Pattern Recognition Letters xxx (2008) xxx–xxx

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Hierarchical classification using a frequency-based weighting and simple visual features

Xin Zhou [a,*], Adrien Depeursinge [a], Henning Müller [a,b]

[a] Service of Medical Informatics, Geneva University Hospitals and University of Geneva, 24, Rue Micheli-du-Crest, 1211 Geneva 11, Switzerland
[b] University of Applied Sciences, Sierre, Switzerland

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Image retrieval
Hierarchical image classification
Automatic annotation
Medical imaging

ABSTRACT

This article describes the use of a frequency-based weighting scheme using low level visual features developed for image retrieval to perform a hierarchical classification of medical images. The techniques are based on a classical *tf/idf* (term frequency, inverse document frequency) weighting scheme of the *GIFT* (GNU Image Finding Tool), and perform classification based on kNN (*k*-Nearest Neighbors) and voting-based approaches. The features used by the GIFT are very simple giving a global description of the images and local information on fixed regions both for colors and textures. We reused a similar technique as in previous years of ImageCLEF to have a baseline for the retrieval performance over the three years of the medical image annotation task. This allows showing the clear increase in quality of participating research systems over the years.

Subsequently, we optimized the retrieval results based on the simple technology used by varying the feature space, the classification method (varying number of neighbors, various voting schemes) and by adding new information such as aspect ratio, which has shown to work well in the past. The results show that the techniques we use have several problems that could not be fully solved through the applied optimizations. Still, optimizations improved results enormously from an error value of 228 to below 150. As a baseline to show the progress of techniques over the years it also works well. Aspect ratio shows to be an important factor to improve results. Performing classification on an axis level performs better than using the entire hierarchy code or not taking hierarchy into account at all. To further improve results, the use of more suitable visual features such as patch histograms or salient point features seems necessary. Small distortions of images of the same class have to be taken into account for very good results. Still, without using any learning technique and high level visual features, the approach performs reasonably well.

© 2008 Elsevier B.V. All rights reserved.

1. Introduction

Medical images are an extremely important part of the diagnosis process in medical institutions. As most hospitals now have computerized patient records and fully digitized image production, new possibilities arise for management of data and the extraction of information from the stored data (Müller et al., 2004a; Tagare et al., 1997; Vannier et al., 2002). At the same time of images becoming digital, the number of images produced and their complexity has increased strongly. The Geneva University Hospitals radiology department alone produced over 70,000 images per day in 2007 (Müller et al., 2007) and these numbers continue to rise.

In other domains, content-based image retrieval has been used for many years to manage the growing amount of visual data (Datta et al., in press; Smeulders et al., 2000; Kato, 1992; Rui et al., 1999). While early approaches used fairly low level features such as global color distributions and texture characteristics (Niblack et al., 1993), more modern systems rather use local features either gained through segmentation (Winter and Nastar, 1999) or in the form of salient points and their relations (Fergus et al., 2004; Tommasi et al., 2007). The latter obtained the best result in ImageCLEF 2007.

Object recognition in images has been another active research area to extract important information from potentially non-annotated images (Everingham et al., 2006; Pinz, 2005). In the medical domain, similar approaches have been used for medical image classification to extract information from these images (Lehmann et al., 2005). The dataset of the IRMA project (Image Retrieval in Medical Applications) is also used in the ImageCLEF[1] benchmark, of which a participation is described in this article. Many of the techniques for image retrieval and for image classification are similar but

* Corresponding author. Fax: +41 22 372 8879.
E-mail addresses: Xin.Zhou@sim.hcuge.ch (X. Zhou), Adrien.Depeursinge@sim.hcuge.ch (A. Depeursinge), Henning.Mueller@sim.hcuge.ch (H. Müller).

[1] http://www.imageclef.org/.

0167-8655/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2008.04.004

whereas for classification, a finite number of classes is regarded and training data are often available, for information retrieval applications, the number of classes occurring in the dataset is often unknown and training data are rarely available.

Several steps can generally be tuned to optimize the final performance.

- Image pre-processing such as segmentation (Antani et al., 2004), normalization of gray levels, or background removal (Müller et al., 2005).
- Extraction of domain-specific visual features (Müller et al., 2004b).
- Optimization of the distance measure or weighting scheme to determine distances between elements.
- Application of a learning strategy (such as Support Vector Machines) (Qiu, 2006).

In our approach, we do not take into account any pre-processing and neither any learning strategy. Efforts are concentrated on the optimization of the feature space and particularly on a classification strategy with our simple features to test the limits of our retrieval engine, the GIFT.[2] This cannot rival in performance with more modern approaches particularly for learning/classification such as the use of Support Vector Machines (SVMs) (Chapelle et al., 2002) or salient point-based visual features (Tommasi et al., 2007).

More on the ImageCLEFmed benchmark, the corresponding classification setup, error calculation, and the other participating techniques can be read in (Deselaers et al., in press).

In Section 2, the methods of our approach are explained in detail. Section 3 presents the results obtained with these methods. In the last section, we critically interpret our results and present the conclusions of this article.

2. Methods

This section describes the data used and the techniques employed.

2.1. Database and task description

We use the dataset of the ImageCLEFmed 2007 automatic classification task containing in total 10,000 training images, 1,000 validation images and 1000 test images. The 1000 test images had to be classified according to the full IRMA code (Lehmann et al., 2003), which is a mono-hierarchical code with four distinct axes (image modality, anatomic region, biosystem under examination, and the body orientation all have their own hierarchy). Classification was allowed to stop at any level of the hierarchy within any of the axes. Non-classified hierarchy levels were regarded as better than incorrectly classified parts to force participants to think about measures of confidence in the classification strategy. A single image can be classified completely incorrectly (error value equal to 1), completely correctly (error value equal to 0) or partly incorrectly (error value between 0 and 1). The maximum error value can be obtained when all the 1000 test images are incorrectly classified, equaling 1000. If all the images are classified as "unknown" the total error value equals 500. A short explanation of this error value calculation is detailed in.[3] More information about the system setup and the error scoring methodology can be found in (Deselaers et al., in press).

2.2. Technical description

The techniques used for visual similarity calculation are mainly those used in the *GIFT* system (Squire et al., 2000). This tool is open source and can be used by other participants of ImageCLEF as well, so all results are reproducible. The image classification is processed in four steps:

(1) indexation of the entire image database with visual features (including the images to be classified);
(2) execution of queries with images to be classified to get similar images with known classification;
(3) re-ordering of the similar images with additional features;
(4) classification of the query image based on the list of similar images and their classes.

Varying parameters were used in steps 1, 3, and 4 to obtain improvement. Several gray level quantizations were used in the indexation step. Varying weights were attributed to the additional features (mainly aspect ratio). These two parts were already studied for a similar task in 2006 (Gass et al., 2007), so this paper investigates rather the effect of varying classification strategies.

2.2.1. Visual features

The four distinct visual feature sets used by *GIFT* are:

- Local color features at different scales by partitioning the images successively into four equally sized regions (four times) and taking the mode color of each region as a binary descriptor.
- Global color features in the form of a color histogram, compared by a simple histogram intersection.
- Local texture features by partitioning the image as before and applying Gabor filters in various scales and directions, quantized into 10 strengths (where the lowest band can be discarded).
- Global texture features represented as a simple histogram of responses of the local Gabor filters in various directions and scales.

The color histogram is originally based on the HSV (Hue, Saturation, Value) color space. Gray levels are added in a varying number as the entire database contains no color images. The texture feature space is based on two parameters: the number of directions and the scale of the Gabor filters. A more detailed description of the *GIFT* feature set can be found in (Squire et al., 1999).

Based on the results from 2006, a varying number of gray levels $(4, 8, 16, 32)$ were tested in this paper. Together with HSV values of $(9, 3, 3)$, this results in a total of 60,833 possible features descriptors, most of them of binary nature. A large part of this feature space is unpopulated as the database contains only gray scale images and no color features are thus possible. A normal image contains around 1000 of these features but the numbers can vary depending on the amount of texture and the number of gray levels present.

2.2.2. Feature weighting

A particularity of *GIFT* is that it uses many techniques well–known from text retrieval. Visual features are quantized and the distributions of the features are fairly similar to those of words in texts (sparsely populated spaces). A simple *tf/idf* weighting is used and the query weights are normalized by the results of the query itself. The features using histograms are compared based on a simple histogram intersection (Swain and Ballard, 1991). The four feature groups are combined in normalized form with an equal weight. Feature groups can also be used directly without separate normalization leading to significantly worse results. This

[2] http://www.gnu.org/software/gift/.
[3] http://www-i6.informatik.rwth-aachen.de/~deselaers/imageclef07/hierarchical.pdf.

UNCORRECTED PROOF

Please cite this article in press as: Zhou, X. et al., Hierarchical classification using a frequency-based weighting ..., Pattern Recognition Lett. (2008), doi:10.1016/j.patrec.2008.04.004

technique was used in our original participation in the classification having a much lower performance.

Visually similar images with known classes are then used to classify images from the test set. In practice, the 100 most similar images for every image of the test set were taken into account, and the similarity scores (see Eq. (1)) of these images were used to perform the classification.

The similarity score for each image k towards a query q is calculated in the following way:

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

The weight of each feature j for a query q is computed by dividing the term frequency (*tf*) of the feature by the squared logarithm of the inverted collection frequency (*cf*).

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

Through normalization, a similarity score is always in the range of $[0;1]$ for single image queries, where this can be slightly different for multiple image queries. The four normalized results of the feature groups are subsequently combined.

2.2.3. Additional features

In GIFT, no scale-invariant features are employed. For ease of similarity calculation all images are transformed to 256×256 pixels. So GIFT does not take into account the *aspect ratio* of the images, which has proven to be a useful criterion in past results (Gass et al., 2007).

The similarity of two images concerning the aspect ratio is calculated as follows:

$$\mathrm{score}_{AR} = |AR_1 - AR_2|, \tag{3}$$

where AR is the aspect ratio of each of the images to be compared.

The function to combine the aspect ratio with the GIFT similarity score is given in Eq. (4). As the similarity is inversely proportional with score_{AR}, the sign of the value is negative. A weighting factor w is used to vary the strength of this feature

$$\mathrm{score}_{\mathrm{final}} = \mathrm{score}_{\mathrm{GIFT}} * (1 + w) - \mathrm{score}_{AR} * w \tag{4}$$

2.2.4. Classification strategies

For our participation in the hierarchical classification of ImageCLEFmed 2007 we decided to not use any learning strategy due to a lack of time in the preparation of the event. The two main classification approaches tested are the following:

- a classical kNN approach using $k = 1,\ldots,20$ nearest neighbors;
- an approach using a voting of the $n = 1,\ldots,100$ most similar images and then a threshold for whether to classify or decide to not classify an image at a certain hierarchy level.

k is thus reserved for the kNN approach and n for the number of votes in the voting-based approach. Based on past experiments we take into account the 100 most similar images for the classification. In the voting-based approach, up to the first $n = 1,\ldots,100$ retrieved images vote for their respective class. This remains a technique fairly similar to standard kNN approaches with integration of information about the confidence of the voting.

Two weight distribution strategies were implemented in the voting approach:

- every retrieved image votes with an equal weight;
- retrieved images vote with decreasing values (from n down to 1) based on their rank.

Confidence of the voting is an additional condition to validate the choice. If the confidence score is not reached the code at a certain level will be classified as "unknown". The total value of the voting weights is shown in the following equation:

$$\mathrm{weight}_{\mathrm{total}} = \sum_{k=1}^{n} \mathrm{weight}_k \tag{5}$$

The weight_k is based on the weight distribution strategy. One choice can be valid only if the sum of the voting weights for one code reached a certain percentage of the total weight. This percentage is named *threshold*.

Three different ways to include hierarchy information into the classification were tested to find out whether it makes sense to use the hierarchy and up to which degree results can improve with the hierarchy information.

- *The total code level:* the entire code is considered to be one single entity.
- *The axis level:* the four code axes are treated separately but each axis is considered to be a single entity.
- *The letter level:* every letter of the code is treated separately.

Most of the best-performing techniques in the benchmark actually did not use the hierarchy at all, so one of our goals was to find out whether hierarchy information can at least be used up to a certain level.

3. Results

This section details the results obtained with the various techniques. The results of all participating research groups are compared with error values in (Deselaers et al., in press).

3.1. Changes in the feature space

In a first step, changes in the feature space were tested to get an optimal setup for further steps in the classification. The classification strategy used in this step is a classical kNN approach with $k = 1,\ldots,20$. The entire code was taken as entity and no hierarchy information was taken into account. Each time the lowest error value with the corresponding k is given and the average over all 20 values. We can see in Table 1 that a very large number of gray levels does not give better results. Average error values show that 8 and 16 gray levels obtain the best results, which was similar in past studies.

3.2. Addition of aspect ratio

Besides variation in the number of gray levels we added the aspect ratio as feature. The results are shown in Table 2. When adding the aspect ratio the performance becomes better (by over 40 points or 20%), underlining the importance of aspect ratio. The average error values show that combined with aspect ratio at all proportions the error value decreases significantly. The optimal value for w varies significantly for the two tested gray level quantizations. As for 16 gray levels, the best value was at 10, so we also tested lower parameters trying to find the local maximum. Two confusion matrices are shown in Figs. 1 and 2 to study the benefit of aspect ratio. Only a subset of the classes received a clear benefit from adding aspect ratio. For classes 40–60 a clear improvement can be observed. The classes with improvement mainly belong to lower extremity/leg part (foot, lower leg, knee, etc.). Aspect ratio is an important criterion for these classes as image are far from quadratic. It can also be shown in the confusion matrices that the classes 98 and 48 are responsible for most of the errors. These two classes are cervical spine images. There are around 300 images of these two classes in the training data (3% of the training dataset)

UNCORRECTED PROOF

Please cite this article in press as: Zhou, X. et al., Hierarchical classification using a frequency-based weighting ..., Pattern Recognition Lett. (2008), doi:10.1016/j.patrec.2008.04.004

Table 1
Varying results for small changes in the feature space

Variation	Lowest error value (at $k = \ldots$)	Average error (at $k = 1,\ldots,20$)
4 gray levels	247.13 ($k = 4$)	263.01
8 gray levels	209.95 ($k = 4$)	226.11
16 gray levels	202.48 ($k = 4$)	224.87
32 gray levels	205.04 ($k = 2$)	249.63

Table 2
Influence of aspect ratio on the classification

Variation	Lowest error (at $k = \ldots$)	Average error (at $k = 1,\ldots,20$)
8 gray levels with $w = 10\%$	183.57 ($k = 3$)	208.87
8 gray levels with $w = 20\%$	182.71 ($k = 4$)	207.25
8 gray levels with $w = 30\%$	184.23 ($k = 3$)	206.76
8 gray levels with $w = 40\%$	180.78 ($k = 4$)	206.27
8 gray levels with $w = 50\%$	179.73 ($k = 3$)	205.95
8 gray levels with $w = 60\%$	180.99 ($k = 5$)	205.90
8 gray levels with $w = 70\%$	181.66 ($k = 5$)	206.02
16 gray levels with $w = 2\%$	175.54 ($k = 3$)	209.24
16 gray levels with $w = 5\%$	162.49 ($k = 3$)	203.62
16 gray levels with $w = 10\%$	160.59 ($k = 2$)	201.87
16 gray levels with $w = 20\%$	162.34 ($k = 2$)	202.03
16 gray levels with $w = 30\%$	163.79 ($k = 2$)	203.35
16 gray levels with $w = 40\%$	166.77 ($k = 2$)	203.72
16 gray levels with $w = 50\%$	170.78 ($k = 5$)	203.76

and 30 images in test data (3% of the test dataset). However, more than 191 test images (20% of the test dataset) are classified into these two classes. By consequence, at least 161 test images (16% of the test dataset) are misclassified. Almost all the classes have images that are confused by the system with cervical spine images, maybe due to the small-scale textures and our global little expressive features.

3.3. Changes in the classification strategy

In a second step, two separate classification strategies further described in Section 2.2.4 were tested. As the aspect ratio improves the result significantly we continue to use it in all further retrieval steps.

3.3.1. kNN approach using the supplied hierarchy

The first strategy is a pure kNN strategy with k being the number of similar images needed to classify a certain IRMA code at a certain level. Fig. 3 shows the error value based on variations of k.

Not surprisingly, small k values lead to best results. For all three hierarchy levels, $k = 2$ delivers the best result with an error value of around 161. The 5 biggest classes contain almost half of the images in the training data (4866 of 11,000). Large k values result in mistakes in classes with a small number of examples in the training data. It can also be noticed that taking into account various hierarchy levels as an entity has an impact on the results. Taking the entire IRMA code into account obtains best results but the scores vary strongly depending on k. When considering every letter in the code as an entity the best result is significantly worse. Classification per axis has fairly good results (slightly worse than for the entire code) but results are more stable concerning changes of k. Table 3 shows the best and average error values.

3.3.2. Voting-based approach using the supplied hierarchy

In a next step, a confidence threshold for classification with voting was introduced. Goal is to find out how to best estimate the confidence in our classification. Two strategies were used for the voting using the same weight for all results or decreasing weights based on rank. In Table 4 all three hierarchical levels were taken into account and impact of the threshold was measured. The best runs for every strategy are shown in Fig. 4.

Table 4 shows results of the voting-based approach. The performance is significantly better than the simple kNN approach, particularly when the classification is performed per axis. Tests were performed taking into account up to 40 images but performance is generally best for values below 10. Lower average error values for voting with descending weights show that the stability with this approach is higher as well.

Fig. 5 shows the percentage of incorrectly classified images for the voting-based approach. The six best runs were selected from the six strategies (two weight distribution strategies combined

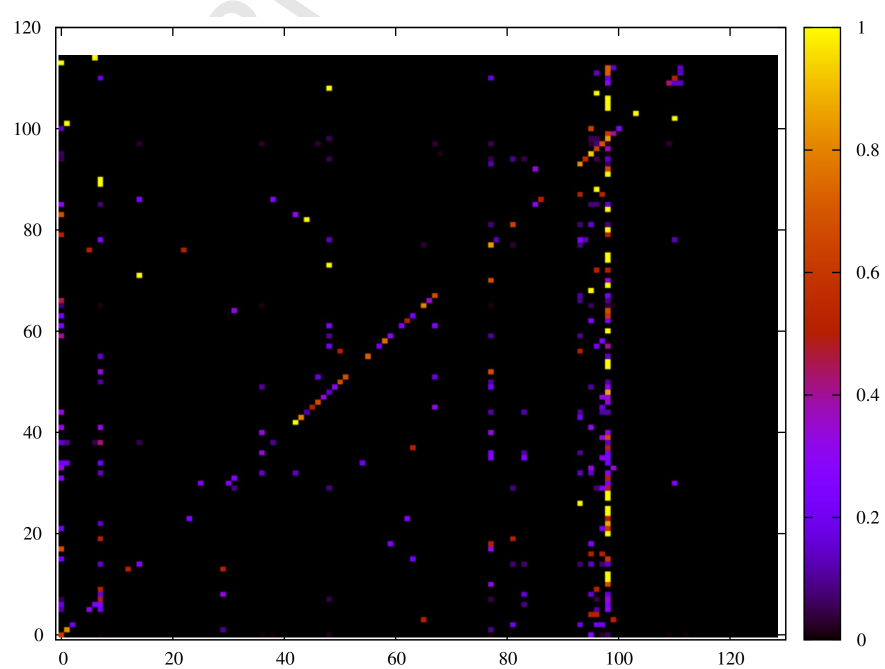

Fig. 1. Confusion matrix: gray levels = 16, without including AR, $k = 4$.

Please cite this article in press as: Zhou, X. et al., Hierarchical classification using a frequency-based weighting ..., Pattern Recognition Lett. (2008), doi:10.1016/j.patrec.2008.04.004

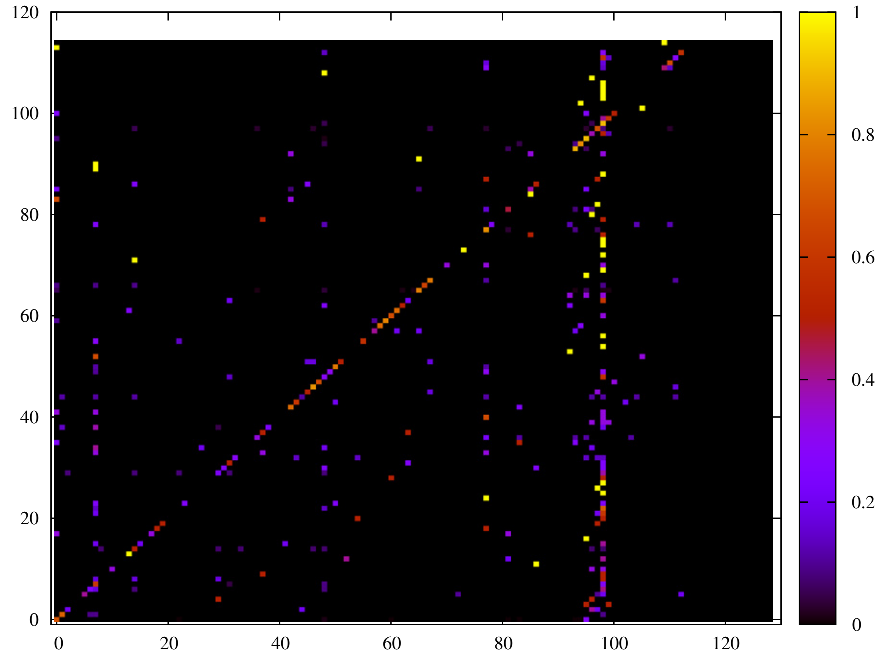

Fig. 2. Confusion matrix: gray levels = 16, w for AR = 10%, k = 2.

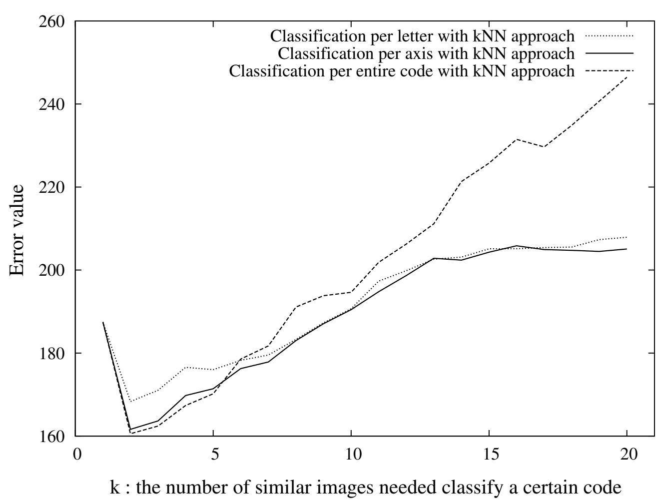

Fig. 3. Results with varying k values using a simple kNN classifier.

Table 3
Error values for classification at various levels of the hierarchy

Variation	Lowest error (at k = 2)	Average error (k = 1,...,20)
Entire code level	160.59	201.87
Axis level	161.62	189.84
Letter level	168.34	191.89

with three hierarchy levels). The results show that the error percentage is not completely in line with the error values based on the hierarchy. This is due to the fact that only fully correctly classified images are regarded as correct, and as a consequence not taking into account the hierarchy obtains best results. The threshold does not improve results when not taking into account the hierarchy.

In total, the dataset contains 116 classes but when dividing them by axis there are only four different classes for technical code,

Table 4
Classification results with varying voting strategies

Strategy	Threshold	Lowest error (at n = ...)	Average error (n = 1,...,40)
Entire code level			
Voting with equal value	0	161.50 (n = 5)	180.28
	0.1	161.50 (n = 5)	180.57
	0.2	171.01 (n = 8)	189.42
	0.3	173.72 (n = 6)	209.38
	0.4	187.48 (n = 1)	244.91
Voting with decreasing value	0	155.66 (n = 9)	168.22
	0.1	155.66 (n = 9)	168.26
	0.2	158.45 (n = 8)	173.98
	0.3	162.61 (n = 5)	192.05
	0.4	183.46 (n = 3)	225.21
Axis level			
Voting with equal value	0.2	161.62 (n = 5)	182.50
	0.3	160.36 (n = 7)	178.36
	0.4	152.67 (n = 3)	174.60
	0.5	152.67 (n = 3)	175.99
	0.6	152.67 (n = 3)	186.13
Voting with decreasing value	0.2	158.02 (n = 6)	170.38
	0.3	153.00 (n = 6)	166.32
	0.4	150.43 (n = 6)	162.54
	0.5	149.34 (n = 5)	163.65
	0.6	158.24 (n = 7)	176.38
Letter level			
Voting with equal value	0.3	172.29 (n = 3)	189.96
	0.4	159.62 (n = 5)	184.50
	0.5	159.62 (n = 5)	175.82
	0.6	159.66 (n = 6)	176.61
	0.7	161.84 (n = 7)	186.35
Voting with decreasing value	0.3	164.67 (n = 6)	176.50
	0.4	161.07 (n = 5)	172.53
	0.5	154.04 (n = 8)	165.28
	0.6	154.69 (n = 7)	167.58
	0.7	164.54 (n = 6)	177.99

26 for orientation, 63 for body region, and five for bio system. Setting a threshold can limit noise when classification is performed on

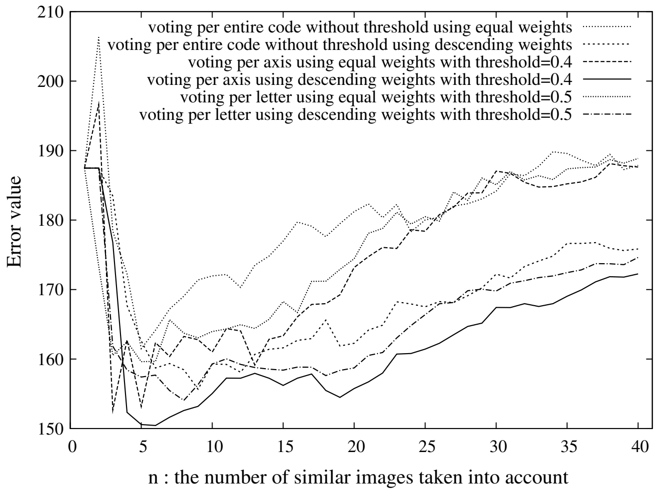

Fig. 4. Classification results taking into account the first *n* images with a voting-based scheme.

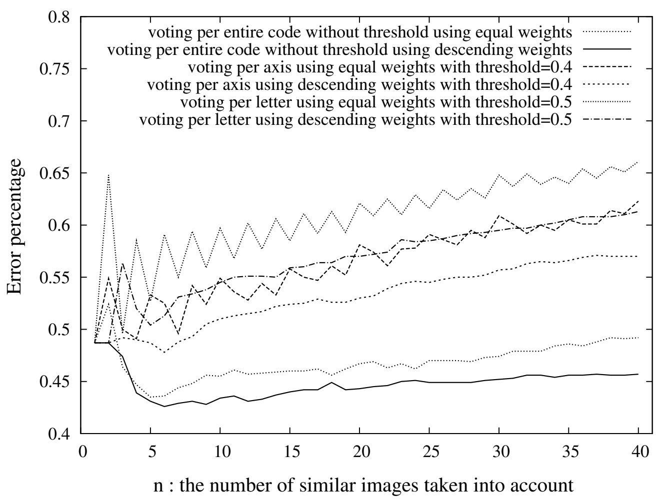

Fig. 5. Error percentages when using the voting-based scheme.

a per axis basis and thus it improves results. Lower thresholds are better on a smaller hierarchy level.

The fact that axis level classification leads to best results is interesting as the best overall system did not at all use this information. A fully detailed letter level classification in our system gives worse result than a per axis classification. An explication for letter level classification not improving results is that the meaning of each axis is independent whereas within a single axis every letter depends on the higher level. For example, within the axis code 940 in body region, the letter 4 means "knee", while for 540, 4 means "mediastinum". Thus taking an entire axis as an entity is a reasonable approach.

The best overall runs with the optimized parameters are listed in Table 5. They are all based on voting on an axis level with descending weights and with a threshold filter.

3.4. Computational analysis

The computation times for the four processing steps described earlier are given in Table 6. These indexing and query times were obtained on a simple server with two DualCore Xeon CPUs with 2.33 GHz, and 4 GB of RAM. The two last steps for the classification were performed on a desktop computer with a CoreDuo CPU with 2.79 GHz and 2 GB of RAM.

Table 5
Best results obtained with the GIFT system and aspect ratio

n	Threshold	hierarchy level	With AR	Error value
5	0.5	Axis	Yes	149.34
6	0.5	Axis	Yes	150.14
6	0.4	Axis	Yes	150.43

Table 6
Time consumption for the processing steps

Processing stage	Time consumption	Activity
Database indexation	2 h	Indexation of 12,000 images
Queries for similar images	10 min	Querying 1000 times
Reordering with AR	3 s	re-ordering of 1000 × 100 similar images
Classification	1–3 s	Classification 1000 times

Table 7
Evaluating the validation dataset with out optimal parameters

n	Threshold	Hierarchy level	With AR	Error value
Runs with the optimal parameters of the test dataset				
5	0.5	Axis	Yes	143.76
6	0.5	Axis	Yes	147.18
6	0.4	Axis	Yes	149.05
The three best runs for the validation dataset				
9	0.4	Axis	Yes	142.61
5	0.5	Axis	Yes	143.76
8	0.4	Axis	Yes	144.18

3.5. Stability of the expected results

Our method does not include any training strategy. Optimizing the parameters as we did directly on the test set introduces a bias compared with systems optimized on the validation dataset. To show the relative stability of our algorithm we also show the optimized parameters for the validation dataset as seen in Table 7. It can be seen the absolute optimums are slightly different on the validation and the test datasets but it can also be seen that the best result on the test dataset obtains the second best result on the validation dataset. This underlines a certain stability of our proposed optimized values across datasets.

4. Interpretation and discussion

In comparison with systems using modern visual techniques and machine learning approaches, the GIFT system with a simple kNN classification and without any learning strategy has a relatively low performance. However, the GIFT runs were initially meant to be a baseline to allow comparison with other techniques. The best overall results were obtained using SIFT (Scale Invariant Feature Transform) features and SVM-based learning approaches. Other top results used histograms of image patches or salient point-based features for the image description, approaches that are much more complex than the simple GIFT features.

Optimizations showed the varying influences of the parameters on the classification quality. Changes in the gray level quantization have an important influence on the classification results, improving results by over 40 points. Best results are obtained with 8 and 16 gray levels. The confusion matrix shows that with the existent features many images (particularly chest images that are similar to cervical spine images) are incorrectly classified. Hand images and non-specified organ tissue are two other classes with a high error rate. The reasons for these two classes can be two-fold. Tissue images might contain small-scale information in the form of absolute texture and thus absolute gray level histograms with no possible variation cannot lead to good results. For hand images two very similar classes are often misclassified among each other.

Aspect ratio improves the result significantly by around 20 error points on average. A few classes profit particularly from this additional information as described in the results section.

Another parameter optimized is the hierarchy level taken into account for the classification. The best-performing systems in the competition all did not take into account the hierarchy information. Most other groups who used hierarchy information tried only to perform the classification per letter. In our kNN approach the classification also obtains best results when not taking into account hierarchy information, albeit the difference between classification on a global and an axis level are not important. The classification when performed on an axis level is more stable with respect to the parameter k.

When using our voting approach, the results with classifying images per axis obtains best results, although only slightly better than when omitting hierarchy information. Classification on the level of the full hierarchy still obtained the worst overall results. The voting strategy obtained better overall results than the kNN classifier. Part of the reason for the axis level classification working better is that errors often occur rather on the axes with more complexity and not on all axes with only few choices. Most often, several axes could be classified correctly although the overall classification was incorrect.

A small number of similar images is sufficient to obtain the lowest error values. For kNN the value for k is usually around 2–4. For the voting approach, less than 10 most similar images obtained the best performance.

This article shows that even with extremely simple techniques and without any learning strategy good results can be obtained, although not in the same league of the results of more sophisticated techniques. Still, the article shows the varying influences of features, classification strategies, and also that the hierarchy information can improve results. The best techniques might actually well profit from taking into account at least the axis information to perform the classification as well.

Acknowledgements

This study was partially supported by the Swiss National Science Foundation (Grant 205321–109304/1) and the European Union in the Sixth Framework Program through the KnowARC project (Grant IST 032691). We also thank the reviewers for their constructive comments that helped to improve this paper.

References

Antani, S., Lee, D., Long, R., Thoma, G., 2004. Evaluation of shape similarity measurement methods for spine X-ray images. J. Visual Comm. Image Represent. 15 (3), 285–303.

Chapelle, O., Vapnik, V., Bousquet, O., Mukherjee, S., 2002. Choosing multiple parameters for support vector machines. Mach. Learn. 46 (1), 131–159.

Datta, R., Joshi, D., Li, J., Wang, J.Z., in press. Image retrieval: Ideas, influences, and trends of the new age. ACM Comput Surveys 65.

Deselaers, T., Müller, H., Deserno, T.M., in press. Automatic medical image annotation in ImageCLEF 2007: Overview, results, and discussion. Pattern Recognition Lett. (Special Issue on Medical Image Annotation in ImageCLEF).

Everingham, M., Zisserman, A., Williams, C.K.I., van Gool, L., Allan, M., Bishop, C.M., Chapelle, O., Dalal, N., Deselaers, T., Dorko, G., Duffner, S., Eichhorn, J., Farquhar, J.D.R., Fritz, M., Garcia, C., Griffiths, T., Jurie, F., Keysers, D., Koskela, M., Laaksonen, J., Larlus, D., Leibe, B., Meng, H., Ney, H., Schiele, B., Schmid, C., Seemann, E., Shawe-Taylor, J., Storkey, A., Szedmak, S., Triggs, B., Ulusoy, I., Viitaniemi, V., Zhang, J., 2006. The 2005 pascal visual object classes challenge. In: Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Textual Entailment (PASCAL Workshop 05). Lecture Notes in Artificial Intelligence, vol. 3944, Southampton, UK, pp. 117–176.

Fergus, R., Perona, P., Zisserman, A., 2004. A visual category filter for google images. In: Proc. 8th European Conf. on Computer Vision (ECCV 2004), vol. 1, Prague, Czech Republic, pp. 242–256.

Gass, T., Geissbuhler, A., Müller, H., 2007. Learning a frequency-based weighting for medical image classification. In: Medical Imaging and Medical Informatics (MIMI) 2007, Beijing, China, pp. 137–147.

Kato, T., 1992. Database architecture for content-based image retrieval. In: Jamberdino, A.A., Niblack, W. (Eds.), Image Storage and Retrieval Systems. SPIE Proc., vol. 1662, San Jose, CA, pp. 112–123.

Lehmann, T.M., Schubert, H., Keysers, D., Kohnen, M., Wein, B.B., 2003. The IRMA code for unique classification of medical images. In: Huang, H.K., Ratib, O.M. (Eds.), Medical Imaging 2003: PACS and Integrated Medical Information Systems: Design and Evaluation. SPIE Proc., vol. 5033, San Diego, CA, USA, pp. 440–451.

Lehmann, T.M., Güld, M.O., Deselaers, T., Keysers, D., Schubert, H., Spitzer, K., Ney, H., Wein, B.B., 2005. Automatic categorization of medical images for content-based retrieval and data mining. Comput. Med. Imaging Graphics 29 (2–3), 143–155.

Müller, H., Michoux, N., Bandon, D., Geissbuhler, A., 2004a. A review of content-based image retrieval systems in medicine – clinical benefits and future directions. Internat. J. Med. Inform. 73, 1–23.

Müller, H., Rosset, A., Vallée, J.-P., Geissbuhler, A., 2004b. Comparing feature sets for content-based medical information retrieval. In: Proc. SPIE Internat. Conf. on Medical Imaging, SPIE, vol. 5371, San Diego, CA, USA, pp. 99–109.

Müller, H., Heuberger, J., Geissbuhler, A., 2005. Logo and text removal for medical image retrieval. In: Meinzer, H.-P., Handels, H., Horsch, A., Tolxdorff, T. (Eds.), Springer Informatik aktuell: Proc. Workshop Bildverarbeitung für die Medizin, Springer, Heidelberg, Germany, pp. 35–39.

Müller, H., Pitkanen, M., Zhou, X., Depeursinge, A., Iavindrasana, J., Geissbuhler, A., 2007. KnowARC: Enabling grid networks for the biomedical research community. Healthgrid 2007, Geneva, Switzerland, pp. 261–268.

Niblack, W., Barber, R., Equitz, W., Flickner, M.D., Glasman, E.H., Petkovic, D., Yanker, P., Faloutsos, C., Taubin, G., 1993. QBIC project: Querying images by content, using color, texture, and shape. In: Niblack, W. (Ed.), Storage and Retrieval for Image and Video Databases. SPIE Proc., vol. 1908, pp. 173–187.

Pinz, A., 2005. Object categorization. Found. Trends Comput. Graph. Vis. 1 (4), 255–353.

Qiu, B., 2006. A refined SVM applied in medical image annotation. In: Evaluation of Multilingual and Multi-modal Information Retrieval, Seventh Workshop of the Cross-Language Evaluation Forum (CLEF 2006). Lecture Notes in Computer Science, vol. 4730, Springer, Alicante, Spain, pp. 690–693.

Rui, Y., Huang, T.S., Chang, S.-F., 1999. Image retrieval: Past, present and future. J. Visual Comm. Image Represent. 10, 39–62.

Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A., Jain, R., 2000. Content-based image retrieval at the end of the early years. IEEE Trans. Pattern Anal. Mach. Intell. 22 (12), 1349–1380.

Squire, D.M., Müller, W., Müller, H., Raki, J., 1999. Content-based query of image databases, inspirations from text retrieval: Inverted files, frequency-based weights and relevance feedback. In: The 11th Scandinavian Conference on Image Analysis (SCIA'99), Kangerlussuaq, Greenland, pp. 143–149.

Squire, D.M., Müller, W., Müller, H., Pun, T., 2000. Content-based query of image databases: Inspirations from text retrieval. In: Ersboll, B.K., Johansen, P. (Eds.). Pattern Recognition Lett. 21 (13–14), 1193–1198.

Swain, M.J., Ballard, D.H., 1991. Color indexing. Internat. J. Comput. Vis. 7 (1), 11–32.

Tagare, H.D., Jaffe, C., Duncan, J., 1997. Medical image databases: A content-based retrieval approach. J. Amer. Med. Inform. Assoc. 4 (3), 184–198.

Tommasi, T., Orabona, F., Caputo, B., 2007. CLEF2007 image annotation task: An SVM-based cue integration approach. In: Working Notes of the 2007 CLEF Workshop, Budapest, Hungary.

Vannier, M.W., Staab, E.V., Clarke, L.C., 2002. Medical image archives – present and future. In: Lemke, H.U., Vannier, M.W., Inamura, K., Farman, A.G., Reiber, J.H.C. (Eds.), Proceedings of the International Conference on Computer-Assisted Radiology and Surgery (CARS 2002), Paris, France, pp. 565–576.

Winter, A., Nastar, C., 1999. Differential feature distribution maps for image segmentation and region queries in image databases. In: IEEE Workshop on Content-based Access of Image and Video Libraries (CBAIVL'99), Fort Collins, Colorado, USA, pp. 9–17.

UNCORRECTED PROOF