Image Classification With a Frequency–Based Information Retrieval Scheme for ImageCLEFmed 2006

Henning Müller[1], Tobias Gass[2], Antoine Geissbuhler[1]

[1] Medical Informatics, University and Hospitals of Geneva, Switzerland
henning.mueller@sim.hcuge.ch
[2] Lehrstuhl für Informatik 6, RWTH Aachen, Germany
gass@informatik.rwth-aachen.de

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). Based on the training data, features appearing in images of the same class are weighted higher than features appearing across classes. These feature weights are added to the classical weights. Several weightings and learning approaches are applied as well as 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 schemes have independent results.

Keywords: Image Retrieval, Classification, Frequency-based

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 of the object annotation task and a photographic retrieval task can be found in [1]. The overview includes a description of the tasks, 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 [2] with all the details of submissions. More on the data can also be found on [3].

This article focuses on the submission for the two image classification tasks. The submissions were slightly after the deadline because of a lack of man power

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

but can be compared with the results in the overview articles. Already in 2005, an automatic medical image annotation task was offered in ImageCLEF [3]. Best results were obtained by systems using classical image classification techniques [4]. Approaches based on information retrieval techniques [5] had lower results but were still among the best five groups, without using any learning data. It was expected that a proper use of learning data could improve results significantly. Such a learning approach is attempted in this paper.

2 Methods

The methods described in this paper rely on the GIFT[4] (GNU Image Finding Tool) [6]. The learning approaches are based on [7].

2.1 Features

GIFT uses four features groups described in more detail in [6].

- Global color histogram based on the HSV (Hue, Saturation, Value) quantised into 18 hues, 3 saturations, 3 values and 4 grey levels.
- 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 scales and 4 directions, quantised into 10 bins.
- Local Gabor blocks by applying the filters above to the smallest blocks created by the recursive partition and using the same quantisation.

This results in 84'362 possible features where each image contains around 1'500. During the experiments, we increased the number of grey levels used for the color block features and color histogram features to 8, 16 and 32.

2.2 Feature Weights

The basic weighting used is *term frequency/inverted document frequency (tf/idf)*, which is well known from text retrieval. Given a query image q and a possible result image k, a score is calculated as the sum of all weights .

$$\text{score}_{kq} = \sum_j (\text{feature weight}_j) \tag{1}$$

The weight of each feature is computed as follows using term frequency(tf) and collection frequency(cf):

$$\text{feature weight}_j = tj_j * \log^2(1/(cf_j)) \tag{2}$$

This results in giving frequent features a lower weight. We use the approach in [7] and add learning strategies to optimise results for classification.

[4] `http://www.gnu.org/software/gift/`

Strategies The learning approach commonly uses log files and finds *pairs* of images marked together. Frequencies can be computed of how often each feature occurs in a pair. A weight can be calculated using the information whether 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 cooccurrence of features. In this paper, we want to train weights focused on classification. This means that we look at class memberships of images. Each result image is marked as relevant if the class matches that of the query image and non–relevant otherwise. We then applied several strategies for extracting the pairs of images. First, each pair of images occurring at least once is considered relevant. In the second approach we aim at discriminating positive and negative results more directly. Only the best positive and worst negative results of a query are taken into account. In a third approach, we pruned all queries which seem *too easy*. If the first N results were already positive we omitted the entire query from further evaluation. This is based on ideas similar to Support Vector Machines, where only information on class boundaries is taken into account.

Computation of Additional Feature Weights For each image pair, we calculate the features they have in common and whether these were positive or negative. We used two ways to compute an additional factor:

- Basic Frequency : Features are weighted by the number of occurrences in positive pairs, normalised by the number of all pairs.

$$\text{factor}_j = \frac{|\{f_j | f_j \in I_a \wedge f_j \in I_b \wedge (I_a \rightarrow I_b)_+\}|}{|\{f_j | f_j \in I_a \wedge f_j \in I_b \wedge ((I_a \rightarrow I_b)_+ \vee (I_a \rightarrow I_b)_-)\}|} \quad (3)$$

  where 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 :

$$\text{factor}_j = 1 + (2 * \frac{pp}{|\{(I_a \rightarrow I_b)_+\}|}) - \frac{np}{|\{(I_a \rightarrow I_b)_-\}|} \quad (4)$$

  where pp (positive probability) is the probability that feature j is important, whereas np (negative probability) denotes the opposite.

The additional factors calculated in this way are then simply multiplied with the already existing feature weights for the calculation of similarity scores.

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 was returned. The class of each result image was computed and the similarity scores were added up for the corresponding classes. The class with the highest accumulated score was assigned to the image. From preliminary experiments it was visible that $N = 5$ produced the best results. This is similar to a typical K–nearest neighbour (k–NN) classifier.

3 Results

3.1 Classification on the LTU Database

The non–medical automatic annotation task consisted of 14'035 training images from 21 classes. Subsets of images such as *computer equipment* were formed, mainly with images crawled from the web with a large variety. The task was hard and only four groups participated. The content of the images was regarded as extremely heterogeneous even for same classes. Without using any learning methods, using a simple 5–NN classifier, GIFT had an error rate of 91,7%. Using the learning method with best/worst pruning and the frequency–based weighting, the error rate decreased to 90,5%. Best results when separately weighting the four feature groups were 88.3%

3.2 Classification on the IRMA Database

10'000 grey level images from 117 classes were made available as training data and 1'000 images as test data. Baseline results of GIFT with various quantisations of grey levels are in Table 1. They show clearly that more grey levels do not help classification, as error rates increase surprisingly.

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

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

Table 3.2 shows results of GIFT using learning approaches. Surprisingly, the effect of learning is small. The only method which improved the error rate at all was the tf/idf weighting combined with best/worst pruning.

Table 2. Error rates on the IRMA database using 4 grey levels.

naive strategy	35,3%	32,4%
use best and worst results	31,7%	32,2%
removing too–easy queries	33,2%	32,5%

We combined eight grey levels with techniques but results were slightly worse. Finally, we accumulated scores of all runs performed resulting in an error rate of 29,7%, which shows that the approaches are combinable/independent.

4 Conclusion and Future Work

The provided tasks proved difficult to optimise for a frequency–based image retrieval system such as the GIFT using very simple features. Thus the features seem to be the main point for potential improvements when using a system similar to the GIFT. The various tested techniques showed that the system can profit from the training but that it needs to be done with much care. It can also be shown that tf/idf works very well on large collections without any knowledge but that classification, particularly when class sizes are very unbalanced needs more than lower weighting frequent features for good results. Some very frequent features might be important for the very large classes.

For future work it seems important to study features and feature groups independently as they are related and not independent. This means that a varying number of grey levels might be useful for local and global color features and that variations in the Gabor filters might also be useful in various ways (not discussed in this paper). Pre–treating images (background removal, normalisation of grey levels) allowing for more variation of the images with respect to object size and position are other approaches that are expected to improve results.

Acknowledgements

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

References

1. Clough, P., Grubinger, M., Deselaers, T., Hanbury, A., Müller, H.: Overview of the ImageCLEF 2006 photo retrieval and object annotation tasks. In: CLEF 2006 Proceedings. Lecture Notes in Computer Science (2007 – to appear)
2. Müller, H., Deselaers, T., Lehmann, T.M., Clough, P., Eugene, K., Hersh, W.: Overview of the imageclefmed 2006 medical retrieval and medical annotation tasks. In: CLEF 2006 Proceedings. Lecture Notes in Computer Science (2007 – to appear)
3. Clough, P., Müller, H., Deselaers, T., Grubinger, M., Lehmann, T.M., Jensen, J., Hersh, W.: The CLEF 2005 cross–language image retrieval track. In: Springer Lecture Notes in Computer Science (LNCS), Vienna, Austria (2006)
4. Deselaers, T., Weyand, T., Keysers, D., Macherey, W., Ney, H.: FIRE in ImageCLEF 2005: Combining content-based image retrieval with textual information retrieval. In: Working Notes of the CLEF Workshop, Vienna, Austria (2005)
5. Müller, H., Geissbuhler, A., Marty, J., Lovis, C., Ruch, P.: The use of MedGIFT and EasyIR for ImageCLEF 2005. In: Working Notes of the 2005 CLEF Workshop, Vienna, Austria (2005)
6. Squire, D.M., Müller, W., Müller, H., Pun, T.: 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** (2000) 1193–1198 B.K. Ersboll, P. Johansen, Eds.
7. Müller, H., Squire, D.M., Pun, T.: Learning from user behavior in image retrieval: Application of the market basket analysis. International Journal of Computer Vision **56(1–2)** (2004) 65–77 (Special Issue on Content–Based Image Retrieval).