Image Retrieval and Annotation Using Maximum Entropy

Thomas Deselaers, Tobias Weyand, and Hermann Ney

Human Language Technology and Pattern Recognition Group
RWTH Aachen University, Aachen, Germany
deselaers@cs.rwth-aachen.de

Abstract. 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. Using our object classification methods, we obtained the best results in the medical and in the object annotation task.

1 Introduction

An important task for obtaining satisfying results in content-based image retrieval and annotation is the combination and weighting of descriptors extracted from the images. Commonly, machine learning algorithms are used to learn a good way of combining descriptors from some training data. Discriminative log-linear models [1] have been successfully used for this purpose in different scenarios: natural language processing [2, 1], data mining [3], and image recognition [4, 5].

Here, we present, how we used the maximum entropy approach to train feature weights in our content-based image retrieval system for the medical image retrieval task in ImageCLEF 2006. We furthermore present our results of the ImageCLEF 2006 photographic retrieval task, and of the ImageCLEF 2006 medical annotation and object recognition tasks. In the annotation and recognition tasks, the maximum entropy method was also extensively used and excellent results were obtained.

2 Retrieval Tasks

ImageCLEF 2006 hosted two independent retrieval tasks: The medical retrieval task [6] and the photo retrieval task [7].

For the retrieval tasks the FIRE image retrieval system[1] 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. It also allows for combining textual information and meta data information with content-based queries.

2.1 FIRE – The Flexible Image Retrieval System

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.

The selection of the weights w_m is a critical step which has so far been done heuristically based on experiences from earlier experiments. However, in the next section we describe, how the maximum entropy framework can be used to learn weights from training data.

2.2 Maximum Entropy Training for Image Retrieval

To obtain suitable feature weights, the maximum entropy approach is promising, because it is ideally suited to combine features of different types and it leads to good results in other areas as mentioned above.

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.

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

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:

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

To allow for prior probabilities, we include a constant feature $f_{i=0}(Q, X) = 1$. Then, the scores $S(q, X)$ from Eq. (1) are replaced by the posterior probability for class $\oplus$ and the ranking and combination of several query images is done as before:

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

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 only the first order features 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 experiments in other domains, features of higher degree have been tested and not found to improve the results, and thus we did not try higher order features.

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.

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

2.3 Medical Retrieval Task

We submitted nine runs to the medical retrieval task [6], 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.

2.4 Photo/Ad-Hoc Retrieval Task

For the photo- and the ad-hoc retrieval task the newly created IAPR TC-12 database [8] was used, which currently consists of 20,000 general photographs,

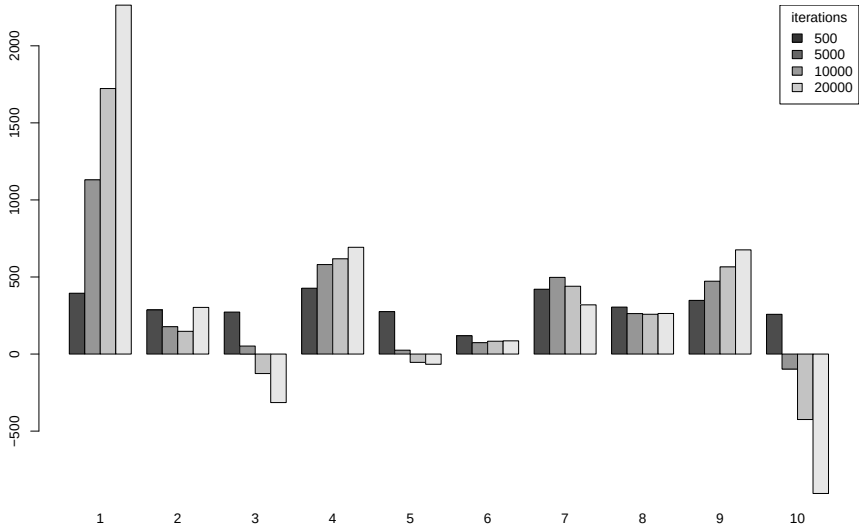

Fig. 1. Trained weights for the medical retrieval task after different numbers of iterations in the maximum entropy training. On the x-axis, the features are given in the same order as in Table 1 and on the y-axis $\lambda_{\oplus i}$ - $\lambda_{\ominus i}$ is given.

mainly from a vacation domain. For each of the images a German and an English description exists. The tasks are described in detail in [7].

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 these tasks.

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.

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 [9].

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 `PatchHisto` we used histograms of vector-quantized image patches with 2048 bins.

All runs we submitted were top-ranked in the category "visual retrieval, no user interaction". A detailed analysis of the results is given in [7].

3 Automatic Annotation Tasks

The following sections describe the methods we applied to the automatic annotation tasks and the experiments we performed.

3.1 Image Distortion Model

The image distortion model [10, 11] 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 [11, 10]. 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 [12].

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[13], 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 [14].

3.3 Patch Histograms & Maximum Entropy Classification

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 [15, 13].

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.

These histograms are then classified using the maximum entropy approach [13].

3.4 Medical Automatic Annotation Task

We submitted three runs to the medical automatic annotation task [6]: 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 [16] and two other runs based on sparse histograms of image patches [14], 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 [6]. Table 3 gives an overview on our submissions and the best competing runs and it can be seen that the runs using the discriminative classifier for the

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 - [13]	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

histograms clearly outperform the image distortion model and that in summary our method performed very good on the task.

Concluding it can be seen that the approach, where local image descriptors were extracted at every position in the image, outperformed our other approaches. Probably the modelling of absolute positions is suitable for radiograph recognition, 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 [7], 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 [17] 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 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

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

find weights that clearly outperform a setting with feature weights chosen from experiences from earlier experiments and thus allowed us to obtain better results than our best old system. The trained features weights can be interpreted to see which features are most important for a given task and this effect is smoothly achieved by an iterative training procedure.

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. Berger, A.L., Della Pietra, S.A., Della Pietra, V.J.: A maximum entropy approach to natural language processing. Computational Linguistics **22** (1996) 39–72
2. Bender, O., Och, F., Ney, H.: Maximum entropy models for named entity recognition. In: 7th Conference on Computational Natural Language Learning, Edmonton, Canada (2003) 148–152
3. Mauser, A., Bezrukov, I., Deselaers, T., Keysers, D.: Predicting customer behavior using naive bayes and maximum entropy – winning the data-mining-cup 2004. In: Informatiktage 2005 der Gesellschaft für Informatik, St. Augustin, Germany (2005) in press
4. Keysers, D., Och, F.J., Ney, H.: Maximum entropy and Gaussian models for image object recognition. In: Pattern Recognition, 24th DAGM Symposium, Zürich, Switzerland (2002) 498–506
5. Lazebnik, S., Schmid, C., Ponce, J.: A maximum entropy framework for part-based texture and object recognition. In: IEEE International Conference on Computer Vision (ICCV 05). Volume 1., Bejing, China (2005) 832–838
6. Müller, H., Deselaers, T., Lehmann, T., Clough, P., Hersh, W.: Overview of the imageclefmed 2006 medical retrieval and annotation tasks. In: Evaluation of Multilingual and Multi-modal Information Retrieval – Seventh Workshop of the Cross-Language Evaluation Forum, CLEF 2006. LNCS, Alicante, Spain (2006) to appear

7. Clough, P., Grubinger, M., Deselaers, T., Hanbury, A., Müller, H.: Overview of the imageclef 2006 photographic retrieval and object annotation tasks. In: Evaluation of Multilingual and Multi-modal Information Retrieval – Seventh Workshop of the Cross-Language Evaluation Forum, CLEF 2006. LNCS, Alicante, Spain (2006) to appear
8. Grubinger, M., Clough, P., Müller, H., Deselaers, T.: 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 (2006) in press
9. Deselaers, T., Keysers, D., Ney, H.: Features for image retrieval – a quantitative comparison. In: DAGM 2004, Pattern Recognition, 26th DAGM Symposium. Number 3175 in Lecture Notes in Computer Science, Tübingen, Germany (2004) 228–236
10. Keysers, D., Gollan, C., Ney, H.: Local context in non-linear deformation models for handwritten character recognition. In: International Conference on Pattern Recognition. Volume 4., Cambridge, UK (2004) 511–514
11. Keysers, D., Gollan, C., Ney, H.: Classification of medical images using non-linear distortion models. In: Proc. BVM 2004, Bildverarbeitung für die Medizin, Berlin, Germany (2004) 366–370
12. Clough, P., Mueller, H., Deselaers, T., Grubinger, M., Lehmann, T., Jensen, J., Hersh, W.: 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 (2005) in press
13. Deselaers, T., Keysers, D., Ney, H.: Discriminative training for object recognition using image patches. In: IEEE Conference on Computer Vision and Pattern Recognition. Volume 2., San Diego, CA (2005) 157–162
14. Deselaers, T., Hegerath, A., Keysers, D., Ney, H.: 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 (2006) 202–211
15. Dorkó, G., Schmid, C.: Object class recognition using discriminative local features. IEEE Transactions on Pattern Analysis and Machine Intelligence (submitted 2004)
16. Deselaers, T., Weyand, T., Keysers, D., Macherey, W., Ney, H.: 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 (2005) in press
17. Hegerath, A., Deselaers, T., Ney, H.: Patch-based object recognition using discriminatively trained gaussian mixtures. In: 17th British Machine Vision Conference (BMVC06). Volume 2., Edinburgh, UK (2006) 519–528