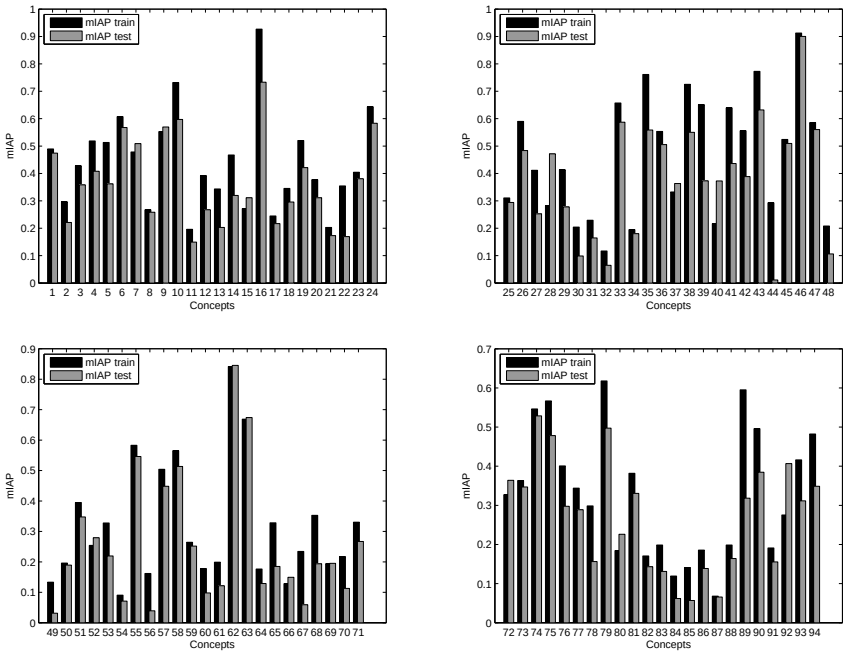

Fig. 3. MiAP per concept for Run 3

#	Descriptor	Dims	MiAP	#	Descriptor	Dims	MiAP
1	GIST	480	0.21026	**13**	HSVSIFT(D)+VL	24576	0.26629
2	GIST-PLSA	100	0.21604	**14**	SIFT(H)+VL	8192	0.2561
3	TOPSURF+BOW	200k	0.1469	**15**	RGBSIFT(H)+VL	24576	0.27177
4	SURF+VL	4096	0.23483	**16**	RGSIFT(H)+VL	24576	0.25235
5	SURF+VL	8192	0.23722	**17**	HUESIFT(H)+VL	10560	0.24541
6	SURF+VL	16384	0.23667	**18**	TOP-TAGS	500	0.2739
7	SIFT(D)+VL	8192	0.26377	**19**	TAGS-BOW	5000	0.29369
8	HUESIFT(D)+VL	10560	0.25883	**20**	PLSA-TOPTAGS	100	0.22639
9	RGBSIFT(D)+VL	24576	0.27672	**21**	PLSA-GIST-TOPTAGS	200	0.24506
10	OPP-SIFT(D)+VL	24576	0.27718	**22**	PLSA-TOP10000TAGS	100	0.20751
11	RGSIFT(D)+VL	24576	0.25368	**23**	TAGS-WNet-TOP500	500	0.27691
12	CSIFT(D)+VL	24576	0.27214	**24**	TAGS-WNET-TOP5712	5712	0.28025
13	HSVSIFT(D)+VL	24576	0.26629	**25**	TAGS-WNET-KRN-TOP500	500	0.22877

Table 1. The MiAP scores for each descriptor, **D** stands for Dense grid and **H** stands for Harris Laplace keypoint Detector, and **VL** stands for VLAD [5].

3.2 Evaluation

Feature comparison: Table 1 compares individual feature performance. Visual features RGBSIFT(D)+VL, OPP-SFIT(D)+VL , CSIFT(D)+VL and RGB-

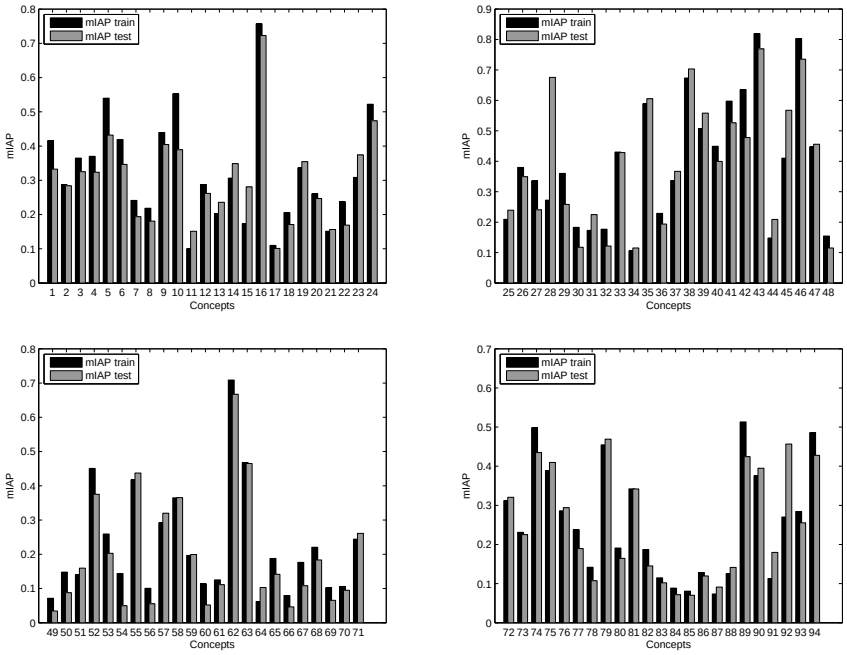

Fig. 4. MiAP per concept for Run 4

SIFT(H)+VL and textual features TOP-TAGS, TAGS-BOW and TAGS-WNET-TOP5712 achieved the best MiAP scores compared to the rest.

Approach 1 vs approach 2: Figure 5 illustrates the MiAP score for each run comparing the performance we achieved in the training set with the one in the test set. Apparently, Runs 3 and 5 suffer from overfitting, while Run 4 appears to generalize better. Furthermore, in some concepts one approach does better than the other. Run 3, based on the first approach is slightly better than Run 4 in the majority of concepts separately (50 concepts). Run 3 does much better in concepts *celestial stars* (6, Figure 3), *weather clearsky* (7), *weather rainbow* (10), *flora grass* (36) and *quality partialblur* (63), while Run 4 does much better in concepts *water underwater* (28, Figure 4), *fauna horse* (39) and *fauna amphibianreptile* (44).

Comparison to competing teams: Comparing per concept our best performance (Run 3) to other competitors, good performance was achieved (in terms of MiAP) in eight concepts and relatively low performance in six concepts. Specifically, our approach yields good performance in concepts *weather rainbow*, *combustion fireworks*, *flora plant*, *fauna spider*, *sentiment euphoric*, *combustion smoke*, *style graycolor* and *transport truckbus*, while it yields low performance in concepts *water other*, *fauna amphibianreptile*, *quantity two*, *quantity three*, *age elderly* and *sentiment unpleasant*. Finally, Tables 2 and 3 provide an impres-

sion of the standing of CERTH's performance against competing teams. Table 2 presents the rank of CERTH's best submission both at run-level (80 runs in total) and at team level (18 competing teams) in terms of the three performance measures. Table 3 presents the ranks of all CERTH runs compared to runs of the same type of features (textual, visual, multimodal).

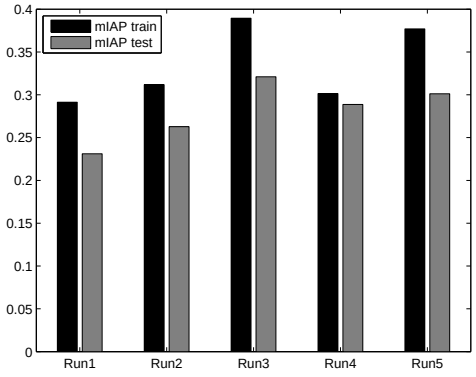

Fig. 5. MiAP for all Runs

measures	**score**	**Run-Level Rank**	**Best-Run Rank**
MiAP	0.3210	28/80	7/18
GMiAP	0.2547	29/80	7/18
F-ex	0.4950	27/80	6/18

Table 2. The test scores from ImageCLEF competition and the best rank

Runs	**features**	**MiAP**	**GMiAP**	**F-ex**
1	textual	0.2311 (5/17)	0.1669 (7/17)	0.3946 (7/17)
2	visual	0.2628 (13/28)	0.1904 (13/28)	0.4838 (10/28)
3	Multimodal All	0.3210 (15/35)	0.2547 (15/35)	0.4899 (18/35)
4	Multimodal gp	0.2887 (18/35)	0.2314 (18/35)	0.2234 (32/35)
5	Multimodal l	0.3012 (17/35)	0.2286 (19/35)	0.4950 (17/35)

Table 3. The test scores from ImageCLEF competition and the level run

4 Discussion

According to the obtained results, CERTH's performance ranks a bit higher than median. This leaves much room for improving performance in the future. An obvious option to achieve this is to use enhanced features. According to Table 3 particular emphasis should be placed on visual features. A second option for improving the performance of the first approach is to avoid overfitting by devising a more robust training process. A further option for improving performance stems from the fact that each image may be related to more than one concepts. For the same class approach, this implies that the average feature for each concept captures not only characteristics of the concept but also some of the characteristics of other concepts frequently co-occuring with it. This could lead to false positives for images not related to the concept but carrying these characteristics due to their relevance to these related concepts. Moreover, in some cases, when these characteristics are very prevalent they may even dominate the representation of the concept, leading to false negatives.

There is a lot of space for improvement considering the fact that we are dealing with a multi-label classification problem [15]. That is, from a probabilistic point of view, the occurrence of many concepts is not independent of the occurrence of other concepts and therefore, the estimates about the occurrence of a concept could be refined using the estimates about the occurrence of other concepts. There have been many approaches for dealing with this problem, for instance [2], which builds a chain of binary classifiers (one for each concept) where the input space of each classifier is augmented by the decisions of previous classifiers and [14] where a set of meta-classifiers are stacked upon the decisions of independent binary classifiers.

Acknowledgements This work was supported by the SocialSensor project, partially funded by the European Commission, under contract number FP7-287975. We also thank Eleftherios Spyromitros-Xioufis for providing us with the VLAD-based features and the features in [13], and Spiros Nikolopoulos for providing us with the pLSA-based features.

References

1. Thomee Bart and Popescu Adrian. Overview of the clef 2012 flickr photo annotation and retrieval task. in the working notes for the clef 2012 labs and workshop. Rome, Italy, 2012.
2. Krzysztof Dembczynski, Weiwei Cheng, and Eyke Hüllermeier. Bayes optimal multilabel classification via probabilistic classifier chains. In Johannes Fürnkranz and Thorsten Joachims, editors, *ICML*, pages 279–286. Omnipress, 2010.
3. C. Fellbaum, editor. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press, 1998.
4. T. Hofmann, editor. *Probabilistic latent semantic analysis, in: Proc. of Uncertainty in Artificial Intelligence*, Stockholm, 1999. UAI99.

5. H. Jégou, M. Douze, C. Schmid, and P. Pérez. Aggregating local descriptors into a compact image representation. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 3304–3311. IEEE, 2010.
6. Tang L. and Liu H. Leveraging social media networks for classification. *Data Min. Knowl. Discov.*, pages 23(3):447–478, Nov 2011.
7. Belkin M. and Niyogi P. Laplacian eignemaps for dimensionality reduction and data representation. *Neural Computing*, pages 15(6):1373–1396, June 2003.
8. Wang M. and Hua X.-S. Beyond distance measurement: Constructing neighborhood similarity for video annotation. pages 11(3):465–476. IEEE Transactions on Multimedia, April 2009.
9. Georgios Petkos, Symeon Papadopoulos, and Yiannis Kompatsiaris. Social event detection using multimodal clustering and integrating supervisory signals. In *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval*, ICMR '12, pages 23:1–23:8, 2012.
10. Macskassy S.A. and Provost F. Classification in networked data: A toolkit and a univariate case study. *JMLR*, pages 8:935–983, May 2007.
11. Patwardhan Siddharth. Incorporating dictionary and corpus information into a context vector measure of semantic relatedness. Master's thesis, August 2003.
12. E. Spyromitros-Xioufis, S. Papadopoulos, I. Kompatsiaris, G. Tsoumakas, and I. Vlahavas. An empirical study on the combination of surf features with vlad vectors for image search. In *Image Analysis for Multimedia Interactive Services (WIAMIS), 2012 13th International Workshop on*, pages 1–4. IEEE, 2012.
13. E. Spyromitros-Xioufis, K. Sechidis, G. Tsoumakas, and I. Vlahavas. Mlkd's participation at the clef 2011 photo annotation and concept-based retrieval tasks. In *ImageClef Lab of CLEF 2011 Conference on Multilingual and Multimodal Information Access Evaluation*, 2011.
14. G. Tsoumakas, A. Dimou, E. Spyromitros-Xioufis, V. Mezaris, I. Kompatsiaris, and I. Vlahavas. Correlation-based pruning of stacked binary relevance models for multi-label learning. In *Proceedings of the 1st International Workshop on Learning from Multi-Label Data (MLD'09)*, pages 101–116, 2009.
15. G. Tsoumakas and I. Katakis. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining (IJDWM)*, 3(3):1–13, 2007.
16. Zhu X. *semi-supervised learning with graphs.* PhD thesis, Pittsburgh, USA, 2005.