Learning a frequency–based weighting for medical image classification

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

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

Abstract. This article describes the use of a frequency–based weightings developed for image retrieval to perform automatic annotation of images (medical and non–medical). The techniques applied are based on a simple *tf/idf* (term frequency, inverse document frequency) weighting scheme of GIFT (GNU Image Finding Tool), which is augmented by feature weights extracted from training data. These additional weights represent a measure of discrimination by taking into account the number of occurrences of these features in pairs of images of the same class or in pairs of images from different classes. This approach is fit to the image classification task by pruning parts of the training data. Since the results were not as good as expected, further investigations were performed showing that these weightings lead to significantly worse classification quality in certain feature domains. Hence, a classifier using a mixture of *tf/idf* weighted GIFT scoring, frequency–based learned feature weights, and regular Euclidean distance proved to give best results using only the simple features GIFT provides. Furthermore, using the aspect–ratio of images as an additional feature improved the results significantly for the medical images.

1 Introduction

Since the amount and importance of visual data in many domains rises each year it is of great interest to find efficient means to seek for visual information. Content–based image retrieval (CBIR) [1, 2] has therefore been one of the most active research areas in computer science over the last 15 years and will probably continue to be of high value. For example, the total amount of cardiology image data produced in the Geneva University Hospital was around 1 TB in 2002, which is impressive considering it is only one subsection of the data produced at the hospital in general [3]. CBIR usually deals with the problem to find images similar to a query consisting of one or more images (Query By Example, QBE). In the medical domain, considering an electronic multimedia patient record, this may help to find similar cases. Especially when using original medical DICOM (Digital Imaging and COmmunication in Medicine)[1] files for processing this can

[1] http://medical.nema.org/

aid in diagnosis and treatment. The GNU Image Finding Tool (GIFT)[2] [4] was developed at the University of Geneva and is suited for these tasks because it treats visual data in the same way as textual data. This makes it easy to incorporate visual and textual features in one processing step. Nonetheless, it is interesting to compare the performance of an information–retrieval based system such as GIFT, which uses very simple generic visual features, to other CBIR systems such as FIRE[3] [5], which is built to be flexible in means of available features and distance measures. The ImageCLEF evaluation campaign [6, 7] provides a platform for such a comparison, containing tasks in retrieval and classification of images in both the medical and non–medical domains. In this paper, we present various approaches to improve classification performance with the GIFT by keeping the simple feature space and learning frequency–based feature weights from the available training data.

2 Methods

The methods described in this paper rely heavily on those used in the GIFT. The learning approaches applied are based on learning algorithms published in [8] using the idea to translate the market basket analysis problem to image retrieval.

2.1 Databases

Two different databases from the ImageCLEF 2006 automatic annotation tasks[4] were used to evaluate classification performance.

IRMA The IRMA (Image Retrieval in Medical Applications, [9]) database of medical images was created at the RWTH Aachen[5]. It consists of 11'000 x–ray pictures of several parts of the human body. Each image is annotated with the label of one out of 116 classes. In the ImageCLEF medical automatic annotation task, 1'000 of these images without class label had to be classified using the 10'000 images with supplied label as training data. Size of the classes varies strongly. A great difficulty is the strong visual similarity between some classes. Since availability of computation power during the experiments was low, a set with 1'000 images was used for system optimisation. Later, when the number of experiments increased a randomly selected test and training set of 500 images each was utilised.

[2] http://www.gnu.org/software/gift/

[3] http://www-i6.informatik.rwth-aachen.de/ deselaers/fire.html

[4] http://ir.shef.ac.uk/imageclef/2006/

[5] http://www.rwth-aachen.de/

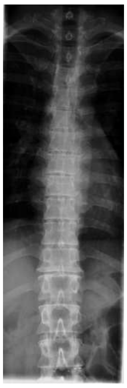

Fig. 1. Example x-ray of the spine.

Fig. 2. Example picture of class "oven"

LTU The LTU database, which was provided by the company LookThatUp[6], consists of images of a wide range of objects such as ashtrays or computer–equipment. A subset of 14'015 images of 21 classes was used for the non–medical automatic annotation task of the ImageCLEF2006 competition, of which 1'000 images served as an unlabelled test set for the evaluation. All experiments on the LTU database were done by using the settings derived from experiments with the IRMA–database without any special optimisation. This was done to show the ability to generalise from the derived results. In general, the non–medical automatic annotation task is hard due to the strong visual dissimilarities within the classes.

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

2.2 Features Used

GIFT itself uses four groups of image features, which are described in more detail in [4]. Here, a basic overview is given:

- A global color histogram, which is based on the HSV (Hue, Saturation, Value) 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 scales and 4 directions, which are quantised into 10 bins with the lowest one 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. Because of this and as a color histogram is usually an effective feature, we decided to increase the grey level 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[7] project. The GIFT uses this extension of the color space for both the block features and the color histogram. This may not be the best approach, since similarity for color blocks with only four different possible bits is already quite low. Hence, a separation of color spaces was tested, only using the enlarged color space for the color histogram features and not for the color block features.

2.3 GIFT Scoring

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 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.

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

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).

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

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

[7] http://www.sim.hcuge.ch/medgift/

other. An example for such a feature is black background being present in a large number of medical images. This applies only to the binary features. For histogram features a generalised histogram intersection is used to compute a similarity score [10].

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 [8]. In this article we use the described weighting approaches and add several learning strategies to optimise results for the classification task, where class membership of the entire training data is known.

Learning Strategies The original learning approach presented in [8] 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 in pairs of images. A weight can then 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 user interaction is not regarded but rather relevance data on class membership 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 these queries. In a first approach, each possible pair of images occurring together at least once is considered relevant. This yields very good results for image retrieval in general [8].

In the second approach we aim at discriminating positive and negative results in a more direct way. To do so, only the best positive and the worst negative result of a query are taken into account when computing pairs of marked images.

As 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. It assumes that all images that are in the middle of the class would be classified correctly anyways.

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.

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

  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 :

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

  Here, pp is the probability that the feature j is important for correct classification, whereas np 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.

2.4 Other Scoring Methods

During the experiments it became obvious that the frequency–based feature weights combined with the scoring method did not improve classification performance as much as hoped. Since GIFT uses four types of features it was necessary to have a more detailed idea of how the methods perform on each group of features. To achieve this, tests were performed where a single feature group was evaluated in GIFT. Experiments where Euclidean distance was used instead of the GIFT scoring were also tried. In the latter case we experimented with applying the learned feature weights to the distances, which worked surprisingly well.

2.5 Classification

With the similarity scores computed for each image, a simple 5–Nearest neighbour algorithm was used to classify unlabelled test data. Each vote was weighted by the similarity score achieved. The selection of using 5-NN was chosen based on manual tests performed in a first stage, where between 1 and 10 images were regarded with sometimes varying results.

3 Experimental Results

All optimisations were done on the IRMA database. Due to the constraints in available computational power partly on small, disjunct subsets as training and test data. The given error rates were obtained by applying the tested methods to the automatic–annotation tasks of the ImageCLEF2006 competition.

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 2. Error rates on the IRMA database using various weighting strategies and 4 levels of grey. 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.

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

3.1 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 to 116. 10'000 images were made available as training data and 1000 images had to be classified.

Enhancing the Color Space The baseline results of the GIFT can be seen in Table 1. They show clearly that a larger number of grey levels does not help the classification, as error rates increase.

Frequency–Based Learned Feature Weights In Table 2, the results of the GIFT using the learning approaches described above can be seen. Surprisingly, the effect of the learning is small in comparison to the good results obtained for retrieval. 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 significant.

We also combined eight grey levels with the described techniques but the results were always worse and thus not worth mentioning. Interestingly, the probabilistic weighting was not as much affected by the selections of relevant results as the frequency–based weighting.

Classification on Single Feature Groups In these experiments we classified the data by using each feature group separately. Then, the varying weighting strategies were performed. The probabilistically learned feature weights were

Table 3. Error rates on the four feature groups using several weighting approaches.

Feature group	unweighted baseline	with tf/idf	learned weights	tf/idf+learned weights
Color block	36,6%	39,6%	35,1%	40,4%
Color hist	74,5%	–	73,8%	–
Gabor block	56,3%	42,3%	50.0%	45,4%
Gabor hist	53,1%	–	51.8%	–

Table 4. Best setup for classification.

Feature group	scoring method	learned feature weights
Color block	L2	–
Color hist	GIFT tf/idf	–
Gabor Block	GIFT Histogram Intersection	used
Gabor hist	L2	used

omitted because of inferior performance in earlier experiments. It turned out that performance varies greatly, so a classification with mixed scoring methods seems most viable.

If the classification in GIFT was performed without any weighting whatsoever, the error rate increased from 32% to 34%, so a more detailed approach is necessary.

Mixed Scoring As described earlier, it is interesting to see how the GIFT scoring method performs in comparison to standard metric–based similarity measures. The first results were interesting as a simple Euclidean–distance–based 5–NN outperformed the GIFT by decreasing the error rate to 29.8%. At this point, several experiments on small test and training sets were conducted in which GIFT scoring, Euclidean distance(L2), and a few other feature weightings were tested. The methods with the best results on these subsets were then performed, improving the error rate significantly to 27.5%. This score was achieved with the following scoring method/weighting approach:

Aspect Ratio In the medical image domain and particularly for x–rays contained in the IRMA database, the aspect ratio of an image is highly correlated to the content of the image. This seems logical since x–rays are usually truncated to show just the region of interest, and bones from the arm, for example, have a significantly different form than a chest. This leads to the idea to use the aspect ratio as a fifth feature group and include it into classification (Figure 3). This approach again improved the classification error rate on the best setup we used from 27,5% to 26.4%.

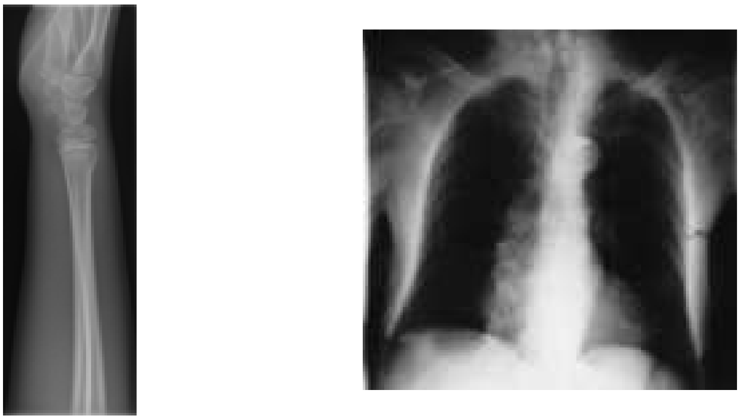

Fig. 3. X–rays of a leg and the chest, with different aspect ratios

Table 5. Error rates on the LTU database using various strategies.

Method used	Error rate
baseline	91,7%
with learned feature weights	90,5%
with mixed scoring	88,3%
classifier combination	89%

3.2 Classification on the LTU Database

The non–medical automatic annotation task consisted of 14'035 training images from 21 classes selected from a total set of more than 200 classes and over 100'000 images. The entire dataset was regarded as too difficult after a few tests, so subsets like *computer equipment* were formed, mainly with images crawled from the web with a large variety for the contained objects. The task remained hard with only three research groups finally submitting results. The content of the images was regarded as extremely heterogeneous even for the same class. Without using any of the described learning methods and 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%. After that, we applied the mixed scoring method derived from the experiments on the IRMA database and achieved an error rate of 88.3%. A combination of available results could not further improve the classification performance.

4 Interpretation

The results show clearly that the approach with very simple visual features taken in GIFT originally is not that perfectly suited for image classification. GIFT uses four groups of simple global and local features, with just two similarity measures (histogram– and non–histogram features). It is, due to the good generalisation of the methods to more than one database, obvious that color and texture features have to be treated differently. Regarding the results, it seems that color features appearing frequently in the entire collection are still necessary to discriminate classes from each other. This can be due to very large classes that have many features in common and misclassifying some of these images can be more costly than loosing performance on very small classes. If these features get reduced in weight too much, the performance decreases. On the other hand, texture features, which occur often throughout the training data are carrying less discriminative information and thus perform better when they are weighted accordingly.

5 Conclusions and Future Work

In this article, we have shown the possibilities to use a frequency–based weighting scheme developed for image retrieval in a classification context. The performance of these weights depends on the kind of features they are applied to, where color features seem to be less weighable or learnable than texture features. In general, the performance of the derived methods is still lower than other CBIR systems available. This results mostly from the very simple feature set used that does not take into account small movements or changes in size of the object in the image. Pre–treatment of images to remove background might be one solution. Another solution is the use of salient features for retrieval.

6 Acknowledgements

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

References

1. Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A., Jain, R.: Content–based image retrieval at the end of the early years. IEEE Transactions on Pattern Analysis and Machine Intelligence **22 No 12** (2000) 1349–1380
2. Eakins, J.P., Graham, M.E.: content–based image retrieval. Technical Report JTAP–039, JISC Technology Application Program, Newcastle upon Tyne (2000)
3. Müller, H., Michoux, N., Bandon, D., Geissbuhler, A.: A review of content–based image retrieval systems in medicine – clinical benefits and future directions. International Journal of Medical Informatics **73** (2004) 1–23
4. 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.
5. 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)

6. 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)
7. 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)
8. 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).
9. Lehmann, T.M., Güld, M.O., Thies, C., Fischer, B., Spitzer, K., Keysers, D., Ney, H., Kohnen, M., Schubert, H., Wein, B.B.: Content–based image retrieval in medical applications. Methods of Information in Medicine **43** (2004) 354–361
10. Swain, M.J., Ballard, D.H.: Color indexing. International Journal of Computer Vision **7** (1991) 11–32