Performing image classification with a frequency–based information retrieval schema for ImageCLEF 2006

Henning Müller, Tobias Gass, Antoine Geissbuhler
University and Hospitals of Geneva, Switzerland
henning.mueller@sim.hcuge.ch

Abstract

This article describes the participation of the University and Hospitals of Geneva at the ImageCLEF 2006 image classification tasks (medical and non–medical). The techniques applied are based on classical tf/idf weightings of visual features as used in the GIFT (GNU Image Finding Tool) image retrieval engine. Based on the training data, features that appear in images of the same class are weighted higher than features appearing across many classes. These feature weights are added to the classical ft/idf weights, making it a mixture of weightings. Several weightings and learning approaches are applied as well as several quantisations of the features space with respect to grey levels. A surprisingly small number of grey levels leads to best results. Learning can improve the results only slightly and does not obtain as good results as classical image classification approaches. A combination of several classifiers leads to best final results, showing that the applied schemes have independent results. For future work it seems important to study in more detail the important features and feature groups as they are not independent in the GIFT system. Pre–treating of the images (background removal) or allowing for more variation of the images with respect to object size and position might be other approaches to further improve results.

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 Managment**]: Languages—*Query Languages*

General Terms

Measurement, Performance, Experimentation

Keywords

Image retrieval, Image classification, frequency–based weights

1 Introduction

ImageCLEF[1] makes available realistic test collections for the evaluation of retrieval and classification tasks in the context of CLEF[2] (Cross Language Evaluation Forum). A detailed description

[1] http://ir.shef.ac.uk/imageclef/
[2] http://www.clef-campaign.org/

of the object annotation task and a photographic retrieval task can be found in [1]. The overview includes a description of the tasks, the submitted results and a ranking of the best systems. A description of a medical image retrieval and automatic image annotation task can be found in [4] with all the details of submissions. More on the data can also be found on [3].

This article will concentrate on the submission of the University and Hospitals of Geneva for the two image classification tasks. The submissions were not in time for the official evaluation because of a lack of man power but can be compared with these results in the overview articles. Already in 2005, an automatic medical image annotation task top was offered in ImageCLEF [2]. Best results were obtained by systems using classical image classification techniques [3]. Approaches based on information retrieval techniques [5] had lower results but were still among the best five groups, and this without using any learning data. It was expected that a proper use of learning data could improve results significantly, although tf/idf weighting already take into account the distribution of features in the collection. Such a learning approach is attempted and described in this paper.

2 Methods

The methods described in this paper rely heavily on those used in the GIFT[4] (GNU Image Finding Tool) [7]. The learning approaches applied are based on learning algorithms published in [6] using the idea to translate the market basket analysis problem to image retrieval.

2.1 Features used

GIFT itself uses four different groups of image features, which are described in more detail in [7].

- A global color histogram which is based on the HSV color space and quantised into 18 hues, 3 saturations, 3 values and usually 4 levels of grey.
- Local color blocks. Each image is recursively partitioned into 4 blocks of equal size, and each block is represented by its mode color.
- A global texture histogram of the responses to Gabor filters of 3 different scales and 4 directions, which are quantised into 10 bins with the lowest one usually being discarded.
- Local Gabor block features by applying the filters mentioned above to the smallest blocks created by the recursive partition and using the same quantisation into bins.

This results in 84'362 possible features where each image contains around 1'500. The Images in the IRMA database are not coloured and thus the number of features is reduced by roughly 50%. Because of this and as a color histogram is usually an effective feature, we decided to increase the color features by extracting not only four levels of grey, but also 8, 16 and 32 levels, resulting in a higher–dimensional space. Such changes in feature space have been used frequently in the medGIFT[5] project.

2.2 Feature weights

Several weighting schemes are implemented in GIFT. The basic one used in this paper is the *term frequency/inverted document frequency*(tf/idf) weighting which is well known from text retrieval (TR) literature. Given a query image q and a possible result image k, a score is calculated as the sum of all weights of features which are occurring in k.

$$score_{kq} = \sum_j (feature\ weight_j)$$

[3] http://ir.ohsu.edu/image/
[4] http://www.gnu.org/software/gift/
[5] http://www.sim.hcuge.ch/medgift/

The weight of each feature is computed by dividing the term frequency(tf) of the feature by the squared logarithm of the inverted collection frequency(cf).

$$feature\ weight_j = tj_j * log^2(1/(cf_j))$$

This results in giving features, which occur very frequently in the collection, a lower weight. These features do not discriminate images very well from each other. An example for such a feature would be black background being present in a very large number of medical images.

The strategy described above does not use much of the information contained in the training data, only the feature frequencies are exploited and not at all the class memberships of the images. For optimising the retrieval of relevant images, learning from user *relevance feedback* was presented in [6]. Here, we use the described weighting approaches and add with several learning strategies to optimise results for the classification task, where class membership of the entire training data is known.

2.2.1 Strategies

The former learning approach was to analyse log files of system use and find *pairs* of images that were marked together in the query process. Afterwards, frequencies can be computed of how often each feature occurs pairs of images. A weight can be calculated by using the information whether or not the images in the pair were both marked as *relevant* or whether one was marked *relevant* and the other as *notrelevant*. This results in desired and non–desired cooccurence of features.

In the approach described in this paper, we want to train weights in a scope more focused on classification. This means that we do not look at user interaction but rather get relevance data on class memberships of images by looking at the class labels of the training data. Each result image for a query is marked as relevant if the class matches that of the query image and non–relevant otherwise. This allows for a more focused weighting than what real users would do with relevance feedback.

We then applied several strategies for extracting the pairs of images for such queries. In the first approach, each possible pair of images which occurs at least once is considered relevant. This yields very good results for image retrieval in general as can be seen in [6]. In the second approach we aim at discriminating positive and negative results in a more direct way. To do this, only the best positive and the worst negative results (images) of a query are taken into account when computing pairs of marked images. In a third approach, we pruned all queries which seemed *too easy*. This means that if the first N results were already positive, we omitted the entire query from further evaluation. Everything else follows the basic approach. This is based on ideas similar to Support Vector Machines (SVM), where only information on the class boundaries is taken into account and all images that are in the middle of the class would be classified correctly anyways.

2.2.2 Computation of additional feature weights

For each image pair detected beforehand, we calculate the features they have in common and whether the image pair was positive (both images in the same class) or negative (images in different classes). This results in positive and negative cooccurence on a feature level. We used two ways to compute an additional weighting factor for the features:

- Basic Frequency : In this weighting scheme, each feature is weighted by the number of occurrences in pairs where both images are in the same class, normalised by the number of occurrences of the feature in all pairs.

  $$factor_j = \frac{|\{f_j | f_j \in I_a \wedge f_j \in I_b \wedge (I_a \rightarrow I_b)_+\}|}{|\{f_j | f_j \in I_a \wedge f_j \in I_b \wedge ((I_a \rightarrow I_b)_+ \vee (I_a \rightarrow I_b)_-)\}|}$$

  In the formula, f_j is a feature j, I_a and I_b are two images and $(I_a \rightarrow I_b)_{+/-}$ denotes that I_a and I_b were marked together positively (+) or negatively (-).

- Weighted Probabilistic :

$$factor_j = 1 + (2 * \frac{pp}{|\{(I_a \rightarrow I_b)_+\}|}) - \frac{np}{|\{(I_a \rightarrow I_b)_-\}|}$$

  Here, pp (positive probability) is the probability that the feature j is important for correct classification, whereas np (negative probability) denotes the opposite.

The additional factors calculated in this way are then simply multiplied with the already existing feature weights using tf/idf for the calculation of similarity scores for all the test images.

2.3 Classification

For each query image q, a set of $N \in \{1, 3, 5, 10\}$ result images k with a similarity score S_k were returned. The class of each result image were computed and the similarity scores were added up for the corresponding classes. The class with the highest accumulated score was then assigned to the query image. From preliminary experiments it was clearly visible that $N = 5$ produced the best results. This schema is very similar to a typical K-nearest neighbour (k-NN) classifier.

3 Results

This section presents our evaluation results based on the ground truth data supplied by the two ImageCLEF classification tasks. These runs were not officially submitted to the ImageCLEF2006 task because of time constraints. They were submitted to the organisers a few weeks late for comparison with the officially submitted results.

3.1 Classification on the LTU database

The non–medical automatic annotation task consisted of 14'035 training images from 21 classes. Data were made available by LookThatUp[6], and a set of more than 200 classes exists but such a task was regarded too difficult after a few tests. Subsets of images such as *computer equipment* were formed, mainly with images crawled from the web with a large variety for the contained objects. The task still remained hard with only three groups participating in it. The content of the images was regarded as extremely heterogeneous even for the same classes. Without using any of the described learning methods, using a simple 5–nearest-neighbour classifier, the GIFT had an error rate of 91,7%. Using the learning method with best/worst pruning and the frequency based weighting described above, the error rate decreased to 90,5%. Best results obtained in the competition by an optimised classification system were 77.3%, and the GIFT was not the worst–performing system submitting results.

3.2 Classification on the IRMA database

The medical image annotation task was done for the second time in 2006, after a first test in 2005. To augment the complexity, the number of classes was raised from 57 in 2005 to 116 in 2006. 10'000 images were made available as training data, and 1000 images with unknown classes had to be classified. The baseline results of the GIFT with various quantisations of grey levels can be seen in Table 1. They show clearly that more levels of grey do not help the classification, as error rates increase with them.

In Table 2, the results of the GIFT using the learning approaches described above can be seen. Surprisingly, the effect of the learning is quite small in comparison to the very good results obtained when aiming at increasing retrieval performance. The only method which improved the error rate at all was the frequency based weighting combined with best/worst pruning of the queries. Even here the difference is statistically not very significant.

[6] `http://www.ltutech.com/`

Number of grey levels	Error rate
4	32,0%
8	32,1%
16	34,9%
32	37,8%

Table 1: Error rates on the IRMA database using a varying number of grey levels.

Used strategy	Frequency weighting	Probabilistic weighting
S_1	35,3%	32,4%
S_2	33,2%	32,5%
S_3	31,7%	32,2%

Table 2: Error rates on the IRMA database using various weighting strategies and 4 grey levels. S_1 corresponds to using the naive strategy, S_2 to pruning the queries which were found too easy, and S_3 means that only the best positive and worst negative result of each query were taken into account.

We also combined eight grey levels with the described techniques but the results were always worse and thus not worth describing in more detail. Interestingly, the probabilistic weighting was not affected by the selections of relevant results in the same way as the frequency based weighting. Finally, we accumulated the scores of nearly all runs we performed. This combination results in an error rate of 29,7%, which shows that the approaches make differing errors and are thus theoretically combinable and at least in part independent.

The best system obtains a classification result of 16.2%, much better than we can obtain with our approach, that is rather good for information retrieval using relevance feedback.

4 Conclusion and Future Work

The provided tasks proved difficult to optimise for a frequency–based image retrieval system such as the GIFT. In comparison to last year's automatic annotation task, the number of classes roughly doubled thus making it more difficult to learn how discriminant features are. The GIFT heavily relies on this analysis, because compared to competitive CBIR systems only low level features not geared towards a specific task are used. We presented and applied a couple of approaches to learn from training data, but given the difficulty of the task could not improve the results significantly. The low level features in GIFT seem to be too sensitive for small changes that occur in such a particular tasks. Possibilities to circumvent this can be on two sides: On the one side it would be possible to normalised images further before they are indexed with the current features. Such a normalisation would include the removal of background as much as possible, the taking into account of the aspect ratio of the images, plus maybe a normalisation of the grey values within a class to have a more constant input. Another part could include changes in the feature space itself. Whereas the block features seem to be working much better with a smaller number of grey levels, the histogram features would rather require a much larger number to contain any good information. More complex features could include shape descriptors after a rough segmentation of the object in the images. Then of course a multi–step approach for classification can be imagined, where images are successively put into smaller and smaller classer and can change features at each stage of the segmentation process. Image classification seems to be hard with an information retrieval approach, but with features of more semantic value this seems to be feasible.

Acknowledgements

This work was partially supported by the Swiss National Foundation (Grant 205321-109304/1).

References

[1] P Clough, M Grubinger, T Deselaers, A Hanbury, and H. Müller. Overview of the imageclef 2006 photo retrieval and object annotation tasks. In *CLEF working notes*, Alicante, Spain, Sep. 2006.

[2] Paul Clough, Henning Müller, Thomas Deselaers, Michael Grubinger, Thomas M. Lehmann, Jeffery Jensen, and William Hersh. The CLEF 2005 cross–language image retrieval track. In *Springer Lecture Notes in Computer Science (LNCS)*, Vienna, Austria, September 2006.

[3] Thomas Deselaers, Tobias Weyand, Daniel Keysers, Wolfgang Macherey, and H. Ney. FIRE in ImageCLEF 2005: Combining content-based image retrieval with textual information retrieval. In *Working Notes of the CLEF Workshop*, Vienna, Austria, September 2005.

[4] H Müller, T Deselaers, T Lehmann, P Clough, and W Hersh. Overview of the imageclefmed 2006 medical retrieval and annotation tasks. In *CLEF working notes*, Alicante, Spain, Sep. 2006.

[5] Henning Müller, Antoine Geissbuhler, Johan Marty, Christian Lovis, and Patrick Ruch. The use of MedGIFT and EasyIR for ImageCLEF 2005. In *Working Notes of the 2005 CLEF Workshop*, Vienna, Austria, September 2005.

[6] Henning Müller, David McG. Squire, and Thierry Pun. Learning from user behavior in image retrieval: Application of the market basket analysis. *International Journal of Computer Vision*, 56(1–2):65–77, 2004. (Special Issue on Content–Based Image Retrieval).

[7] David McG. Squire, Wolfgang Müller, Henning Müller, and Thierry Pun. Content–based query of image databases: inspirations from text retrieval. *Pattern Recognition Letters (Selected Papers from The 11th Scandinavian Conference on Image Analysis SCIA '99)*, 21(13-14):1193–1198, 2000. B.K. Ersboll, P. Johansen, Eds.