Image Retrieval and Annotation Using Maximum Entropy

Thomas Deselaers, Tobias Weyand, and Hermann Ney
Human Language Technology and Pattern Recognition
Lehrstuhl für Informatik 6, RWTH Aachen University, Aachen, Germany
surname@informatik.rwth-aachen.de

Abstract

In this work, we present and discuss our participation in the four tasks of the ImageCLEF 2006 Evaluation. In particular, we present a novel approach to learn feature weights in our content-based image retrieval system FIRE. Given a set of training images with known relevance among each other, the retrieval task is reformulated as a classification task and then the weights to combine a set of features are trained discriminatively using the maximum entropy framework. Experimental results for the medical retrieval task show large improvements over heuristically chosen weights. Furthermore the maximum entropy approach is used for the automatic image annotation tasks in combination with a part-based object model. The best results are achieved in the medical and the object annotation task.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.3 Information Search and Retrieval ; I.5 [**Pattern Recognition**]: I.5.4 Applications

Keywords

content-based image retrieval, object recognition, textual information retrieval

1 Introduction

Image retrieval and automatic classification or annotation of images are highly related research fields. Obviously, image retrieval can be "solved" by image annotation straightforwardly: given a database of images, annotate all of them and use textual information retrieval techniques. Multi modal information retrieval, another highly related field allows to use e.g. visual and textual information to retrieve relevant documents. All these tasks have in common that somehow the semantic gap has to be bridged and that therefore large amounts of data have to be processed. Features and descriptors are extracted from the data and these have to be combined to obtain a satisfying solution.

In the domain of feature combination, machine learning algorithms are used quite commonly, among them log-linear models that are discriminatively trained under the maximum entropy criterion are very successful [2]. Maximum entropy, or logistic, models are commonly used in natural language processing [1, 2], data mining [27], and image processing [19, 21, 25].

In this work, we present how the maximum entropy approach can on the one hand be used for object recognition and classification of images and on the other hand for discriminative training of feature weights in an image retrieval system and thus for learning to combine textual information sources with visual information sources in a unified framework. In particular, we describe how we

used maximum entropy training for our submissions to the 2006 ImageCLEF image retrieval and classification/annotation evaluation. The main contribution of this paper is a method to learn feature weights for image retrieval from a given set of queries and relevant documents.

The remainder of this paper is structured as follows: Section 2 describes the retrieval framework, the application of the maximum entropy approach to feature weight training and the experiments we performed for the two image retrieval tasks in ImageCLEF 2006: medical retrieval and photo/ad-hoc retrieval. Section 3 describes the experiments that were performed for the automatic annotation tasks.

2 Retrieval Tasks

ImageCLEF 2006 hosted two independent retrieval tasks: The medical retrieval task [28] and the photo retrieval task [3].

2.1 FIRE – The Flexible Image Retrieval System

For the retrieval tasks the Flexible Image Retrieval Engine (FIRE) developed in our group was used. FIRE is a research image retrieval system that was designed with extensibility in mind and allows to combine various image descriptors and comparison measures easily.

Given a set of positive example images Q^+ and a (possibly empty) set of negative example images Q^- a score $S(Q^+, Q^-, X)$ is calculated for each image X from the database:

$$S(Q^+, Q^-, X) = \sum_{q \in Q^+} S(q, X) + \sum_{q \in Q^-} (1 - S(q, X)). \tag{1}$$

where $S(q, X)$ is the score of database image X with respect to query q and is calculated as $S(q, X) = e^{-\gamma D(q,X)}$ with $\gamma = 1.0$. $D(q, X)$ is a weighted sum of distances calculated as

$$D(q, X) := \sum_{m=1}^{M} w_m \cdot d_m(q_m, X_m). \tag{2}$$

Here, q_m and X_m are the mth feature of the query image q and the database image X, respectively. d_m is the corresponding distance measure and w_m is a weighting coefficient. For each d_m, $\sum_{X \in \mathcal{B}} d_m(Q_m, X_m) = 1$ is enforced by re-normalization.

Given a query (Q^+, Q^-), the images are ranked according to descending score and the K images X with highest scores $S(Q^+, Q^-, X)$ are returned by the retriever.

Weights were chosen heuristically based on experiences from earlier experiments; furthermore we used the weights of our run that performed best in the 2005 ImageCLEF medical retrieval evaluation.

Another way to obtain suitable weights is described in Section 2.3 which requires slight modifications of the decision rule.

2.2 Features

In the following we describe the image features we used in the evaluation. These features are extracted offline from all database images.

Appearance-based Image Features. The most straight-forward approach is to directly use the pixel values of the images as features. For example, the images might be scaled to a common size and compared using the Euclidean distance. In optical character recognition and for medical data improved methods based on image features usually obtain excellent results [20, 23, 24].

In this work, we used 32×32 versions of the images, these were compared using Euclidean distance. It has been observed, that for classification and retrieval of medical radiographs, this method serves as a not-too-bad baseline.

Color Histograms. Color histograms are widely used in image retrieval [5, 13, 29, 31]. Color histograms are one of the most basic approaches and to show performance improvements, image retrieval systems often are compared to a system using only color histograms. The color space is partitioned and for each partition the pixels with a color within its range are counted, resulting in a representation of the relative frequencies of the occurring colors. In accordance with [29], we use the Jeffrey divergence to compare histograms.

Tamura Features. In [32] the authors propose six texture features corresponding to human visual perception: *coarseness*, *contrast*, *directionality*, *line-likeness*, *regularity*, and *roughness*. From experiments testing the significance of these features with respect to human perception, it was concluded that the first three features are very important. Thus in our experiments we use coarseness, contrast, and directionality to create a histogram describing the texture [5] and compare these histograms using the Jeffrey divergence [29]. In the QBIC system [13] histograms of these features are used as well.

Global Texture Descriptor. In [5] a texture feature consisting of several parts is described: *Fractal dimension* measures the roughness or the crinkliness of a surface. In this work the fractal dimension is calculated using the reticular cell counting method [16]. *Coarseness* characterizes the grain size of an image. Here it is calculated depending on the variance of the image. *Entropy* is used as a measure of disorderedness or information content in an image. The *Spatial gray-level difference statistics* (SGLD) describes the brightness relationship of pixels within neighborhoods. It is also known as co-occurrence matrix analysis [17]. . The *Circular Moran autocorrelation function* measures the roughness of the texture. For the calculation a set of autocorrelation functions is used [15].

Invariant Feature Histograms. A feature is called invariant with respect to certain transformations if it does not change when these transformations are applied to the image. The transformations considered here are translation, rotation, and scaling. In this work, invariant feature histograms as presented in [30] are used. These features are based on the idea of constructing features invariant with respect to certain transformations by integration over all considered transformations. The resulting histograms are compared using the Jeffrey divergence [29]. Previous experiments have shown that the characteristics of invariant feature histograms and color histograms are very similar and that invariant feature histograms often outperform color histograms [7]. Thus, in this work color histograms are not used.

Patch Histograms. In object recognition and detection currently the assumption that objects consist of parts that can be modelled independently is very common, which led to a wide variety of bag-of-features approaches [11, 8, 26].

Here we follow this approach to generate histograms of image patches for image retrieval. The creation is a 3-step procedure:

1. in the first phase, sub-images are extracted from all training images and the dimensionality is reduced to 40 dimensions using PCA transformation.

2. in the second phase, the sub-images of all training images are jointly clustered using the EM algorithm for Gaussian mixtures to form 2000-8000 clusters.

3. in the third phase, all information about each sub-image is discarded except its closest cluster center. Then, for each image a histogram over the cluster identifiers of the respective patches is created, thus effectively coding which "visual words" from the code-book occur in the image.

2.3 Maximum Entropy Training for Image Retrieval

We propose a novel method based on maximum entropy training using the generalized iterative scaling algorithm (GIS) to obtain feature weightings tuned toward a specific task. The maximum entropy approach is promising here, because it is ideally suited to combine features of different types and it yields good results in other areas like natural language processing [2] and image recognition [21, 19]. In [19], the maximum entropy approach is used for automatic image annotation. The authors partition the image into rectangular parts and consider these patches as "image terms" similar to the usage of words in [2].

We consider the problem of image retrieval to be a classification problem. Given the query image, the images from the database have to be classified to be either relevant (denoted by $\oplus$) or irrelevant (denoted by $\ominus$). As classification method we choose log-linear models that are trained using the maximum entropy criterion and the GIS algorithm.

As features f_i for the log-linear models we choose the distances between the m-th feature of the query image Q and the database image X:

$$f_i(Q, X) \quad := \quad d_i(Q_i, X_i).$$

To allow for prior probabilities, we include a constant feature $f_{i=0}(Q, X) = 1$. Then, the score is replaced by the posterior probability for class $\oplus$:

$$\begin{aligned} S(Q, X) &:= p(\oplus|Q, X) \\ &= \frac{\exp\left[\sum_i \lambda_{\oplus i} f_i(Q, X)\right]}{\sum_{k \in \{\oplus, \ominus\}} \exp\left[\sum_i \lambda_{ki} f_i(Q, X)\right]} \end{aligned} \tag{3}$$

Given these scores, we return the K images from the database that have the highest score $S(Q, X)$, i.e. the K images that are most likely to be relevant according to the classifier. Note that here in comparison to the score calculation from Equation (1), the w_i are replaced by the $\lambda_{\oplus i}$ and the $\lambda_{\ominus i}$ and an additional renormalization factor is introduced to assure that the probabilities sum up to one. Alternatively, Eq. 3 can easily be transformed to be of the form of Eq. 1 and the w_i can be expressed as a function of $\lambda_{\oplus i}$ and $\lambda_{\ominus i}$. In addition to considering the first order features alone as they are described above, we propose to use supplementary second order features (i.e. products of distances) as this usually yields superior performance on other tasks. Given a query image Q and a database image X we use the following set of features:

$$\begin{aligned} f_i(Q, X) &:= d_i(Q_i, X_i) \\ f_{i,j}(Q, X) &:= d_i(Q_i, X_i) \cdot d_j(Q_j, X_j), \quad i \geq j, \end{aligned}$$

again including the constant feature $f_{i=0}(Q, X) = 1$ to allow for prior probabilities. The increased number of features results in more parameters to be trained. In earlier experiments, features of higher degree have been tested and not found to improve the results.

In the training process, the values of the λ_{ki} are optimized. A sufficiently large amount of training data is necessary to do so. We are given the database $\mathcal{T} = \{X_1, \ldots, X_N\}$ of training images with known relevances. For each image X_n we are given a set $R_n = \{Y \mid Y \in \mathcal{T}$ is relevant, if X_n is the query.$\}$.

Because we want to classify the relation between images into the two categories "relevant" or "irrelevant" on the basis of the distances between their features, we choose the following way to derive the training data for the GIS algorithm: The distance vectors $D(X_n, X_m) = (d_1(X_{n1}, X_{m1}), \ldots, d_I(X_{nI}, X_{mI}))$ are calculated for each pair of images $(X_n, X_m) \in \mathcal{T} \times \mathcal{T}$. That is, we obtain N distance vectors for each of the images X_n. These distance vectors are then labeled according to the relevances: Those $D(X_n, X_m)$ where X_m is relevant with respect to X_n, i.e. $X_m \in R_n$, are labeled $\oplus$ (relevant) and the remaining ones are labeled with the class label $\ominus$ (irrelevant).

Given these N^2 distance vectors and their classification into "relevant" and "irrelevant" we train the λ_{ki} of the log-linear model from Eq. (3) using the GIS algorithm.

The GIS algorithm proceeds as follows to determine the free parameters of the model (3). First an initial parameter set $\Lambda^{(0)} = \{\lambda_{ki}^{(0)}\}$ is chosen, and then for each iteration $t = 1, \ldots, T$ the parameters are updated according to

$$\begin{aligned}
\lambda_{ki}^{(t)} &= \lambda_{ki}^{(t-1)} + \Delta\lambda_{ki}^{(t)} \\
&= \lambda_{ki}^{(t-1)} + \frac{1}{F} \log \frac{N_{ki}}{Q_{ki}^{(t)}}, \\
Q_{ki}^{(t)} &:= \sum_{X_n, X_m} p_{\Lambda^{(t)}}(k|X_n, X_m) f_i(X_n, X_m), \\
N_{\oplus i} &:= \sum_{X_n, X_m \in R_n} f_i(X_n, X_m) \\
N_{\ominus i} &:= \sum_{X_n, X_m \notin R_n} f_i(X_n, X_m)
\end{aligned}$$

Here, F is a constant that depends on the training data. In some cases, this method is problematic due to the high computational demands. Here, the number of parameters to be estimated is small, i.e. $2I+1$, thus performance is not a problem. Due to the low computational demands, this method can also be used to incorporate relevance estimates gathered from user interaction. To do so, the current state of the classifier can be used as a starting point for further training iterations with the training set enlarged by the newly gathered data. This process can e.g. be performed once a day. As the training is performed in an offline manner, the speed of the image retrieval engine is hardly decreased because the calculation of Equation (3) takes barely longer than the calculation of Equation (1).

2.4 Medical Retrieval Task

We submitted nine runs to the medical retrieval task [28], one of these using only text, three using only visual information, and five using visual and textual information. For one of the combined runs we used the above-described maximum entropy training method. To determine the weights, we used the queries and their qrels from last year's medical retrieval task as training data. Table 1 gives an overview of the runs we submitted to the medical retrieval task and the results obtained.

In Figure 1 the trained feature weights are visualized after different numbers of maximum entropy training iterations. It can clearly be seen that after 500 iterations the weights hardly differ from uniform weighting and that thus not enough training iterations were performed. After 5000 iterations, there is a clear gain in performance (cp. Table 1) and the weights are not uniform any more. For example, the weight for feature 1 (English text) has the highest weight. With more iterations, the differences between the particular weights become bigger; after 10.000 iterations no additional gain in performance is yielded anymore.

2.5 Photo/Ad-Hoc Retrieval Task

For the photo- and the ad-hoc retrieval task the newly created IAPR TC-12 database [14] was used, which currently consists of 20,000 general photographs, mainly from a vacation domain. For each of the images a German and an English description exists. The task is described in detail in [3].

Two tasks were defined on this dataset: An ad-hoc task of 60 queries of different semantic and syntactic difficulty, and a photo task of 30 queries, which was based on a subset aiming to investigate the possibilities of purely visual retrieval. Therefore, some semantic constraints were removed from the queries. All queries were formulated by a short textual description and three positive example images.

Due to short time, we were unable to tune any parameters and just chose to submit two purely visual, full-automatic runs to both of these tasks.

Table 1: Summary of our runs submitted to the medical retrieval task. The numbers give the weights (empty means 0) of the features in the experiments and the columns denote: *En*: English text, *Fr*: French text, *Ge*: German text, *CH*: color histogram, *GH*: gray histogram, *GTF*: global texture feature, *IH*: invariant feature histogram, *TH*: Tamura Texture Feature histogram, *TN*: 32x32 thumbnail, *PH*: patch histogram. The first group of experiments uses only textual information, the second group uses only visual information, the third group uses textual and visual information, and the last group both types of information and the weights are trained using the maximum entropy approach. The last column gives the results of the evaluation. The last three lines are unsubmitted runs that were performed after the evaluation ended.

run-tag	En	Fr	Ge	CH	GH	GTF	IFH	TH	TN	PH	MAP
En	1										0.15
SimpleUni				1	1	1	1	1	1		0.05
Patch										1	0.04
IfhTamThu							2	2	1		0.05
EnIfhTamThu	1						2	2	1		0.09
EnFrGeIfhTamThu	2	1	1				2	2	1		0.13
EnFrGePatches	2	1	1							1	0.17
EnFrGePatches2	2	1	1							2	0.16
ME [500 iterations]	*	*	*	*	*	*	*	*	*	0	0.07
ME [5000 iterations]	*	*	*	*	*	*	*	*	*	0	0.15
ME [10000 iterations]	*	*	*	*	*	*	*	*	*	0	0.18
ME [20000 iterations]	*	*	*	*	*	*	*	*	*	0	0.18

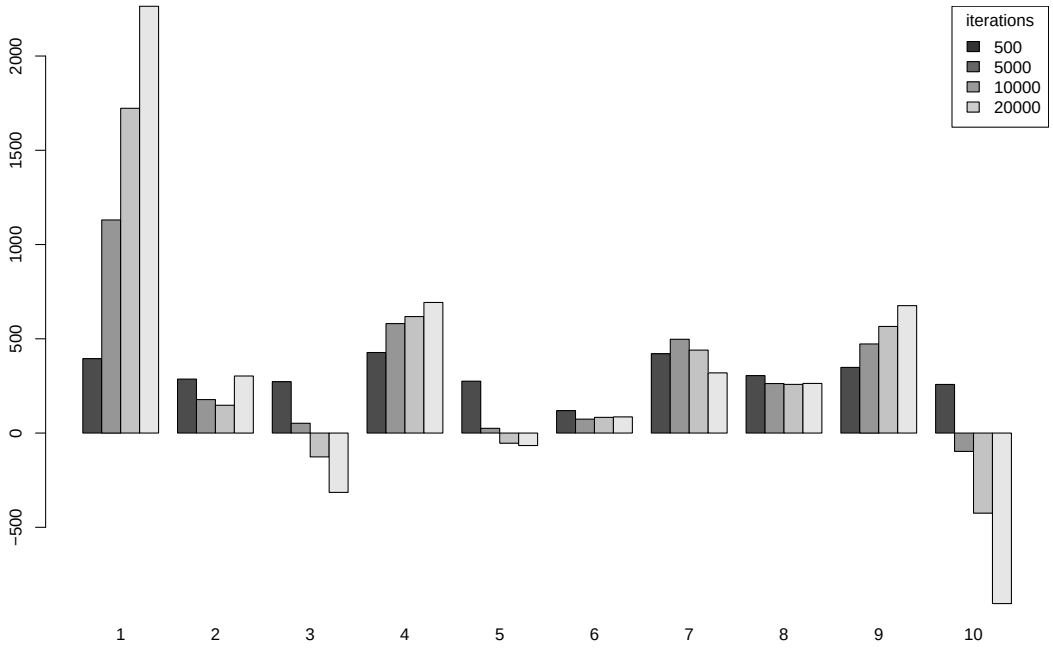

Figure 1: Trained weights for the medical retrieval task after different numbers of iterations in the maximum entropy training. On the x-axis, the features are given in the same order as in Table 1 and on the y-axis $\lambda_{\oplus i}$ - $\lambda_{\ominus i}$ is given.

Table 2: Results from the AdHoc and the Photo task.

(a) Results from the adhoc retrieval task with 60 queries in the category "visual only, full automatic, no user interaction".

task	run-tag	map	rank
RWTHi6	IFHTAM	0.06	1
RWTHi6	PatchHisto	0.05	2
CEA	mPHic	0.05	3
CEA	2mPHit	0.04	4
IPAL	LSA	0.03	5
IPAL	MF	0.02	6

(b) Results from the photo retrieval task with 30 queries. All submissions to this task were submitted as full automatic, visual only submissions without user feedback.

task	run-tag	map	rank
RWTHi6	IFHTAM	0.11	1
RWTHi6	PatchHisto	0.08	2
IPAL	LSA3	0.07	3
IPAL	LSA2	0.06	5
IPAL	LSA1	0.06	4
IPAL	MF	0.04	6

For the runs entitled `IFHTAM`, we used a combination of invariant feature histograms and Tamura texture histograms. Both histograms are combined by Jeffrey divergence and the invariant feature histograms are weighted by a factor of 2. This combination has been seen to be a very effective combination of features for databases of general photographs like for example the Corel database [7]. For the runs entitled `PatchHisto` we used histograms of vector-quantized image patches with 2048 bins.

In Table 2 we summarize the outcomes of the two tasks using the IAPR TC-12 database. The overal MAP values are rather low, but the combination of invariant feature histograms and Tamura texture features clearly outperforms all competing methods.

3 Automatic Annotation Tasks

In ImageCLEF 2006, two automatic annotation tasks were arranged. One dealing with the automatic classification of medical radiographs [28] and one tackling the problem of automatic classification of everyday objects like backpacks, clocks, and plates [3]. The medical annotation task was very similar to last year's task, but the number of images was slightly raised and the number of classes was raised from 57 to 116. The automatic annotation task was somehow similar to the PASCAL visual object classes challenge [12]. Here, 20 classes had to be discriminated at once. The following sections describe the methods we applied to these classification tasks and the experiments we performed.

The task of the medical automatic annotation task and the object annotation tasks are very similar, but differ in some critical aspects:

- Both tasks provide a relatively large training set and a disjunct test set. Thus, in both cases it is possible to learn a relatively reliable model for the training data (this is somewhat proven for the medical annotation task, and below we also show this for the object annotation task).
- Both tasks are multi-class/one object per image classification tasks. Here they differ from the PASCAL visual classes challenge which addresses a set of object vs. non object tasks where several objects (of equal or unequal type) may be contained in an image.
- The medical annotation task has only gray scale images, whereas the object annotation task has mainly color images. This is probably most relevant for the selection of descriptors.
- The images from the test and the training set are from the same distribution for the medical task, whereas for the object annotation task, the training images are rather clutter-free and the test images contain a significant amount of clutter. This is probably relevant and should be addressed when developing methods for the object annotation task. Unfortunately, our models currently do not address this issue which probably has a significant impact on the results.

3.1 Image Distortion Model

The image distortion model [23, 20] is a zeroth-order image deformation model to compare images pixel-wise. Here, classification is done using the nearest neighbor decision rule: to classify an image, it is compared to all training images in the database and the class of the most similar image is chosen. To compare images, the Euclidean distance can be seen as a very basic baseline, and in earlier works it was shown that image deformation models are a suitable way to improve classification performance significantly e.g. for medical radiographs and for optical character recognition [22, 23]. Here we allow each pixel of the database images to be aligned to the pixels from a 5×5 neighborhood from the image to be classified taking into account the local context from a 3×3 Sobel neighborhood.

This method is of particular interest as it outperformed all other methods in automatic annotation task of ImageCLEF 2005 [4].

3.2 Sparse Patch Histograms & Discriminative Classification

This approach is based on the widely adopted assumption that objects in images can be represented as a set of loosely coupled parts. In contrast to former models[8, 9], this method can cope with an arbitrary number of object parts. Here, the object parts are modelled by image patches that are extracted at each position and then efficiently stored in a histogram. In addition to the patch appearance, the positions of the extracted patches are considered and provide a significant increase in the recognition performance.

Using this method, we create sparse histograms of 65536 ($2^{16} = 8^4$) bins, which can either be classified using the nearest neighbor rule and a suitable histogram comparison measure or a discriminative model can be trained for classification. Here, we used a support vector machine with a histogram intersection kernel and a discriminatively trained log-linear maximum entropy model.

A detailed description of the method is given in [6].

3.3 Patch Histograms & Maximum Entropy Classification

In this approach, we use the histograms of image patches as described in Section 2.2 and maximum entropy training [8, 9].

This method has performed very well in the 2005 annotation task of ImageCLEF [4] and in the 2005 and 2006 visual object classes challenges of PASCAL [12].

3.4 Medical Automatic Annotation Task

We submitted three runs to the medical automatic annotation task [28]: one run using the image distortion model `RWTHi6-IDM`, with exactly the same settings as the according run from last year, which clearly outperformed all competing methods [10] and two other runs based on sparse histograms of image patches [6], where we used a discriminatively trained log-linear maximum entropy model (`RWTHi6-SHME`) and support vector machines with a histogram intersection kernel (`RWTHi6-SHSVM`) respectively. Due to time constraints we were unable to submit the method described in Section 3.3, but we give comparison results here.

Results. The results of the evaluation are given in detail in the overview paper. Table 3 gives an overview of the results and it can be seen that the runs using the discriminative classifier for the histograms clearly outperform the image distortion model and that in summary our method performed very good on the task.

The table also gives the result for the method presented in [8, 9], which we were unable to submit in time. Interestingly, the results of this method are not very good although it is strongly related to the sparse histogram method.

Table 3: An overview of the results of the medical automatic annotation task. The first part gives our results (including the error rate of an unsubmitted method for comparison to the results of last year); the second part gives results from other groups that are interesting for comparison

rank	run-tag	error rate[%]
1	RWTHi6 SHME	16.2
2	RWTHi6 SHSVM	16.7
11	RWTHi6 IDM	20.5
-	RWTHi6 - [8]	22.4
2	UFR ns1000-20x20x10	16.7
4	MedIC-CISMef local+global-PCA335	17.2
12	RWTHmi rwthmi	21.5
23	ULG sysmod-random-subwindows-ex	29.0

Table 4: Results from the object annotation task.

rank	Group ID	run-tag	Error rate
1	RWTHi6	SHME	77.3
2	RWTHi6	PatchHisto	80.2
3	CINDI	SVM-Product	83.2
4	CINDI	SVM-EHD	85.0
5	CINDI	SVM-SUM	85.2
6	CINDI	Fusion-knn	87.1
7	DEU-CS	edgehistogr-centroid	88.2
8	DEU-CS	colorlayout-centroid	93.2

Interesting conclusions can be drawn when comparing our results to the results of other groups: the medical informatics division of the RWTH Aachen University (`RWTHmi`) method uses the image distortion model as a significant part of their method and combines it with various other global image descriptors, which seem not to help the classification. The ULG run is interesting, as it was one of the best performing methods from last year and is also closely related to our unsubmitted run: it uses sparsely extracted sub-images and a discriminative classification framework. The runs of University Freiburg (UFR) and INSA Rouen (MedIC) are included for comparison with the best results from other groups. A more detailed overview of the results can be found in the track overview paper [3, 28].

Concluding it can be seen that the approach, where local image descriptors were extracted at every position in the image, outperformed our other approaches, and that probably the modelling of absolute positions is suitable for radiograph recognition. This is because it seems to be a suitable assumption that radiographs are taken under controlled conditions and that thus the geometric layout of images showing the same body region can be assumed to be very similar.

3.5 Object Annotation Task

We submitted two runs to this task [3], one using the method with vector quantized histograms described in Section 3.3 (run-tag `PatchHisto`) and the other using the method with sparse histograms as described in Section 3.2 (run-tag `SHME`). These two methods were also used in the PASCAL visual object classes challenge 2006. The third method [18] we submitted to the PASCAL challenge could not be applied to this task due to time and memory constraints.

Results. Table 4 gives the results of the object annotation task. On the average, the error rates are very high. The best two results of 77.3% and 80.2% were achieved with our discriminative classification method. For the submissions of the CINDI group, support vector machines were used and the DEU-CS group used a nearest neighbor classification. Obviously, the results are not

satisfactory and large improvements should be possible.

4 Conclusion and Outlook

We have presented our efforts for the ImageCLEF 2006 image retrieval and annotation challenge. In particular, we presented a discriminative method to train weights to combine features in our image retrieval system. This method allows to find weights that clearly outperform a setting with feature weights chosen from experiences from earlier experiments and thus allows us to obtain better results than our best old system. We give an interpretation of the trained weights and show the development of the weights given different number of training iterations.

The maximum entropy principle was futhermore used for automatic image annotation and very good results were obtained.

Acknowledgement

This work was partially funded by the DFG (Deutsche Forschungsgemeinschaft) under contract NE-572/6.

References

[1] O. Bender, F. Och, and H. Ney. Maximum Entropy Models for Named Entity Recognition. In *7th Conference on Computational Natural Language Learning*, Edmonton, Canada, pages 148–152, May 2003.

[2] A. L. Berger, S. A. Della Pietra, and V. J. Della Pietra. A Maximum Entropy Approach to Natural Language Processing. *Computational Linguistics*, 22(1):39–72, March 1996.

[3] P. Clough, M. Grubinger, T. Deselaers, A. Hanbury, and H. Mller. Overview of the ImageCLEF 2006 photographic retrieval and object annotation tasks. In *CLEF working notes*, Alicante, Spain, September 2006.

[4] P. Clough, H. Mueller, T. Deselaers, M. Grubinger, T. Lehmann, J. Jensen, and W. Hersh. The CLEF 2005 Cross-Language Image Retrieval Track. In *Workshop of the Cross–Language Evaluation Forum (CLEF 2005)*, Lecture Notes in Computer Science, Vienna, Austria, page in press, September 2005.

[5] T. Deselaers. Features for Image Retrieval. Diploma thesis, Human Language Technology and Pattern Recognition Group, RWTH Aachen University, Aachen, Germany, December 2003.

[6] T. Deselaers, A. Hegerath, D. Keysers, and H. Ney. Sparse Patch-Histograms for Object Classification in Cluttered Images. In *DAGM 2006, Pattern Recognition, 26th DAGM Symposium*, volume 4174 of *Lecture Notes in Computer Science*, Berlin, Germany, pages 202–211, September 2006.

[7] T. Deselaers, D. Keysers, and H. Ney. Features for Image Retrieval – A Quantitative Comparison. In *DAGM 2004, Pattern Recognition, 26th DAGM Symposium*, number 3175 in Lecture Notes in Computer Science, Tbingen, Germany, pages 228–236, September 2004.

[8] T. Deselaers, D. Keysers, and H. Ney. Discriminative Training for Object Recognition using Image Patches. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, San Diego, CA, pages 157–162, June 2005.

[9] T. Deselaers, D. Keysers, and H. Ney. Improving a Discriminative Approach to Object Recognition using Image Patches. In *DAGM 2005, Pattern Recognition, 26th DAGM Symposium*, number 3663 in Lecture Notes in Computer Science, Vienna, Austria, pages 326–333, August 2005.

[10] T. Deselaers, T. Weyand, D. Keysers, W. Macherey, and H. Ney. FIRE in ImageCLEF 2005: Combining Content-based Image Retrieval with Textual Information Retrieval. In *Workshop of the Cross–Language Evaluation Forum (CLEF 2005)*, Lecture Notes in Computer Science, Vienna, Austria, page in press, September 2005.

[11] G. Dork and C. Schmid. Object class recognition using discriminative local features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, submitted 2004.

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

[13] C. Faloutsos, R. Barber, M. Flickner, J. Hafner, W. Niblack, D. Petkovic, and W. Equitz. Efficient and Effective Querying by Image Content. *Journal of Intelligent Information Systems*, 3(3/4):231–262, July 1994.

[14] M. Grubinger, P. Clough, H. Mller, and T. Deselaers. The IAPR Benchmark: A New Evaluation Resource for Visual Information Systems. In *LREC 06 OntoImage 2006: Language Resources for Content-Based Image Retrieval*, Genoa, Italy, page in press, May 2006.

[15] Z. Q. Gu, C. N. Duncan, E. Renshaw, M. A. Mugglestone, C. F. N. Cowan, and P. M. Grant. Comparison of Techniques for Measuring Cloud Texture in Remotely Sensed Satellite Meteorological Image Data. *Radar and Signal Processing*, 136(5):236–248, October 1989.

[16] P. Habercker. *Praxis der Digitalen Bildverarbeitung und Mustererkennung*. Carl Hanser Verlag, Mnchen, Wien, 1995.

[17] R. M. Haralick, B. Shanmugam, and I. Dinstein. Texture Features for Image Classification. *IEEE Transactions on Systems, Man, and Cybernetics*, 3(6):610–621, November 1973.

[18] A. Hegerath, T. Deselaers, and H. Ney. Patch-based Object Recognition Using Discriminatively Trained Gaussian Mixtures. In *17th British Machine Vision Conference (BMVC06)*, Edinburgh, UK, page in press, September 2006.

[19] J. Jeon and R. Manmatha. Using Maximum Entropy for Automatic Image Annotation. In *Proceedings of the 3rd International Conference on Image and Video Retrieval*, pages 24–32, 2004.

[20] D. Keysers, C. Gollan, and H. Ney. Classification of Medical Images using Non-linear Distortion Models. In *Bildverarbeitung für die Medizin*, Berlin, Germany, pages 366–370, March 2004.

[21] D. Keysers, F.-J. Och, and H. Ney. Maximum Entropy and Gaussian Models for Image Object Recognition. In *Pattern Recognition, 24th DAGM Symposium*, Zürich, Switzerland, pages 498–506, September 2002.

[22] D. Keysers, C. Gollan, and H. Ney. Classification of Medical Images using Non-linear Distortion Models. In *Proc. BVM 2004, Bildverarbeitung für die Medizin*, Berlin, Germany, pages 366–370, March 2004.

[23] D. Keysers, C. Gollan, and H. Ney. Local Context in Non-linear Deformation Models for Handwritten Character Recognition. In *International Conference on Pattern Recognition*, volume 4, Cambridge, UK, pages 511–514, August 2004.

[24] D. Keysers, W. Macherey, H. Ney, and J. Dahmen. Adaptation in Statistical Pattern Recognition using Tangent Vectors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(2):269–274, February 2004.

[25] S. Lazebnik, C. Schmid, and J. Ponce. A Maximum Entropy Framework for PArt-Based Texture and Object Recognition. In *IEEE International Conference on Computer Vision (ICCV 05)*, volume 1, Bejing, China, pages 832–838, October 2005.

[26] R. Mare, P. Geurts, J. Piater, and L. Wehenkel. Random Subwindows for Robust Image Classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 34–40, 2005.

[27] A. Mauser, I. Bezrukov, T. Deselaers, and D. Keysers. Predicting Customer Behavior using Naive Bayes and Maximum Entropy – Winning the Data-Mining-Cup 2004. In *Informatiktage 2005 der Gesellschaft fr Informatik*, St. Augustin, Germany, page in press, April 2005.

[28] H. Mller, 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, September 2006.

[29] J. Puzicha, Y. Rubner, C. Tomasi, and J. Buhmann. Empirical Evaluation of Dissimilarity Measures for Color and Texture. In *International Conference on Computer Vision*, volume 2, Corfu, Greece, pages 1165–1173, September 1999.

[30] S. Siggelkow. *Feature Histograms for Content-Based Image Retrieval*. PhD thesis, University of Freiburg, Institute for Computer Science, Freiburg, Germany, 2002.

[31] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-Based Image Retrieval at the End of the Early Years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1349–1380, December 2000.

[32] H. Tamura, S. Mori, and T. Yamawaki. Textural Features Corresponding to Visual Perception. *IEEE Transaction on Systems, Man, and Cybernetics*, 8(6):460–472, June 1978.