# Comparison of Various AVEIR Visual Concept Detectors with an Index of Carefulness

H. Glotin[1], A. Fakeri-Tabrizi[3], P. Mulhem[4], M. Ferecatu[5], Z. Zhao[1,2], S. Tollari[3], G. Quenot [4], H. Sahbi [5], E. Dumont[1], P. Gallinari[3]

[1] Univ. Sud Toulon-Var, Systems & Information S ences, LSIS UMR CNRS 6168, Toulon, France

[2] School of Computer & Information, Hefei Univ. of Technology, China

glotin@univ-tln.fr; zhongqiuzhao@gmail.com; emilie.r.dumont@gmail.com

[3] Université Pierre et Marie Curie - Paris 6, UMR CNRS 7606 LIP6, F-75016 Paris, France

firstname.lastname@lip6.fr

[4] Univ. Joseph Fourier, Lab. d'Informatique de Grenoble, LIG UMR CNRS, Grenoble, France

firstname.lastname@imag.fr

[5] Institut TELECOM ParisTech, LTCI UMR CNRS 5141, Paris, France

Marin.Ferecatu@telecom-paristech.fr; Hichem.Sahbi@telecom-paristech.fr

**Abstract**

Visual annotation is still an open issue. The Content Based community admits that a plurality of features and systems shall be considered. We present in this paper four very different strategies using not only visual information but also text, to implement ImageCLEF2009 Photo Annotation Task. The visual features are various, such as HSV, Gabor, EDGE, SIFT, and some more recent. Then we study each model performances, and propose a new measure, the Carefulness Index (Q) computed on the histogram of the model's outputs. Q seems to be correlated with the model performances.

## 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**]: Cross-Language Retrieval in Image Collections (ImageCLEF)

## Keywords

Carefulness Index, SVM, Fusion, SIFT, Gabor, HSV, Profile Entropy Feature, Ontology

## 1 Introduction

This year, the annotation task focuses on scaling the algorithms to thousands of images and possibly more, which is a very difficult task. Indeed, image annotation is still an unsolved problem and recent state of the art algorithms perform less than satisfactorily on most image databases. The image annotation task uses 53 concepts, many of them being holistic, that is they are not associated with some part of an image, but with the visual impressions extracted from the whole image. Furthermore, even the concepts corresponding to objects are associated with the entire image and not to some part of it. Local methods, for example those based on the extraction of keypoints or image regions, are less likely to function correctly in this case.

In order to analyse which strategy shall be optimal for this kind of task, we depict four very different models that have been built independently to each other. We give their performances,

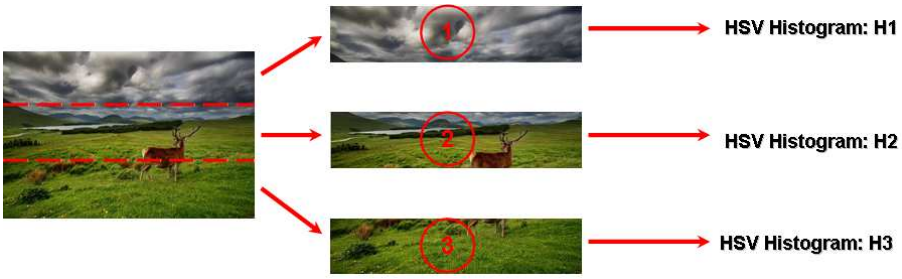


Figure 1: Dividing image on three horizontal segment to extract the histogram (HSV) of each part

and propose a new measure, the Carefulness Index (Q) computed on the histogram of the model's outputs. Q seems to be correlated with the model performances. We also analyse simple fusion models. In average the best model is the simplest, the arithmetic average, compared to the selection of the a priori best model, or an early fusion model.

The next section presents the four models, then the results are analysed and the Carefulness Index measure is proposed in section 4. Other comments on the models performances are given before to conclude.

# 2 The four different models

## 2.1 Model 1: SVM based on HSV with ROC loss function

We used the color-based visual descriptors in this model. As in [7], we segment the images into 3 horizontal regions with the same sizes. We believe that these visual segmentation is particularly interesting for general concepts (i.e. not objects), as for instance: sky, sunny, vegetation, sea... (see figure 1). For each region, we compute a color histogram in the HSV space.

We train a SVM classifier[1] which has a linear kernel.Because of the imbalanced class problem, we use a ROC area as the loss function as proposed in [1]. So we consider not only the misclassification in each learning iteration, but also the number of positive and negative examples in order to avoid the fault ignorance. ROCarea can be computed from the number of swapped pairs

$$SwappedPaires = \|\{\ (i,j) : (y_i > y_j) and (w^T x_i < w^T x_j)\}\|$$

i.e. the number of pairs of examples that are in the wrong order

$$ROCarea = 1 - \frac{SwappedPaires}{\#pos.\#neg}$$

. Here 1-ROCarea is used as the value of misclassification in loss function for each iteration. More details on this model can be found in [2].

## 2.2 Model 2: RBG, SIFT, Gabor, and ontology SVMs

This model uses three sets of features: The first one is based on 512 bins RGB histogram of the three horizontal stripes (same height of 1/3 of the image height, whole width of the image as presented in previous section). Histograms are normalized and they result is a 1536 histogram. The second set of features are SIFT, using software provided by K. van de Sande [SAND08]. The SIFT features are extracted from regions selected according to Harris-Laplace feature points detection. Each feature is a 128-dimension vector. A visual vocabulary containing 4000 dimensions was then generated using the SIFT features of the learning set, yielding to a 4000 dimensions vector for each

[1] `http://svmlight.joachims.org/svm_perf.html`

image. The third feature set, called HSVGAB, is an early fusion of colour and texture features. We used a 64 dimensions HSV colour histogram concatenated with a 40 dimensions vector describing gabor filters energy (7 dimensions, 5 scales). For the RGB, SIFT, and the HSVGAB features we used then a simple one against all SVM (RBF kernel) that learns the probability for one sample of belonging to each concept. For the SIFT features, we used additionally a multiple SVM leaning process. Consider one concept C having pc positive samples, and nc negative samples (ni = 5000 - pc ¿¿ pc). We define Nc SVM with all the positive samples and 2*pc negative samples, so that union the negative samples as all SVMs cover all the pc negatives samples of C. Each of these SVM learns the probability of belonging to each class concept/non concept. For one concept, we sum-up then the results for all the NC SVMs. We applied then a scaling in a way to fit the learning set a priori probabilities. Then we select the best feature/learning combination for each concept. We took into account the hierarchy of concept in the following way: a) when conflicts occur (for instance the tag Day and the tag Night are associated to one image of the test set), we keep unchanged the larger value tag, and we decrease (linearly) the value all the other conflicting tags, b) we propagated the concepts values in a bottom-up way if the values of the generic concept is increased, otherwise we do not update the values. More details can be found in [4].

### 2.3 Model 3: average of Gabor-HSV SVMs and of Visual Dictionary

This model is an average of only visual information models, based on SVM and Visual Dictionary approaches on some new features depicted in [5]. As some of these models were proposed for the first time, we decided to build for this paper an average model that is the arithmetic average of three sub-models.

In sum, the Model 3 is built from various visual features: HSV, EDGE, Gabor, and the recent DF and Profile Entropy Features (PEF) [7]. Firstly for each concept, we compute Linear Discriminant Analysis (LDA) and we train support vector machines (SVMs) [5]. We also consider the SVM trained on the PEF. Third, we merge a Visual Dictionary (VD) model, which constructs a concept visual dictionary composed by visual words [5]. We notice after the evaluation that this average is suboptimal, it is below the 8th AUC rank that is taken by one of its component (LSIS best run). However, it produces complementary estimates to the other models proposed in this paper.

### 2.4 Model 4: fast (unprecised) Canonical Correlation model

This model 4 is focused on global image descriptors and favors fast algorithms that can scale to thousands of images and annotation concepts. First, we represent each image using a text descriptor and a global visual descriptor. As visual descriptors we use global color, texture and shape features, similar to those presented in MPEG7. We use Canonical Correlation Analysis (CCA) to infer a latent space where the two representation are most correlated. Given the visual features of an unseen image, we fist project it to the CCA space and then we infer the linear combination of input concepts that is most correlated with it. We then back-project the result into the input space and we normalize it to [0, 1]. A value close to 1 means that the corresponding concept is likely to be found in the image, while a value close to zero suggests the contrary.

The tradeoff in our method is a slight loss of precision, but we make up for this in speed (we use less than 1 sec. for both training and prediction on an average 2.5 GHz PC). Moreover, adding new concepts to our method is straightforward and do not require training separate models for each concept. More details can be found in [6].

## 3 Results and discussions on a carefulness index

### 3.1 Global performances

We give the average Area Under the Curve (AUC) of the four models in figure 2 for each topics, and their average in table 1 including comparison to the best runs of each team participating to

the campaign. We see that AUC(Model 1) > AUC(Model 2) > AUC(Model 3) > AUC(Model 4).

For comparison, three basic fusion models are computed. The first one, called early fusion, is a SVM trained on the merged features of the four models. The second is the simple arithmetic average of the outputs of the four models (late fusion). The last one, called 'best1' is the selection of the best model according to the training performances.

Table 1: The official table results including the four models and the fusion AVEIR model, and the best runs of each submitting team of the ImageCLEF2009 Photo Annotation Task. The Rank is given only for the best run of each team.

| RANK | LAB | Average EER | AUC |
|---|---|---|---|
| 1 | ISIS | 0.234476 | 0.838699 |
| 2 | LEAR | 0.249469 | 0.823105 |
| 3 | FIR2 | 0.253566 | 0.817159 |
| 4 | CVIUI2R | 0.253296 | 0.813893 |
| 5 | XRCE | 0.267301 | 0.802704 |
| 6 | bpacad | 0.291718 | 0.773133 |
| 7 | MMIS | 0.312366 | 0.744231 |
| 8 | LSIS | 0.330819 | 0.720931 |
| 9 | IAM | 0.330401 | 0.714825 |
| + 10 | Model 1 | 0.372169 | 0.673089 |
| + 11 | Model 2 | 0.382840 | 0.644589 |
| + | Model 3 | 0.430236 | 0.600746 |
| * 12 | AVEIR | 0.440589 | 0.550866 |
| - | Random | 0.500280 | 0.499307 |
| 13 | CEA | 0.500495 | 0.469035 |
| + 14 | Model 4 | 0.526302 | 0.459922 |
| 15 | Wroclaw | 0.446024 | 0.220957 |
| 16 | KameyamaLab | 0.452374 | 0.164048 |
| 17 | UAIC | 0.479700 | 0.105589 |
| 18 | INAOE | 0.484685 | 0.099306 |
| 19 | apexlab | 0.482693 | 0.070400 |

The best fusion of the four models is the late fusion which gives an average AUC of 0.55, and occupies the 12th rank among the 19 teams in the official VCDT evaluation. Anyway, it is worst than the best model. We analyse in detail each model performances.

### 3.2 Performances are correlated with a Carefulness Index

In order to analyse each model results, we depict in figure 3 the histograms of the outputs of each model M1, ..., M4 on the test set. The shape of each histogram largely differs from one model to another. We then investigate a simple statistics that may indicates from this shape the quality of the model.

A detailed analyse of central and extreme values of these histograms reveal that for the best model, the center (bins 5 and 6) is bigger than the extremities (bins 1 and 10). We then compute a simple ratio:

$$Q = h(center)/h(extremities),$$

where $h$ is the histogram here of 10 bins, so $h(center) = h(5) + h(6)$ and $h(extremities) = h(1) + h10)$.

Table 2: Lists of the 10 topics having the lowest STD between the four model (LEFT), and the biggest STD (RIGHT)

| Ten topics with lowest STD | Ten best topics with highest STD |
|---|---|
| Fancy | Underexposed |
| Aesthetic-Impression | Beach-Holidays |
| Motion-Blur | Sunset-Sunrise |
| Partly-Blur | Night |
| Overall-Quality | Sea |
| No-Blur | Neutral-Illumination |
| Canvas | Clouds |
| Sunny | Landscape-Nature |
| Plants | Sky |
| Still-Life | Water |

Q is high if the border estimates of a model are rare, that is if the model if 'careful' (most of the decision are close to the decision boundary). Thus we call this index the 'Carefulness Index'.

In figure 4 we give the log(Q) values and the AUC results for each of the four models. We see that when Q decreases, AUC is also decreasing, moreover the ranks given by Q are similar to the AUC ranks.

# 4 Conclusion

Depicting the results of very different models we enlighted a simple statistics on the raw model outputs that seems to be tied to its performances. The very different models we tested have different carefulness index. The experiments show that more careful is a model, more it AUC increases. This result shall be confirmed on other raw distributions of other model outputs. This kind of global shape statistics are interesting for scaled systems, where fast and unsupervised estimates of visual detector quality shall be possible. Further work will be conducted in this field in the AVEIR group.

# Acknowledgment

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

# References

[1] Joachims, T.: A Support Vector Method for Multivariate Performance Measures, In Proceedings of the International Conference on Machine Learning (ICML) (2005)

[2] Fakeri-Tabrizi, A., Tollari, S., Denoyer, L., Gallinari, P.: UPMC/LIP6 at ImageCLEFannotation 2009, Large Scale Visual Concept Detection and Annotation, In CLEF working notes 2009 (2009).

[3] VandeSande, Gevers T., and Snoek C.: Evaluation of Color Descriptors for Object and Scene Recognition, In Proceedings of CVPR. Anchorage, Alaska, USA (2008)

[4] Mulhem et al.: MRIM-LIG at ImageCLEF2009 photo annotation, In CLEF working notes 2009 (2009)

[5] Zhao, Q., Glotin, H. and Dumont, E.: LSIS Scaled Photo Annotations - Discriminant Features SVM vs Visual Dictionary based on Image Frequency, In CLEF working notes 2009 (2009)

[6] Ferecatu, M. and Sahbi, H.: TELECOM ParisTech at ImageClef 2009: Large Scale Visual Concept Detection and Annotation Task, In CLEF working notes 2009 (2009)

[7] Glotin, H., Zhao, Z.Q., Ayache, S.: Efficient Image Concept Indexing by Harmonic & Arithmetic Profiles Entropy, IEEE International Conference on Image Processing, Cairo, Egypt, November 7-11 (2009)

[8] Nowak S., Dunker P.: Overview of the CLEF 2009 Large Scale - Visual Concept Detection and Annotation Task, CLEF working notes 2009, Corfu, Greece, (2009).

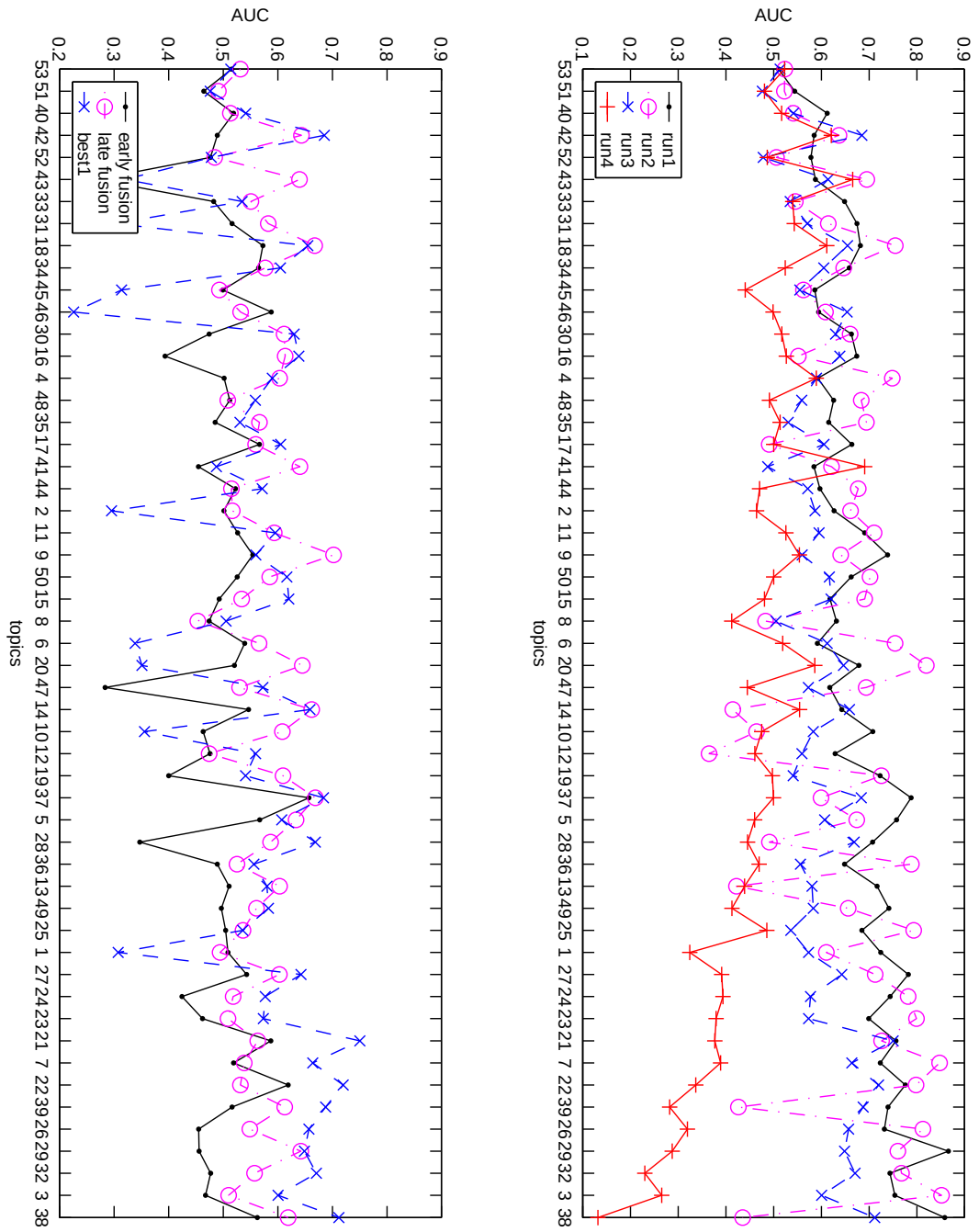


Figure 2: Area Under the Curve (AUC) evaluations for each Model and topic (number are the original ones given by the organizers). The topics are here sorted according to the STD between the four models (Right). The early, late and best1 fusions are depicted in the Left figure. We see that all these naive fusions are always worst than the best model. The best fusion is best1 for high STD between the four models, and the early fusion is nearly the worst for all topics.

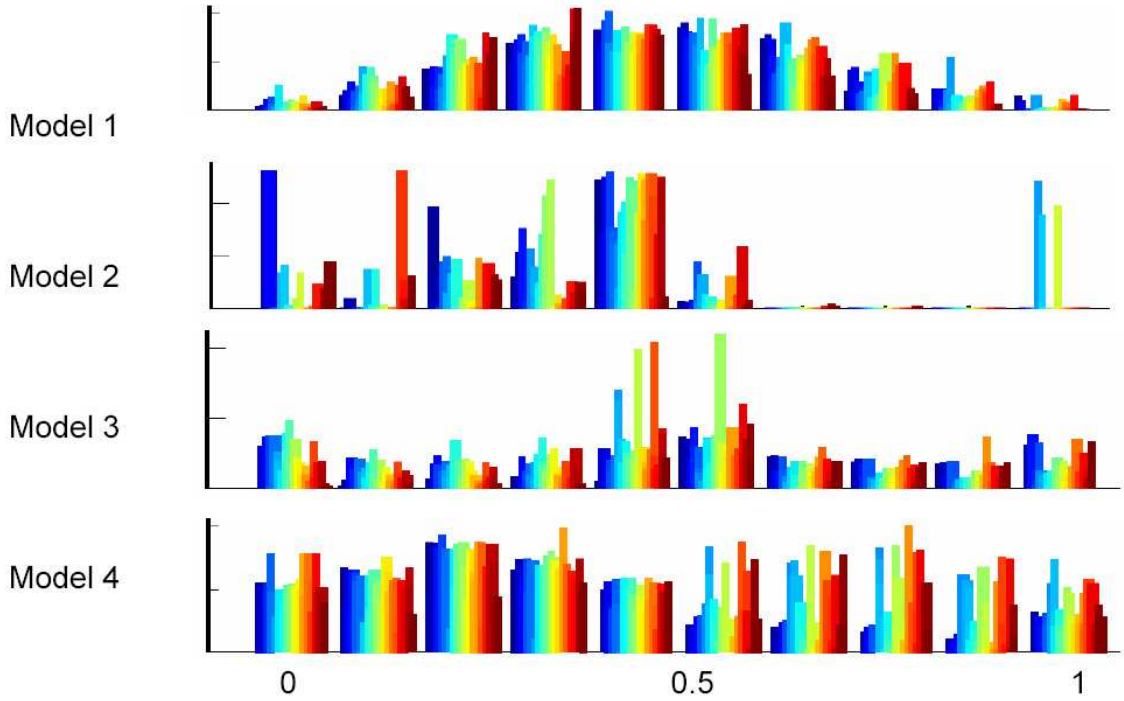


Figure 3: Histograms of the similarities of the concepts, generated by each of the four individual models (M1 to M4), and for each topics. The 53 topics are represented by incremental colors from blue to red. These histograms give the raw behavior of each models. The experiment shows that their central and extreme values have a simple relation with the AUC of the model: AUC(Model 1) > AUC(Model 2) > AUC(Model 3) > AUC(Model 4).

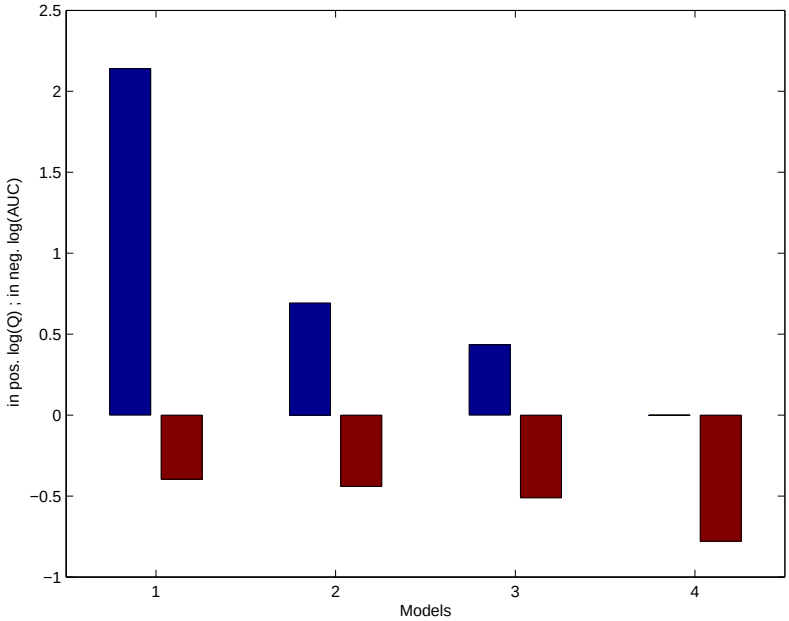


Figure 4: The relation between Q index and the AUC for the four models. Log(Q) are the positive (blue) values, while the negative (red) are the log(AUC). We see that when the carefulness index Q decreases, AUC is also decreasing, moreover the ranks given by Q are similar to the AUC ranks.