

TIA-INAOE's Participation at ImageCLEF 2009

Hugo Jair Escalante, Jesús A. González, Carlos A. Hernández, Aurelio López,
Manuel Montes, Eduardo Morales, Elias Ruiz, Luis E. Sucar, Luis Villaseñor

Research Group on Machine Learning for
Image Processing and Information Retrieval
Department of Computational Sciences
National Institute of Astrophysics, Optics and Electronics
Luis Enrique Erro No. 1, 72840, Puebla, México
hugojair@ccc.inaoep.mx

Abstract

This working note describes the participation of TIA-INAOE in the Photographic Retrieval and the Large Scale Image Annotation tracks at ImageCLEF2009. We developed specific methods for each track with the goal of exploiting the information available while maximizing annotation and retrieval performance. On the one hand, for the retrieval track, we proposed a post processing technique for re-ranking documents according to different diversity categories. With this formulation we considered both visual and textual features and we incorporated information of the different categories by which topics are clustered. Results obtained with this technique suggest it is a promising method for result diversification. However, we still need to deal with issues that affect the retrieval performance. On the other hand, for the annotation task, we adopted a simple annotation technique based on KNN classification. Only global features were considered under this formulation. The output of the KNN method was then refined by means of an energy-based model that attempts to maximize the *semantic cohesion* among labels assigned to each image. We considered information of the annotation hierarchy to constraint certain labeling configurations. Results obtained with this technique give evidence that the KNN approach is an effective annotation technique despite being rather simple. The refinement strategy resulted useful for improving the labeling performance of an ineffective annotation method; although we could not improve the labeling performance of a strong baseline. Summarizing, the results obtained at ImageCLEF2009 are encouraging and motivate further research in several directions that we are currently exploring.

Categories and Subject Descriptors

H.3 [Information Storage and Retrieval]: H.3.1 Content Analysis and Indexing; H.3.3 [Information Systems and Applications]: Information Search and Retrieval—*Retrieval models; Selection process; Information Filtering*

General Terms

Performance, Experimentation

Keywords

Multimedia image retrieval; Image annotation; Result diversification; Semantic cohesion modeling; Document re-ranking.

1 Introduction

This working note describes the participation of TIA-INAOE in the Photographic Retrieval and the Large Scale Image Annotation tracks at ImageCLEF2009. A total of 10 runs were submitted that comprise different settings of our proposed methods for image annotation and retrieval. The proposed methods aim exploiting the information available while maximizing annotation and retrieval performance.

On the one hand, for the retrieval task, we adopted a two stages retrieval process. In the first stage, an initial image search is performed by using only textual information to obtain potentially relevant documents. In a second stage the candidate documents are re-ranked by considering the different diversity categories (clusters) as provided by the organizers. For re-ranking we considered both visual and textual information. Results with this approach are mixed: whereas the re-ranking technique resulted helpful for diversifying retrieval results, the best retrieval performance (MAP) was obtained with the baseline retrieval method. This result suggests that textual retrieval methods are better suited for this collection.

On the other hand, for the annotation task, we adopted a three stages methodology. In a first stage, candidate labels for test images are selected with a KNN approach, which consists of obtaining the labels assigned to the K-nearest neighbors (in the training set) of the test image. In the second stage, an energy-based model is used to select, among the candidate labels, the *disjoint* labels for the image, see [4]; this model uses the output of the KNN method and label co-occurrence statistics. In a third stage, we consider the labels selected in the second stage to obtain the optional labels for the test image. Applying this method took about 0.25 seconds per image, which makes it attractive for large scale annotation. Annotation results are contradictory: we obtained acceptable performance when considering the evaluation measure based on the annotation hierarchy [5, 4]; however, the performance of our methods is rather limited in terms of EER and area under the ROC curve, we analyze these results below.

The rest of this document is organized as follows. In the next section we describe the approach we adopted for the photographic retrieval task and the results obtained with this technique. Next, in Section 3, we present the annotation method we proposed as well as official results of this method. Finally, in Section 4, we describe the conclusions derived from this work and we outline future work directions.

2 Photographic retrieval

We proposed a two stage approach for the photographic retrieval task at ImageCLEF2009; our methodology is depicted in Figure 1. In the first stage a set of m potentially relevant documents is retrieved by using a text-based image retrieval technique. In the second stage the m documents are re-ranked by taking into account the initial score assigned to documents and the similarity of candidate documents to the diversity clusters provided by the organizers. We proposed this formulation for two main reasons: 1) we wanted to take advantage of the topic clusters for diversifying retrieval results and 2) we wanted to make more efficient the search process. The latter was accomplished because the initial search can be performed efficiently (over the 500,000 documents) and once the set of potentially relevant documents is reduced to a subset of documents we can compare images in acceptable time and we can apply more complex strategies over this reduced set of documents. The rest of this section describes in detail our method. Further details on the task and on the collection are described by Lestari et al. [6].

2.1 Feature extraction

The proposed method considers both textual and visual information for representing documents. As textual features we consider a *tf-idf* weighting scheme over words for representing documents, see the next section. As visual features we use color histograms on both RGB (256 bins) and HSI (128 bins per channel), for a total of 640 visual attributes. We also performed preliminary

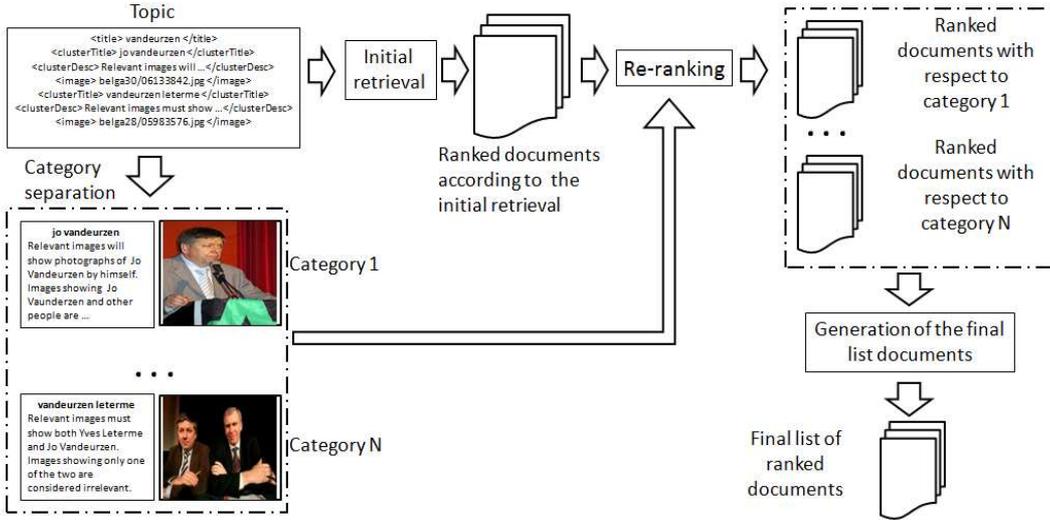


Figure 1: Diagram of the proposed approach to image retrieval.

experiments with other visual features, including: edge histograms, color histograms in CIE-Lab, texture features and local descriptors; however, the RGB and HSI color histograms resulted more effective (according to an empirical evaluation we conducted) for retrieving images under the Euclidean distance.

2.2 Initial retrieval

For the initial retrieval we considered the vector space model (VSM) under *tf-idf* weighting for representing documents [1]; we used the cosine similarity for comparing documents. In particular, we used the TMG Matlab^R toolbox for indexing and retrieval [9]. Because of the size of the collection we indexed the collection by batches of 20,000 documents.

We considered textual information only, because computing the Euclidean distance, between query images and the 500,000 images that compose the collection, would be very computationally expensive. Thus, the re-ranking approach was also used for efficiency reasons (query images and a reduced set of m -images can be compared in acceptable time).

For querying we used all of the textual information available in topics (i.e. title, cluster titles and cluster descriptions); this is motivated because at this stage we wanted to retrieve documents that were related to the query as a whole, so that in the next stage the search can be refined. We ranked the documents by their similarity to the query and we keep the $m = 1,000$ top ranked documents for the second stage; our baseline run consists of returning these 1,000 documents.

2.3 Re-ranking based on multimedia features

In the second stage we re-ranked the m -documents obtained by the initial retrieval method. For each topic category (cluster), j , we assigned a score to each of the m -documents, $d_{i \in \{1, \dots, m\}}$, as follows:

$$sc_{final}^X(d_i^X, q_j^X) = \lambda \times sc_{initial}(d_i) + \times S^X(d_i^X, q_j^X) \quad (1)$$

where $sc_{initial}$ is the similarity score obtained from the initial retrieval stage and $S^X(d_i^X, q_j^X)$ is an estimate of the similarity between the i^{th} -document and the j^{th} -sub-query under modality X . λ is a scalar weighting the contribution of the first term. A sub-query q_j^X is the part of the topic corresponding to the j^{th} diversity cluster, where $j \in \{1, \dots, C\}$ and C the number of categories associated with the topic. The superscript X indicates which information modality is used: $X = T$ means that textual information was considered (e.g. cluster title), $X = V$ indicates

that visual information was used (i.e. cluster image) and $X = M$ means that both textual and visual information were considered. When $X = T$, we used the cosine similarity as S^T ; when $X = V$ we used the (normalized) inverse of the Euclidean distance as S^V ; when $X = M$ we used $S^M = w_{m_1} \times S^T + w_{m_2} \times S^V$, where the scalars w_{m_1} and w_{m_2} weight the contribution of each modality.

For each category j , the score assigned to the m documents was sorted in descending order; thus for each category we had a different ranking for the m -documents. The C -rankings were combined (by means of round robin) to generate a final ranking for the m -documents. The top 1,000 documents according to the final ranking were submitted for evaluation. For the topics that do not have textual information (i.e. topics 25 to 50) we used the provided query images.

2.4 Submitted runs and results

Table 1 summarizes the five runs we submitted for the photographic retrieval task; while Table 2 shows the results obtained by these runs. We report the (average across topics of) mean average precision (**MAP**), cluster recall at 20 documents (**C20**), R-precision (**RP**), precision at 20 documents and the ratio of relevant retrieved / relevant documents (**RR/R**). The parameters involved in our method were set empirically, analyzing the results of each configuration, Table 2 shows the parameter settings used for each run.

ID	Par.	Description
R-1	-	Textual retrieval method, see Section 2.2.
R-2	$w_{m_1} = 4$; $w_{m_2} = 1$	The m -documents obtained by the initial retrieval method are re-ranked according to the score in Equation (1); although the topic was not separated into categories under this formulation.
R-3	$\lambda = 0.25$	Re-ranking technique with $X = V$, see Section 2.3.
R-4	$\lambda = 0.25$	Re-ranking technique with $X = T$, see Section 2.3.
R-5	$\lambda = 0.25$; $w_{m_1} = 4$; $w_{m_2} = 1$	Re-ranking technique with $X = M$, see Section 2.3.

Table 1: Runs submitted by our team to the photographic retrieval track.

ID	MAP	C-20	RP	P20	RR/R
R-1	0.2901	0.4299	0.3411	0.5550	263.20/697.74
R-2	0.2723	0.4737	0.3300	0.5540	262.92/697.74
R-3	0.2710	0.5787	0.3298	0.5580	262.92/697.74
R-4	0.2706	0.4855	0.3274	0.5340	262.92/697.74
R-5	0.2645	0.5534	0.3242	0.5690	262.92/697.74

Table 2: Official retrieval results obtained with the submitted runs.

The obtained results are mixed. The best retrieval performance, in terms of **MAP** and **RP**, was obtained with the baseline (i.e. a textual retrieval technique); **P20** was higher with the R-5 configuration. Nevertheless, the difference in retrieval performance between the baseline and the other runs was of less than 0.03.

In terms of results diversification (i.e. **C20**), it is observed an improvement over the baseline for all of the runs (rows 3-6 in Table 2). The largest improvement in **C20** was obtained with the methods that considered visual information (i.e. R-3 and R-5). The run R-3, which used only visual information for re-ranking documents, resulted particularly helpful for diversifying retrieval results (this run was ranked 67 out of 84). These results suggest that the re-ranking technique

can be helpful for diversifying results and that using different modalities for the initial search and the re-ranking technique results in better performance. Note that the performance of the re-ranking technique depends on the initial retrieval, thus we expect better diversification of results when better retrieval methods are considered for the initial search. We are studying this research direction. Retrieval performance is slightly affected by applying the re-ranking method. However, the performance of the initial search method was rather limited: this method was ranked 47 out of the 85 submitted runs.

2.5 Discussion

The results obtained by the TIA-INAOE team in the photographic retrieval task at ImageCLEF2009 may seem discouraging in a first instance. However, interesting findings can be drawn from our participation: the proposed re-ranking technique resulted helpful for result diversification, although it slightly affects the retrieval performance; better diversification performance was obtained when the re-ranking was based on visual information only; better diversification performance is expected if a better search engine is used for the initial retrieval; as whole, the proposed formulation can be helpful for efficient multimedia retrieval in large scale image collections.

3 Large scale image annotation

We proposed a three steps methodology for facing the annotation task at ImageCLEF2009. For each test image, we identified a subset of *candidate labels* by comparing the test image to the training ones. Then, we selected the disjoint labels for the image by means of an energy-based model. Next, optional labels were assigned by taking into account co-occurrence statistics.

The method described in this section is based in the assumption that similar images have associated similar labels. Thus, for each test image we considered the labels assigned to the K -most similar training images; then, we applied different strategies for selecting disjoint and optional labels for a test image. The benefits of adopting this methodology are annotation efficiency, implementation simplicity and the competitive performance that can be obtained with the proposed formulation. The rest of this section describes our methodology and the obtained results. Further details on the task and on the collection are described by Nowak et al. [4].

3.1 Feature extraction

We used global features to represent images. In particular, we considered an RGB color histogram (256 bins), texture features extracted from the co-occurrence matrix (88 values), an edge histogram (360 bins) and an HSI color histogram with 128 bins per channel for a total of 1,088 attributes. Each image was represented by its vector of features; thus, hereafter, we will refer as images to both the images themselves and the vectors of features representing the images. For comparing images we used a weighted Euclidean distance, where a different weight is used for each subset of features (i.e. RGB, texture, edge, HSI). The weights were set empirically by trial and error.

3.2 KNN for image annotation

The first step in our methodology (depicted in Figure 2) is to obtain the k -most similar training-images to each test image I^T , we denote this set of images by I_{NN}^T . We are interested in the labels associated with the images in I_{NN}^T , we denote the corresponding set of labels by L_{NN}^T and we call it the set of candidate labels for I^T . We call the positions of label l_i^T to the set of positions, in the sorted set I_{NN}^T , occupied by images that have l_i^T as annotation. Then, we assigned a score to each label $l_i^T \in L_{NN}^T$ as follows:

$$\Psi(l_i^T) = \alpha_r \times w_r(l_i^T, L_{NN}^T) + \alpha_a \times w_a(l_i^T, L_{NN}^T) + \alpha_s \times w_s(l_i^T, L_{NN}^T) \quad (2)$$

We defined an energy-based model for the disjoint categories (i.e. Seasons, Place, TimeOfDay, Illumination, Blurring and Persons) using a variable a_j per category, with $j \in \{1, \dots, 6\}$. Each random variable can take values from its corresponding set of possible labels (e.g. the variable corresponding to the category TimeOfDay, can take the values: Day, Night or No-Visual-Time); we denote the assignment of label l_x to variable a_j with a_j^x , thus a_j^x can be considered a label itself. Additionally, we restricted the values that each variable can take, by considering as possible values only to those labels that appear in the set of candidate labels (i.e. L_{NN}^T). Figure 3 depicts the modeling process for a particular test image.

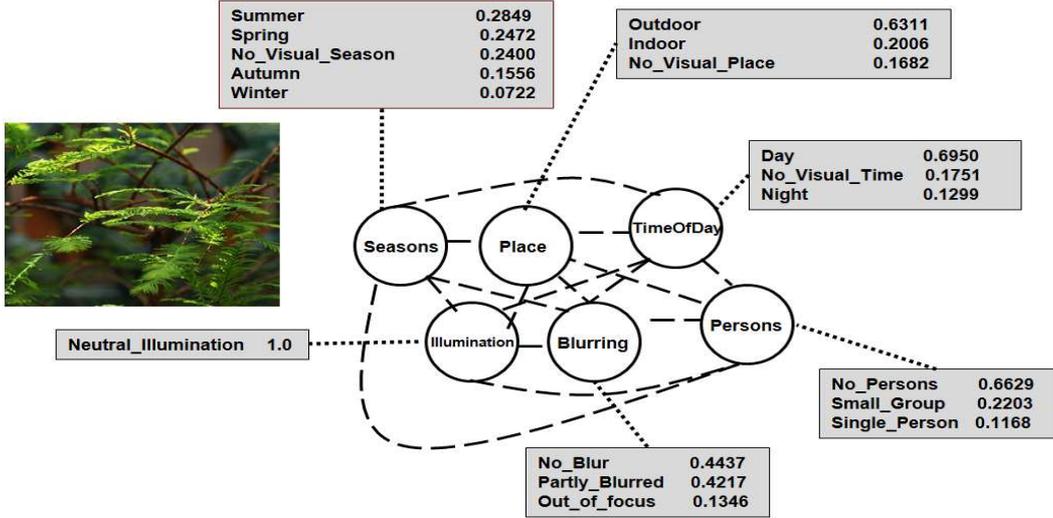


Figure 3: Illustration of the proposed energy-based model. Each variable (node) represents a disjoint category. A variable can take values labels from the corresponding category only (shaded box). The estimate from Equation (2) is normalized and used as input for the model. We considered a fully connected model.

The goal of the model is to select the configuration of labels A (i.e. a label assignment per category) that maximizes the cohesion among the labels assigned to the image. Accordingly, we assigned an energy value to each configuration of labels as follows:

$$E(A) = - \left(\sum_{a_j \in A} \delta \times \Psi(a_j^x) + \times \sum_{a_j \in A} \sum_{a_h \in \eta_{a_j}} \rho(a_j^x, a_h^y) \right) \quad (3)$$

Where $\Psi(a_j^x)$ as in Equation (2) and $\rho(a_j^x, a_h^y)$ is a factor that weights the association between labels l_x and l_y , assigned to categories a_j and a_h , respectively; η_{a_j} is the set of neighbors², under the model, of category a_j , as we are using a fully complete graph the set of neighbors for a_j is $\eta_{a_j} = a_p \in A : p \neq j$. δ weights the contribution of the initial ranking to the energy of the configuration. We used co-occurrence statistics to estimate the association between labels. Specifically, we estimate $\rho(a_j^x, a_h^y)$ as follows:

$$\rho(a_j^x, a_h^y) = \frac{\#(l_x, l_y)}{\#(l_x)\#(l_y)} \quad (4)$$

where $\#(l_x)$ is the number of images in the training set in which label l_x occurs and $\#(l_x, l_y)$ is the number of images in which both l_x and l_y co-occur. For specific labels l_x and l_y , the higher $\rho(a_j^x, a_h^y)$ the more both labels are associated. Note that $\rho(a_j^x, a_h^y)$ can be calculated for any pair of labels (disjoint and optional); thus we use this association information in the next section for

²Note that these neighbors are different from the neighbors considered in Section 3.2.

selecting optional labels as well. Equation (3) assigns low energy values to correct configurations and large values to incorrect ones. Therefore, the problem of selecting disjoint labels for a test image is reduced to that of finding the configuration of labels that minimizes Equation (3); for this work we used iterated conditioned modes (ICM) for this task [8].

The energy-based model returns labels for each disjoint category. Intuitively, the model selects the combination of labels that maximizes their semantic cohesion. This method is based on the method proposed in [2] for region-labeling; in this paper we extend such a model to work for image-level annotation.

3.4 Annotating images

Once we selected the disjoint labels we assigned optional labels to each test image as follows. We assigned a score to each candidate label l_i^T (that does not belong to any disjoint category), identified in Section 3.2, as follows:

$$\zeta(l_i^T) = \Psi(l_i^T) \times \prod_{j=1}^6 \rho(l_i^T, l_{d_j}) \quad (5)$$

where $\Psi(l_i^T)$ and $\rho(l_i^T, l_{d_j})$ are defined as above and l_{d_j} is the label assigned to the j^{th} disjoint category. We ranked labels according to $\zeta(l_i^T)$ and used the top- n labels for labeling the test image.

3.5 Postprocessing

For generating the final annotation for a test image we applied the following postprocessing. First, regarding the number of labels, we assigned the top 4 optional labels to each test image, as 4 is the average number of optional labels that were used for annotating the training images. Second, when a leaf-label was chosen as optional label for an image, we also included its parent label, as appear in the annotation hierarchy defined by the organizers. Thus, for example, if the label Lake was considered as optional label for the image, we also included the label Water. Of course, this is only applicable to optional labels that appear in the hierarchy as leaves.

3.6 Submitted runs and results

Table 3 summarizes the five runs we submitted for the large scale annotation task; whereas Table 4 shows the results obtained by such runs. We show the following performance measures: the hierarchical measure described in [5], two variants are available: **H-A** is the hierarchical performance with annotator agreement, whereas **H-WA** is the performance without annotator agreement, the higher the values of **H-A** and **H-WA** the better the annotation performance, see [4, 5] for further details. Also, the average of equal error rate (**EER**) and the area under the ROC curve (**AUC**) were considered for evaluation. The parameters involved in our method were set by using cross validation using the training set of images.

From Table 4 we can see that the baseline method is a rather strong baseline, confirming the results reported by Makadia et al. [3]. The performance of the baseline was not improved by applying the energy-based model. This can be due to the fact that we did not use a good set of parameters for the model. In cross validation experiments we obtained better performance in both **EER** and **AUC**, thus it seems we overfitted the data.

The worst performance, in terms of **H-A** and **H-WA**, was obtained when local information was used for ranking labels according to the KNN approach. Hence these local features were not helpful for re-ranking. An interesting result, however, is that when the energy-based model was applied with the labels ranked according to local features, the energy-based model was able to improve the performance of the former significantly (compare the performances of runs A-2 and A-4). This result suggest the energy-based model can be helpful when the initial labeling is not good.

ID	Description
A-1	KNN; the score in Equation (2) is used for assigning labels.
A-2	KNN-RR; we use the alternative re-ranking with local features (see Section 3.2.1) and the score in Equation (2) for assigning labels.
A-3	KNN + EBM; candidate labels selected as in A-1, the energy-based model is used to select the disjoint labels and optional labels are selected as described in Sections 3.3, 3.4 and 3.5
A-4	KNN-RR + EBM; candidate labels selected as in A-2, the energy-based model is used to select the disjoint labels and optional labels are selected as described in Sections 3.3, 3.4 and 3.5
A-5	KNN-RRW + EBM; same as A-4, but we use a larger δ value.

Table 3: Runs submitted by our team to the large scale annotation track.

ID	H-A	H-WA	EER	AUC	Time (s)
A-1	0.7592	0.7317	0.4862	0.1008	0.15
A-2	0.5329	0.5125	0.4847	0.0993	0.24
A-3	0.7281	0.6966	0.4929	0.0442	0.25
A-4	0.7323	0.7018	0.4924	0.0622	0.26
A-5	0.7418	0.7127	0.4872	0.0947	0.23

Table 4: Official annotation results obtained with the submitted runs.

It is interesting to note that whereas the performance of our runs in **H-A** and **H-WA** was, to some extent, satisfactory, the performance in terms of **EER** and **AUC** was rather limited. Our best run (A-1) was ranked 26 out of 74 submitted runs in terms of **H-A** and **H-WA**. However, the same run was ranked 59 and 66 out of 74 in terms of **EER** and **AUC**, respectively. This result suggest that the method can label images as whole satisfactorily, although its per-label performance is limited; which is not surprising as we have not developed visual concept detectors per class. Note that the main goal of assigning labels to images is to support annotation based image retrieval methods, which use labels assigned to the images as a whole. Thus, it seems that our method could support effectively this form of image retrieval, we will study this aspect as future work, currently we are conducting an more in-deep analysis of the results.

Finally, the processing time³ of our methods is quite acceptable, this time could be further reduced if we use software that is less computationally expensive (we used Matlab^R for all of our experiments) and if we optimize our code.

3.7 Discussion

The results obtained by the TIA-INAOE team are encouraging. The KNN approach to image labeling proved to be a very effective method for image annotation, regardless of its simplicity and generality. Despite the proposed energy-based model did not improve the performance of the KNN method, it was able to improve significantly the performance of the KNN-RR method. Thus suggesting the energy-based model can be helpful when the initial method is not very effective; this is a desired behavior of the model. Note that the energy-based model is still under development and that we have fixed the number of labels that are assigned to an image, also the parameter selection process can be improved. In general terms, our annotation methodology offers a good tradeoff in terms of annotation performance (in **H-A** and **H-WA**) and processing time. We would like to emphasize that the proposed approach for labeling refinement is not restricted to our KNN annotation method. It can be applied as a postprocessing step with any annotation method,

³The reported time does not include feature extraction from images.

provided that the labels can be ranked, hence, showing the generality of the method and the potential impact it can have. Summarizing, the energy-based model is intuitively sound and is a promising method in which we are still working on.

4 Conclusions

We have described the participation of TIA-INAOE at ImageCLEF2009. Our team submitted runs for the Photographic Retrieval and for the Large Scale Annotation tracks. We proposed specific methods for each track. On the one hand, we described a re-ranking approach that aims at maximizing the diversity of retrieved documents at the first positions. Our results show that whereas the proposed technique can improve the diversity of results, the base retrieval system still needs to be improved. On the other hand, we adopted a simple method for image annotation, and introduced a labeling refinement technique with the goal of improving the annotations as obtained with the former method. Our results suggest the KNN approach is effective for annotation and very efficient. However, there are still several issues with our refinement method that we are currently working on.

Acknowledgements. The authors thank the organizers of *ImageCLEF2009* because of their support. This work was partially supported by CONACyT under project grant 61335 and scholarship 205834.

References

- [1] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Pearson E. L., 1999.
- [2] H. J. Escalante, M. Montes, and E. Sucar. Maximizing the semantic cohesion for region labeling. *Submitted to International Journal of Computer Vision*, 2009.
- [3] A. Makadia, V. Pavlovi, and S. Kumar. A new baseline for image annotation. In *ECCV'08: Proceedings of the 10th European Conference on Computer Vision*, volume 5304 of *LNCS*, pages 316–329, Marseille, France, 2008. Springer.
- [4] S. Nowak and P. Dunker. Overview of the clef 2009 large scale - visual concept detection and annotation task. In F. Borri, A. Nardi, and C. Peters, editors, *CLEF working notes*, Corfu, Greece, October 2009.
- [5] S. Nowak and H. Lukashevich. Multilabel classification evaluation using ontology information. In *Proceedings of the First ESWC Workshop on Inductive Reasoning and Machine Learning on the Semantic Web*, volume 474 of *CEUR Workshop Proceedings*, Heraklion, Greece, 2009.
- [6] M. Paramita, M. Sanderson, and P. Clough. Diversity in photo retrieval: Overview of the imageclef-photo task 2009. In F. Borri, A. Nardi, and C. Peters, editors, *CLEF working notes*, Corfu, Greece, October 2009.
- [7] T. Serre, M. Kouh, C. Cadieu, U. Knoblich, G. Kreiman, and T. Poggio. A theory of object recognition: computations and circuits in the feedforward path of the ventral stream in primate visual cortex. Technical report, AI Memo #2005-036, Massachusetts Institute of Technology, Cambridge, MA, USA, Dec 2005.
- [8] G. Winkler. *Image Analysis, Random Fields and Markov Chain Monte Carlo Methods*. Number 27 in Applications of Mathematics. Springer, 2006.
- [9] D. Zeimpekis and E. Gallopoulos. Tmg: A matlab toolbox for generating term-document matrices from text collections. In C. Nicholas J. Kogan and M. Teboulle, editors, *Grouping Multidimensional Data: Recent Advances in Clustering*, pages 187–210. Springer, 2005.