Consortium AVEIR at ImageCLEFphoto 2008: on the fusion of runs

Sabrina Tollari[1], Marcin Detyniecki[1], Marin Ferecatu[2], Hervé Glotin[3], Philippe Mulhem[4], Massih-Reza Amini[1], Ali Fakeri-Tabrizi[1], Patrick Gallinari[1], Hichem Sahbi[2], Zhong-Qiu Zhao[3]

[1]Université Pierre et Marie Curie-Paris6, UMR CNRS 7606-LIP6, Paris, firstname.lastname@lip6.fr
[2]TELECOM ParisTech, UMR CNRS 5141 LTCI, Paris, firstname.lastname@telecom-paristech.fr
[3]Université du Sud Toulon-Var, UMR CNRS 6168 LSIS, Toulon, name@univ-tln.fr
[4]Université Joseph Fourier, UMR CNRS 5217 LIG, Grenoble, firstname.lastname@imag.fr

Abstract

In this working note, we present the submission of the AVEIR consortium, composed of 4 French laboratories, to ImageCLEFphoto 2008. The submitted runs correspond to different fusion strategies applied to four individual ranks, each proposed by an AVEIR consortium partner. In particular, we study the complete, and partial, average of the ranking values, the minimum of these values, and a random based diversification. We first briefly describe the individual run of each partner, then we describe the fusion runs. The official results classed one of the runs, the MEAN fusion, as the third best in the automatic text-image run category. This run gives better results than the best partner run.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software; H.3.7 Digital Libraries; H.2.3 [**Database Management**]: Languages—*Query Languages*

General Terms

Measurement, Performance, Experimentation

Keywords

Rank Fusion, Image Retrieval, Multimodal Information Retrieval

1 Introduction

AVEIR (Automatic annotation and Visual concept Extraction for Image Retrieval) is the name of a project supported by the French National Agency of Research (ANR-06-MDCA-002). A consortium of four French CNRS research laboratories are involved in the project:

LIG Laboratoire d'Informatique de Grenoble at the Université Joseph Fourier (UJF),

LIP6 Laboratoire d'Informatique de Paris 6 at the Université Pierre et Marie Curie-Paris 6 (UPMC),

LSIS Laboratoire des Sciences de l'Information et des Systèmes at the Université du Sud Toulon-Var (USTV),

LTCI Laboratoire Traitement et Communication de l'Information at the TELECOM ParisTech.

The overall goal of the project is to enrich image retrieval systems with semantic indexation and annotation, and with symbolic relational description, all being automatically extracted and built from the textual and image content extracted from documents or web pages. This semantic and symbolic information are, then, used to reduce the visual ambiguity in images and to enhance the retrieval of images from large databases. The project develops 3 research axes. The first axis focuses on image analysis, feature extraction and visual feature representations. The second axis is concerned with automatic labeling of image components or objects with textual concepts. The third axis considers image retrieval and evaluation of the proposed algorithms. For more details please refer to `http://aveir.lip6.fr`.

In order to compare the state of the art approaches, each of the partners participated individually to ImageCLEFphoto (cf. [1, 2, 3, 4]).

The particularity of the 2008 ImageCLEFphoto edition was its focus on diversity. The evaluation was based on two measures: precision at 20 and instance recall at rank 20 (also called cluster recall or S-recall), which calculates the percentage of different classes or clusters represented in the top 20. The idea behind these measures was to focus on relevant but diverse - in terms of clusters - images.

In order to analyze if combining different runs improves the diversity, a submission under the label AVEIR was proposed. In this paper we briefly discuss the former submission, in particular the different fusion strategies, and the results.

2 Description of individual runs

Although each of the partners had its own diversification strategy, for the fusion we used the non diversified runs. In table 1, we briefly describe each the used runs. For more details please refer to the specific papers:

LIG_histo_3_p_o_1.5_4_0_NOCLUST_EN-EN-AUTO-TXTIMG [3]: this run is based on the linear combination of the scores provided by a language model using Dirichlet smoothing on the text and by a Jeffrey-Divergence correspondence on the images.

UPMC-LIP6_r3tfidf_VCDTWN_EN-EN-AUTO-TXTIMG [4]: the text processing is based on standard TF-IDF with cosine similarity. Forest of Fuzzy Decision Trees (FFDT) trained on VCDT ImageCLEF task 2008 are used for a visual concept filtering of the textual results. The matching of the concepts and the topics text used WordNet

LSIS_EN-EN-AUTO-TXTIMG-AUTO_GLOZHA_ar_12_NOCLUST [2]: the visual features are entropic features. Lots of SVMs are trained and generated with different parameters using the sample images provided. Then the first 20 images of the LIG run are used as the positive samples for each topic, and the others as the negative samples to construct the validation set for selecting the best one among the generated SVMs.

PTECH-EN-EN-AUTO-TXTIMG-AMKNR [1]: the run uses a combination of text and image descriptors. For a given topic, a separate query is performed for each modality (text and image). The results are merged by a minimum rank criterion: each image keeps the best rank.

3 Description of AVEIR runs

The AVEIR consortium proposed, for ImageCLEFphoto2008, 4 runs each with a different fusion strategy. Since for each partner's run, we have at most 1000 images ranked by topic, some images are sometimes not ranked.

Partner	Text preprocessing	Visual descriptors	Approach
LIG [3]	use of the <narr> field, stopwords, Porter's lemmatization	grid segmentation into 9 regions, RGB histograms, Jeffrey-divergence	language model with Dirichlet smoothing, linear combination of text and image results
LIP6 [4]	<narr> without sentences containing "not", stopwords adapted to image retrieval	segmentation into 9 overlapping regions, HSV histo for VCDT task, no other visual in ImageCLEFphoto	TF-IDF, Forest of Fuzzy Decision Trees (FFDT) used for learning VCDT concepts, use of WordNet for the matching of VCDT concepts and the topics, visual filtering using VCDT concepts
LSIS [2]	–	RGB entropic features	use of LIG's results to perform visual queries with a two-class SVM with Gaussian Kernel
PTECH [1]	<narr> field, stopwords, Porter's lemmatization, linear PCA	color, texture, shape	visual queries performed with a two-class SVM with Laplacian Kernel

Table 1: Short description of runs used by AVEIR for the fusion. For more details please refer to partners' papers.

AVEIR_LIG_LIP6_LSIS_PTECH_EN-EN-AUTO-TXTIMG_MIN: for each image, the fusion-rank corresponds to the minimum rank observed on each of the 4 partner's runs. This strategy corresponds to creating a rank by alternatively choosing an image from each of the partners' runs. The first image of the fusion rank corresponds to the first image of the first partner; the second image corresponds to the first image of the second partner; the fifth corresponds to the second image of the first partner, and so on.

AVEIR_LIG_LIP6_LSIS_PTECH_EN-EN-AUTO-TXTIMG_MEAN: for each image, the fusion-rank corresponds to the average rank observed on each of the 4 partner's runs. This strategy corresponds to a compromise taking into account all the systems. Images not present in one of the ranked lists are considered as having rank 1001.

AVEIR_LIG_LIP6_LSIS_PTECH_EN-EN-AUTO-TXTIMG_MEAN2on4: here only images that were ranked by at least two partners where considered. The fusion-rank correspond to the average of the available ranks. The idea behind this strategy is to avoid fusionning images returned only by one partner.

AVEIR_LIG_LIP6_LSIS_PTECH_EN-EN-AUTO-TXTIMG_MEAN_DIVALEA40: the first 40 images of the MEAN run were randomly shuffled. The objective of this run is to observe how randomness affects diversity and to provide a baseline for the instance recall.

4 Results and discussion

Figure 1 compares Precision and Cluster Recall when considering the first n retrieved images. The average precision at 20 (P20) and the average cluster recall at 20 (CR20), of the best 4 runs from each participating group (25 groups and 100 runs), was respectively P20= 0.32 and CR20= 0.35. All the fusion strategies are above these scores. This may be explained by the fact that some of the partners' runs performed very well.

When comparing MEAN and MEAN DIV (for $n < 40$) on figure 1, we conclude that a random diversification worsens the results as well for the precision as for the cluster recall. In terms of precision, the best fusion strategy is the MEAN, the worse being the MIN. In other words, from

Run	P20	Gain %	CR20	Gain %	MAP	Gain %
Run of partner X	0.260	-	0.293	-	0.191	-
Run of partner Y*	0.292	-	0.383	-	0.155	-
Run of partner Z	0.303	-	0.380	-	0.212	-
Best individual run (PTECH)	0.400	(ref)	**0.487**	(ref)	0.264	(ref)
AVEIR MIN	0.337	-16	0.462	-5	0.236	-11
AVEIR MEAN2on4	0.346	-13	0.431	-11	0.244	-8
AVEIR MEAN	**0.420**	**+5**	0.463	-5	**0.303**	**+15**
AVEIR MEAN DIVALEA40	0.377	-6	0.458	-6	0.274	+11
ImageCLEFphoto Average	0.320	-20	0.353	-28	0.219	-17
Best EN-AUTO-TXTIMG run	**0.512**	+28	0.426	-13	**0.366**	+39

Table 2: Individual runs, AVEIR's fusion runs, ImageCLEFphoto Average and Best ImageCLEFphoto EN-AUTO-TXTIMG run in terms of precision at 20 (P20), cluster recall at 20 (CR20) and Mean Average Precision (MAP). Partner runs are ordered from worst to best precision at 20. The gains are calculated in function of the best individual run score (ref) *this run was not submitted

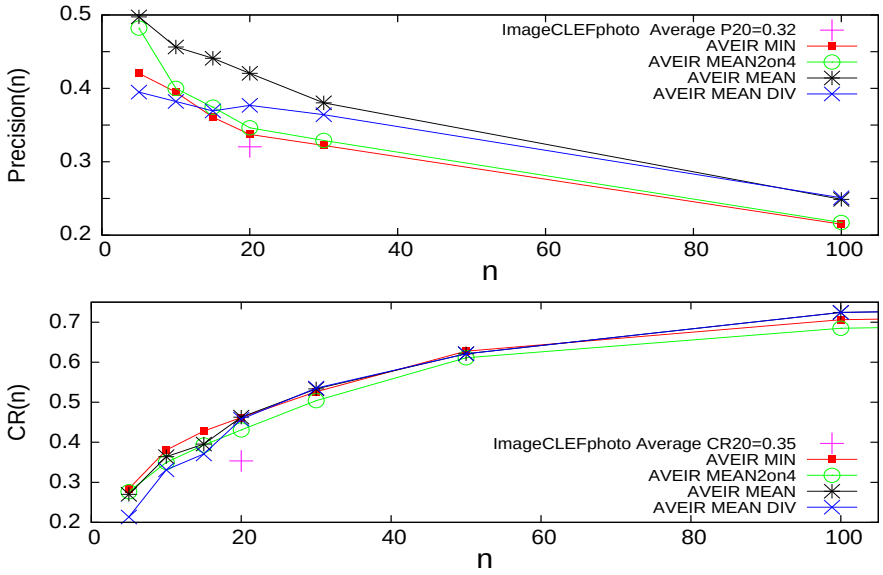


Figure 1: Precision and Cluster Recall when considering the n first retrieved images

the precision point of view, it is more interesting to base the fusion on a *compromise*. In fact, the MIN strategy considers an image as very good as long as one of the partners, independently of the others, ran it high. The best images, when using the MEAN strategy, correspond to images that were highly ranked by *all* the systems.

Surprisingly, from the cluster precision perspective, in average (over all the topics), there is not much difference between the runs, although the MIN slightly outperforms the other strategies (in particular when considering the very first images). If we look at figure 2(a), we discover that there are topics for which the MIN strategy is better and topics for which the MEAN is better. Although the reasons behind this behaviour needs further research, it explains why in average there is no difference between the two strategies.

Table 2 compares the *best individual run* with the AVEIR fusion runs. Only the MEAN strategy shows an improvement with respect to the best of the individual runs. The Mean Average Precision is clearly improved. There is no improvement in the cluster recall, actually there is a slight drop. The explanation lies in the behaviour per topic. On figure 2(b), we observe that for some topics the best individual run outperforms any fusion, while for others the fusion improves beyond the

best individual run. The fusion is not correlated to best individual run. Furthermore, the topics that have a high cluster recall score (i.e. with a score higher than 0.5 for as well for the MEAN as for the Best Individual) are better served by the best individual run, while the ones with a low score are better with the MEAN fusion. The compromise pays, in terms of diversity, when the problem is difficult. This may be explained by our previous observation that the MEAN improves the precision. In fact, for difficult topics, the MEAN brigs new images up, increases the precision and the cluster recall, since a new relevant image belongs with a high probability to a new class.

5 Conclusion

In this working note, we presented the submission, of the AVEIR consortium, to ImageCLEFphoto 2008. The particularity of this year edition was its focus on diversity. The evaluation was based on the relevance, measured by the precision at 20 and by the diversity measured by the cluster recall at rank 20. The idea behind these two measures was to focus on relevant but diverse images.

The submitted runs correspond to different fusion strategies applied to four individual ranks, each proposed by a partner. In particular we study the complete, and partial, average of the rank values (MEAN and MEAN2on4), the minimum of these values (MIN), and a random based diversification (DIVALEA40). The official results[1] classed one of the runs, the MEAN fusion, as the third best. Our experiments showed why this fusion particularly improves the precision and, even more, the mean average precision. We also observed that it only slightly affects the diversity.

Furthermore, the MIN fusion - which corresponds to alternating images from each individual run - despite its weak precision at 20, improves slightly the overall diversity. The weak precision at 20 may be partially explained by the disparity, in terms of quality, of the runs. In fact, low scoring runs bring non-relevant images, lowering the precision, but also keeping the diversity of the runs' information.

Finally, the experiments also pointed out that the diversity is strongly affected by the relevance, in particularly for difficult queries. Although the experiments showed that, in terms of diversity, the *best individual* run performs better than any type of fusion, we observed that for low precision topics it is more interesting to perform a MEAN fusion (that increases the mean average precision) and that for high precision topics it is more interesting to fusion with the MIN (as long as the runs have similar performance). In other terms diversity comes after a good relevance.

Acknowledgment

This work was supported by the French National Agency of Research (ANR-06-MDCA-002).

References

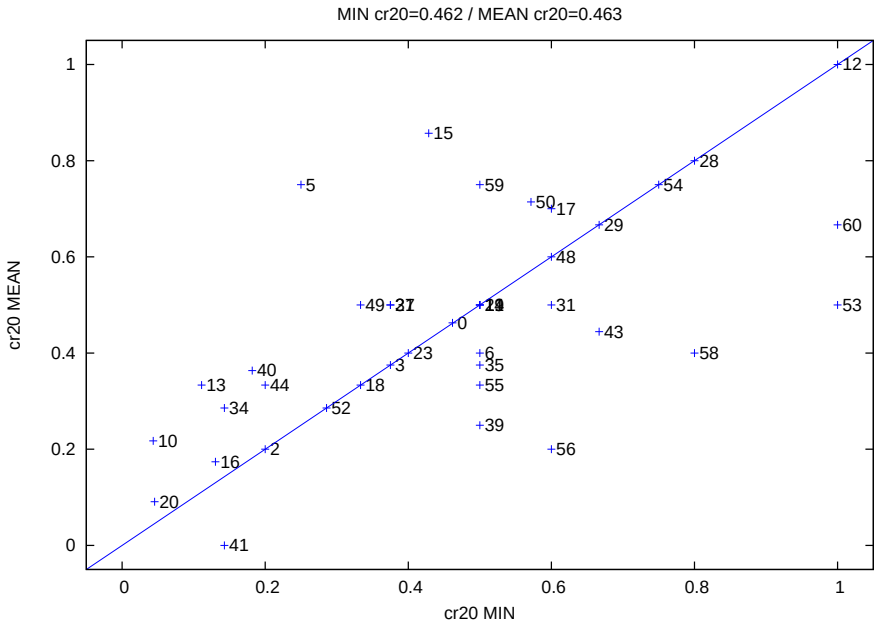
[1] Marin Ferecatu and Hichem Sahbi. TELECOM ParisTech at ImageClefphoto 2008: Bi-modal text and image retrieval with diversity enhancement. In *Working Notes of ImageCLEFphoto2008*, 2008.

[2] Hervé Glotin and Zhong-Qiu Zhao. Affinity propagation promoting diversity in visuo-entropic and text features for clef photo retrieval 2008 campaign. In *Working Notes of ImageCLEFphoto2008*, 2008.

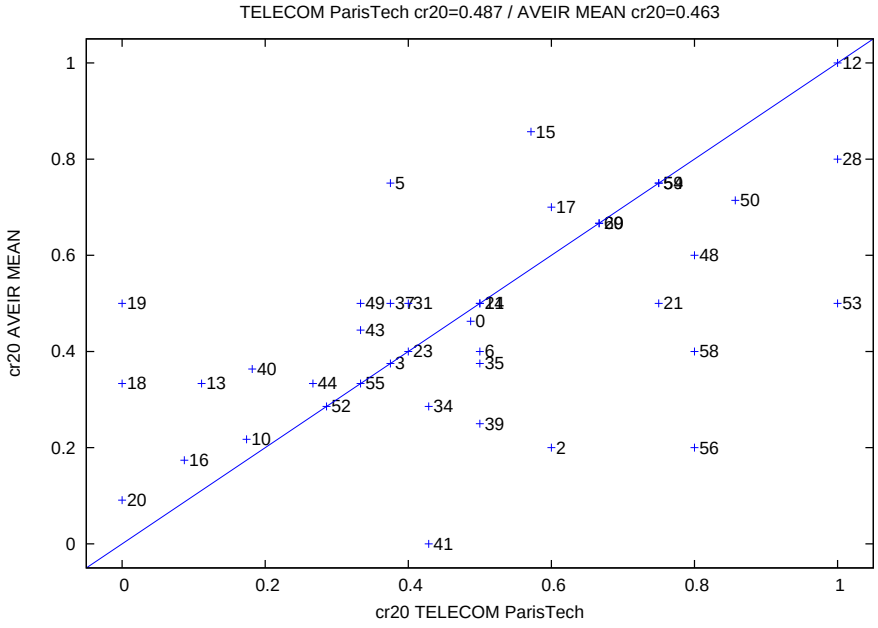
[3] Philippe Mulhem. LIG at ImageCLEFphoto 2008. In *Working Notes of ImageCLEFphoto2008*, 2008.

[4] Sabrina Tollari, Marcin Detyniecki, Ali Fakeri-Tabrizi, Massih-Reza Amini, and Patrick Gallinari. UPMC/LIP6 at ImageCLEFphoto 2008: on the exploitation of visual concepts (VCDT). In *Working Notes of ImageCLEFphoto2008*, 2008.

[1] http://www.imageclef.org/2008/results-photo



(a) MIN vs MEAN strategies



(b) Best individual run vs best fusion strategy (MEAN)

Figure 2: Comparison of some AVEIR results by topic in terms of Cluster Recall at 20 (CR20). Points are labeled by topic numbers. Point labeled 0 corresponds to Cluster Recall on all topic