MMIS at ImageCLEF 2009: Non-parametric Density Estimation Algorithms

Ainhoa Llorente, Suzanne Little and Stefan Rüger
Knowledge Media Institute
The Open University, Milton Keynes
MK7 6AA, United Kingdom
{a.llorente, s.little, s.rueger}@open.ac.uk

Abstract

This paper presents the work of the MMIS group done at ImageCLEF 2009. We submitted five different runs to the Photo Annotation task. These runs were based on two non-parametric density estimation models. The first one evaluates a set of visual features and proposes a better, weighted set of features. The second approach uses keyword correlation to compute semantic similarity measures using several knowledge sources. The knowledge sources used are, the training set of the collection, Google Web search engine, WordNet and Wikipedia. Evaluation of results is done under two different metrics, one based on ROC curves and the other in a hierarchical measure proposed by the organisers. Our results are quite encouraging; under the first metric our best run was located between the median and the top quartile and under the second metric our best run was between the first quartile and the median.

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; I.4 [**Image Processing and Computer Vision**]: I.4.8 Scene Analysis; I.4.9 Applications

General Terms

Algorithms, Experimentation, Performance, Measurement

Keywords

Content Based Image Retrieval, Object Recognition, Thesauruses

1 Introduction

In this paper, we describe the experiments performed by the MMIS group at ImageCLEF 2009. We participated in the Large Scale Visual Concept Detection and Annotation Task. The main goal of this task is, as described in [13], given a training set of 5,000 images manually annotated with words coming from a vocabulary of 53 visual concepts, to automatically provide annotations for a test set of 13,000 images. The visual concepts are organized in a small ontology so participants may take advantage of the hierarchical order of the concepts and the relations among them for better accomplishing the annotation task. Another important goal of this year competition is to reflect about the influence of large amount of data and concepts in the annotation task and about whether or not an ontology can help.

We submitted five runs in total. Each one of them is based on a different non-parametric density estimation model but placing emphasis on different aspect of the research field. For instance, the run MMIS_33_2_1245434554581.txt is evaluating a sequence of possible image feature selections in order to propose a better, weighted set of features while the other four runs MMIS_33_2_1245586552541.txt, MMIS_33_2_1245601239738.txt, MMIS_33_2_1245611281967.txt, and, MMIS_33_2_1245674693001.txt attempt to improve a baseline probabilistic model taking advantage of the correlation between keywords computing semantic similarities measures using different knowledge bases.

Evaluation of results have been done under two different metrics, one is based on ROC curves [4] and proposes as measures Equal Error Rate (EER) and the Area under Curve (AUC) while the second metric is the hierarchical measure proposed by [14] that considers the relations between concepts and the agreement of annotators on concepts. All in all, our results are quite encouraging, under the first metric our best run was located between the median and the top quartile and under the second metric our best run was between the first quartile and the median.

The rest of this paper is organised as follows. Section 2 provides an introduction on non-parametric density estimation. Section 3 describes the first approach followed while Section 4 illustrates the second one. Then, our evaluation results are discussed in Section 5. Finally, Section 6 shows our conclusions.

2 Non-parametric Density Estimation

Both approaches followed in this research are variations of the probabilistic framework developed by Yavlinsky et al. [16] who used global features together with a non-parametric density estimation. This approach is based on the Bayes rule being the ultimate goal to model $f(x|\omega)$ for each annotation keyword ω, being x a feature vector representing a test image. The non-parametric approach is employed because the distributions of image features have irregular shapes that do not resemble *a priori* any simple parametric form.

The function $f(x|\omega)$ is estimated following a kernel k based approach as represented in:

$$f(x|\omega) = \frac{1}{nC} \sum_{i=1}^{n} k \frac{x - x_{\omega}^{(i)}}{h}, \tag{1}$$

where $x_{\omega}^{(1)}$, $x_{\omega}^{(2)}$,..., $x_{\omega}^{(n)}$, is a sample of feature vectors from the training set labelled with the keyword ω and $x = (x_1, ..., x_d)$ is a vector of real-valued image features.

The approach explained in Section 3 places a d-dimensional Laplacian kernel over point $x^{(i)}$ as expressed in:

$$k_L(t; h) = \prod_{l=1}^{d} \frac{1}{h_l} \, e^{-\frac{t_l}{h_l}}, \tag{2}$$

while the approach described in Section 4 uses a Gaussian kernel as shown in:

$$k_G(t; h) = \prod_{l=1}^{d} \frac{1}{\sqrt{2\pi} h_l} e^{-\frac{1}{2}(\frac{t_l}{h_l})^2}, \tag{3}$$

where $t = x - x^{(i)}$ and h_l is the bandwidth of the kernel which is set by scaling the sample standard deviation of feature component l by the same constant λ.

2.1 Image Features

A key aspect of the non-parametric density estimation approach is the global visual features used. The algorithm described in Section 3 used four features: CIELAB and HSV colour descriptors

combined with Tamura and Gabor texture descriptors while our second algorithm 4 combines the CIELAB color feature with the Tamura texture.

CIE L*a*b* (CIELAB) [7] is the most complete colour space specified by the International Commission on Illumination (CIE). Its three coordinates represent the lightness of the colour (L*), its position between red/magenta and green (a*) and its position between yellow and blue (b*). The histogram was calculated over two bins for each coordinate.

HSV is a cylindrical colour space with H (hue) being the angular, S (saturation) the radial and V (brightness) the height component. The H, S and V axes are subdivided linearly (rather than by geometric volume) into two bins each. The HSV colour histogram is normalised so that this components add up to one.

The Tamura texture feature [15] is computed using three main texture features called "contrast", "coarseness", and "directionality". Contrast aims to capture the dynamic range of grey levels in an image. Coarseness has a direct relationship to scale and repetition rates and it was considered by Tamura et al. as the most fundamental texture feature and finally, directionality is a global property over a region. The histogram was calculated over two bins for each feature.

The process for extracting each of these features is as follows, each image is divided into nine equal rectangular tiles, the mean and second central moment feature per channel are calculated in each tile. The resulting feature vector is obtained after concatenating all the vectors extracted in each tile.

The final feature extracted is a texture descriptor produced by applying a Gabor filter to enable filtering in the frequency and spatial domain. Our implementation is based on [10]. To each image we applied a bank of four orientation and six scale sensitive filters that map each image point to a point in the frequency domain. This feature was calculated on the whole image rather than using the tiling approach.

3 Weighted Global Features

The original implementation of this algorithm used two features: CIELAB and Tamura. Subsequent work evaluated a sequence of possible feature selections [8] and proposed a better, weighted set of features. The feature sets to be evaluated were constructed based on information from existing literature about visual feature selection and attempted to avoid descreasing performance due to redundant features or multi-variate prediction.

The feature set proposed from this set of evaluations added two additional features, HSV colour and Gabor texture, to the original CIELAB and Tamura descriptors. These features were weighted at CIELAB - 0.75, HSV - 0.5, Tamura - 0.5 and Gabor - 0.5. This set improved the mean average precision when evaluated on the standard Corel5k dataset [3] and the IAPR TC12 dataset used for ImageCLEF 2006 [6].

The run is labelled **MMIS_33_2_1245434554581.txt** and is based on the approach used in [16]. The four chosen features were extracted from the training set to train the non-parametric density estimation annotator which then provided the probability of each concept being present in the test image. Manhattan distance was used for all features.

This algorithm represented a straight-forward approach that exploited only the global low-level features and the supervised learning of a prediction model. We predicted that this set of features would provide a good coverage of the colour and texture space and sufficient details without placing an excessive calculation burden on the system. Initial tests using ten-fold cross-validation on the training set re-enforced this expectation.

4 Exploiting Word Correlations to Compute Semantic Similarities

The early attempts in automated image annotation were focused on algorithms that explored the correlation between words and image features. More recently, there are some efforts which attempt

to benefit from exploiting the correlation between words computing semantic similarity measures. Among the many uses of the concept "semantic similarity", we refer to the definition by Miller and Charles [11] who consider it as the degree of contextual interchangeability or the degree to which one word can be replaced by another in a certain context. Consequently, two words are similar if they refer to entities that are likely to co-occur together like "mountains" and "vegetation", "beach" and "water", "buildings" and "road", etc. In this research we will use indistinctly the term semantic similarity and semantic relatedness.

This non-parametric density estimation model exploits the statistical correlation between words by computing semantic similarity measures using different knowledge bases. We propose four versions of this model that differ on the knowledge base used as source of information and the semantic similarity measure employed. The knowledge bases used are, the training set of the collection, Google Web search engine, WordNet, and Wikipedia. The semantic similarity measures used are explained in Section 4.2.

The process can be described as follows. We calculate the probability value of each concept being present in each image of the test set following the non-parametric density estimation described in Section 2. Then, a statistical keyword correlation is computed using the corresponding knowledge base. With the help of the semantic similarity measures and applying some rules the accuracy of the final annotations is improved.

4.1 Parameter Estimation

We divided the dataset into three parts: a training set, a validation set and a test set. The validation test is used to find the parameters of the model. Thus, we performed a 10-fold cross validation on the training set. After that, the training and validation set are merged to form a new training set of 5,000 images that is used to predict the annotations in the test set of 13,000 images.

4.2 Submitted Runs

In this subsection, we describe the four submitted runs based on this approach:

MMIS_33_2_1245586552541.txt This run is based on the approach developed in [9] where the training set is computed to generate a co-occurrence matrix that represents the probabilities of the frequency of two vocabulary words appearing together in a given image. This algorithm was previously tested on the Corel5k collection and in the collection provided by the last edition of ImageCLEF, in 2008.

MMIS_33_2_1245611281967.txt The semantic similarity measure used in this run is called *web-based semantic relatedness measure* as it uses Google Web search engine as knowledge base. It was developed by Gracia and Mena [5] who defined the semantic relatedness between the concepts x and y, as:

$$\mathrm{rel}(x, y) = e^{-2\,\mathrm{NWD}(x,y)}, \tag{4}$$

where NWD stands for Normalized Web Distance which is a generalisation of the Normalized Google Distance (see Equation 5) extended to any web-based search engine as source of frequencies. The Normalized Google Distance (NGD) between two terms x and y, was expressed by Cilibrasi and Vitányi [2] as:

$$\mathrm{NGD}(x, y) = \frac{\max\{\log f(x), \log f(y)\} - \log f(x, y)}{\log N - \min\{\log f(x), \log f(y)\}}, \tag{5}$$

where $f(x)$ and $f(y)$ are the counts for search terms x and y using Google and $f(x, y)$ is, the number of web pages found on which both x and y occur. N is the total number of web pages searched by Google which, in 2007, was estimated to be more than 8bn pages.

MMIS_33_2_1245674693001.txt This run uses the adapted Lesk measure applied to WordNet proposed by Banerjee and Pedersen in [1]. They defined the *extended gloss overlap measure* which computes the relatedness between two synsets c_1 and c_2 by comparing the glosses of synsets related to them through explicit relations provided by WordNet:

$$\text{rel}(c_1, c_2) = \sum score(R_1(c_1), R_2(c_2)),\ \forall (R_1, R_2) \in relPairs. \tag{6}$$

Thus, the set $relPairs$ is defined as follows:

$$relPairs = \{(R_1, R_2) \mid R_1, R_2 \in rels;\ if(R_1, R_2) \in relPairs,\ then(R_1, R_2) \in relPairs\}, \tag{7}$$

being $rels$ a non-empty set of relations that consists of one or more of the following relations:

$$rels \subset \{r \mid r\ is\ a\ relation\ defined\ in\ WordNet\}. \tag{8}$$

MMIS_33_2_1245601239738.txt This run computes the semantic relatedness between two concepts applying the Wikipedia measure defined by Milne and Witten. In [12], they proposed their *Wikipedia Link-based Measure* (WLM) which extracts semantic relatedness measure between two concepts using the hyperlink structure of Wikipedia. The semantic relatedness between concepts x and y is estimated by the angle between the vectors of the links found between the Wikipedia articles whose title matches each one of the concepts:

$$\text{rel}(x, y) = \frac{\vec{x} \cdot \vec{y}}{|\vec{x}| \cdot |\vec{y}|}, \tag{9}$$

where the vectors for article x and y are built using link counts weighted by the probability of each link occurring, as seen in:

$$\vec{x} = (w(x \to l_1), w(x \to l_2), ..., w(x \to l_n)), \tag{10}$$

and, in:

$$\vec{y} = (w(y \to l_1), w(y \to l_2), ..., w(y \to l_n)). \tag{11}$$

Thus, the weighted value w for the link $a \to b$ can be defined as:

$$w(a \to b) = |a \to b|\ \cdot log\left(\sum_{x=l}^{t} \frac{t}{|x \to b|}\right), \tag{12}$$

being t is the total number of articles within Wikipedia.

5 Evaluation Measures and Results

We used two metrics to determine the quality of the annotations. The first metric is based on ROC curves [4]. Initially, a receiver operating characteristic (ROC) curve was used in signal detection theory to plot the sensitivity versus (1 - specificity) for a binary classifier as its discrimination threshold is varied. Later on, ROC curves were applied to information retrieval in order to represent the fraction of true positives (TP) against the fraction of false positives (FP) in a binary classifier. The Equal Error Rate (EER) is the error rate at the threshold where FP=FN. The area under the ROC curve, AUC, is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one. Note that, the lower the EER, the better the annotations.

In Table 1, we show the results for all our submitted runs under the EER and AUC metric. Our best run corresponds to MMIS_33_2_1245434554581.txt that follows the first approach of weighted global features. This run achieved a reasonable EER across all concepts of just over 31% which was consistent with the predicted performance from our earlier ten-fold cross validation. Table 2

Table 1: EER and AUC results for all the runs of MMIS group

Algorithm	**EER**	**AUC**
Random	0.500280	0.499307
MMIS_33_2_1245434554581.txt	0.312366	0.744231
MMIS_33_2_1245586552541.txt	0.352478	0.689410
MMIS_33_2_1245601239738.txt	0.356945	0.684821
MMIS_33_2_1245611281967.txt	0.352485	0.689407
MMIS_33_2_1245674693001.txt	0.352612	0.689342

Table 2: Top 10 best concepts best MMIS run

Concept	**EER**	**AUC**
Sunset-Sunrise	0.181372	0.889386
Clouds	0.192238	0.875234
Underexposed	0.207616	0.871133
Sky	0.211369	0.865876
Night	0.212004	0.860332
Sea	0.213746	0.854027
Mountains	0.217231	0.846818
Landscape-Nature	0.225471	0.84186
Desert	0.232747	0.810619
Food	0.251319	0.828054

shows that the ten best concepts identified by this annotator are those which have previously performed well using only global visual features. With the exception of "Underexposed", the best performing concepts belong to fairly common visual categories, primarily landscape elements.

Regarding the four runs based on keyword correlation, we observe that the best performance is achieved using the training set as corpora. Not surprisingly, the second best is the run based on Google Normalized Distance that uses Google Web search engine as knowledge source. This is due to the fact that both approaches do not reply on a prior disambiguation process like WordNet and Wikipedia.

The worst result corresponds to the run based on Wikipedia. The reasons behind it might be found in the strong dependency of the semantic relatedness measure on doing a proper word disambiguation. The disambiguation in Wikipedia is automatically performed by selecting the sense of the word more probable according to the content store on Wikipedia database.

The 53 concepts of the proposed vocabulary belong to one of the following categories: Scene description, Seasons, Place, Landscape Elements, Time of the day, Picture representation, Illumination, Quality Blurring, Picture Objects, and Quality Aesthetics.

Most of the categories do not correspond to real visual features and the best way of predicting them is making use of the "exif" metadata. As our focus is on visual features we have not incorporated them in any of our algorithms. Consequently, we predicted and posteriorly checked, lower results for concepts classified into categories such as Seasons, Time_of_the_day, Picture representation, Illumination, Quality Blurring and specially, the most subjective one, Quality Aesthetics.

The second metric is the proposed hierarchical measure [14] that considers the relations between concepts and the agreement of annotators on concepts. In Table 3, the results of all our submitted runs are shown. The best run is the one based on Google Web search engine followed by the co-occurrence, and WordNet approaches. This makes sense as all these runs are employing semantic similarity measures on external ontologies, which is exactly the criteria that the hierarchical score attempts to evaluate. The run which applied weighted global features relied less on the hierarchical

Table 3: Average Annotation Score for all the runs of MMIS group

Algorithm	**With Annotator**	**Without Annotator**
Random	0.3843171	0.35097164
MMIS_33_2_1245434554581.txt	0.5479666	0.49800622
MMIS_33_2_1245586552541.txt	0.6179764	0.57577974
MMIS_33_2_1245601239738.txt	0.4205571	0.35027474
MMIS_33_2_1245611281967.txt	0.6180272	0.57583610
MMIS_33_2_1245674693001.txt	0.6172693	0.57497290

information and therefore did not perform as well using the metric.

6 Conclusions

While it is difficult to make conclusive statements about the submitted runs as the differences in their performance are minimal, the results do re-enforce previous expectations.

The performance of our best run (according to EER) supports previous findings about the impact of feature selection and weighting on the non-parametric density algorithm as it outperforms the other four runs which used the original features.

With respect to the second metric (hierarchical measure), the best runs are those that use as knowledge base the training set and Google Web search engine because the rest of the approaches (WordNet and Wikipedia) have been penalised as a result of the prior disambiguation process that follow.

Interestingly the metric to distinguish the performance of annotators based on measurement of the hierarchical distribution isolates the feature weighting run. This alternative method of ranking performance gives valuable insight into the influence and impact of the analysis of hierarchical labels in image annotation. It is likely that annotators that achieve a higher ranking using the hierarchical measure have better distribution across the concepts. Further analysis is needed to determine if annotators with a better hierarchical measure are also more robust overall.

Acknowledgments.

This work was partially funded by the EU Pharos project (IST-FP6-45035) and by Santander Corporation.

References

[1] S. Banerjee and T. Pedersen. Extended gloss overlaps as a measure of semantic relatedness. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence*, 2003.

[2] Rudi Cilibrasi and Paul Vitanyi. The Google similarity distance. *IEEE Transactions on Knowledge and Data Engineering*, 19(3):370–383, 2007.

[3] P. Duygulu, Kobus Barnard, J. F. G. de Freitas, and David A. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Proceedings of the European Conference on Computer Vision*, pages 97–112, 2002.

[4] Tom Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006.

[5] Jorge Gracia and Eduardo Mena. Web-based measure of semantic relatedness. In *Proceedings of 9th International Conference on Web Information Systems Engineering*, volume 5175, pages 136–150, 2008.

[6] Michael Grubinger. *Analysis and Evaluation of Visual Information Systems Performance*. PhD thesis, School of Computer Science and Mathematics, Faculty of Health, Engineering and Science, Victoria University, Melbourne, Australia, 2007.

[7] A Hanbury and J Serra. Mathematical morphology in the CIELAB space. *Image Analysis & Stereology*, 21:201–206, 2002.

[8] Suzanne Little and Stefan Rueger. Conservation of effort in feature selection for image annotation. In *IEEE Workshop on Multimedia Signals Processing (MMSP2009)*, Rio De Janeiro, Brazil, October 5–7 2009.

[9] Ainhoa Llorente and Stefan Rüger. Using second order statistics to enhance automated image annotation. In *Proceedings of the 31st European Conference on Information Retrieval*, volume 5478, pages 570–577, 2009.

[10] B. Manjunath and W Ma. Texture features for browsing and retireval of image data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18:837–842, 1996.

[11] George A. Miller and Walter G. Charles. Contextual correlates of semantic similarity. *Journal of Language and Cognitive Processes*, 6:1–28, 1991.

[12] D. Milne and I.H. Witten. An effective, low-cost measure of semantic relatedness obtained from wikipedia links. In *Proceedings of the first AAAI Workshop on Wikipedia and Artifical Intellegence*, 2008.

[13] Stefanie Nowak and Peter Dunker. Overview of the CLEF 2009 Large Scale -Visual Concept Detection and Annotation Task. In *Cross-Language Evaluation Forum Working Notes*, Corfu, Greece, 2009.

[14] Stefanie Nowak and Hanna Lukashevich. Multilabel classification evaluation using ontology information. In *Proceedings of ESWC Workshop on Inductive Reasoning and Machine Learning on the Semantic Web*, 2009.

[15] H. Tamura, T. Mori, and T. Yamawaki. Textural features corresponding to visual perception. *IEEE Transactions on Systems, Man and Cybernetics*, 8(6):460–473, 1978.

[16] Alexei Yavlinsky, Edward Schofield, and Stefan Rüger. Automated image annotation using global features and robust nonparametric density estimation. In *Proceedings of the International ACM Conference on Image and Video Retrieval*, pages 507–517, 2005.