Deformations, Patches, and Discriminative Models for Automatic Annotation of Medical Radiographs

Thomas Deselaers *, Hermann Ney

RWTH Aachen University, Computer Science Department,
Human Language Technology and Pattern Recognition Group, Aachen, Germany

Abstract

In this paper, we describe three different methods for the classification and annotation of medical radiographs. The methods were applied in the medical image annotation tasks of ImageCLEF in 2005, 2006, and 2007. Image annotation can be used to access and find images in a database using textual queries when no textual image description is available. One of the methods is a non-linear model taking into account local image deformations to compare images which are then classified using the nearest neighbour decision rule. The other two methods use local image descriptors for a *bag-of-features* approach. The bags of local image features are classified using discriminative classifiers. Our methods performed best in the 2005 and 2006 evaluations and second best in 2007.

Key words: medical image annotation, local features, bag of features, image deformation model

1 Introduction

Automatic annotation of images is a technique to provide textual labels for images which are not labelled and thus can be used to access images using textual queries. In the medical domain huge amounts of images are produced daily and the manual creation of textual labels is expensive and error prone.

* Corresponding author.
Email addresses: deselaers@cs.rwth-aachen.de (Thomas Deselaers), ney@cs.rwth-aachen.de (Hermann Ney).
URL: www-i6.informatik.rwth-aachen.de/~deselaers (Thomas Deselaers).

The automatic medical image annotation task of ImageCLEF, as it was established in 2005, offered a great opportunity to compare our method to other state of the art approaches. The method we applied in 2005 is a non-linear deformation model accounting for local image deformations. The method was originally developed for optical character recognition by Keysers et al. (2007) but it was already observed that it could be applied in other areas such as recognition of medical radiographs (Keysers et al., 2004a) and video analysis (Dreuw et al., 2006). Image comparison methods accounting for variabilities have already been investigated for a long time in different areas, mainly for OCR applications since this was one of the first important research areas in computer vision. One of the first approaches to a deformation invariant image comparison measure is the tangent distance which was introduced by Simard et al. (1993) and later extended and further investigated by Keysers et al. (2004b). The tangent distance accounts for global image transformations and can be approximated using a linear approximation of the transformation with a Taylor expansion. Later, transformations invariant with respect to non-linear and local image deformations were introduced. Uchida and Sakoe (1998) present a dynamic programming algorithm for the two-dimensional image deformation model which takes into account continuity and monotonicity constraints. Other two-dimensional image warping models and efficient approximations are discussed by Keysers et al. (2007).

In 2006 and 2007, we applied two novel methods which have originally been developed for the recognition of objects in cluttered scenes (Deselaers et al., 2005, 2006). Both of these approaches are based on the common assumption that objects consist of parts which can be modelled more or less independently. This assumption offers the immediate advantage that, in principle, objects can still be detected if most of the object is occluded and only a single part can be identified. This approach allows us to recognise radiographs even if parts are missing, occluded, or if the view-port is chosen differently. In 2006 and 2007, an increasing share of the participating approaches were part-based and on the average these methods have achieved good results.

The idea of part-based image retrieval goes back to (Schmid and Mohr, 1997), it has also been used successfully for texture recognition (Mikolajczyk and Schmid, 2001) and object recognition (Schiele and Crowley, 1996).

Similar techniques were later made popular for object recognition and detection tasks. Fergus et al. (2003) and Leibe and Schiele (2003) made this approach popular for object recognition. Now, most approaches to recognising objects in cluttered scenes follow the assumption of objects being composed of parts. One common and successful way to create object detectors is to create a visual vocabulary, to represent the images by histograms over this vocabulary, and to classify these histograms (Dorkó and Schmid, 2005; Perronnin et al., 2006). Another common approach is based on boosting and Haar fea-

tures and was first presented by Viola and Jones (2001) for face detection. Variants of this approach use decision trees or different features (Opelt et al., 2006; Shotton et al., 2006).

Image annotation, (content-based) image retrieval, and image classification are strongly connected areas of research. Similar approaches as given above for object recognition have been proposed for each of these applications (Li et al., 2005; Duygulu et al., 2002; Jeon and Manmatha, 2004).

The remainder of this paper is structured as follows: section 3 introduces the different methods that we applied to the medical image annotation task in ImageCLEF 2005, 2006, and 2007. First, we describe simple baseline methods, then a zero-order deformation model, and finally two discriminative models that use local image descriptors. Section 2 provides a short overview of the ImageCLEF medical annotation tasks 2005, 2006, and 2007 and highlights the difficulties with respect to the methods. In section 4, experimental results of the evaluation events and of some additional experiments are presented and discussed. Finally, the paper is concluded in section 5.

2 A Short Overview of the Tasks

ImageCLEF hosts medical annotation tasks each year since 2005 with consecutive tasks building on the predecessors. The tasks increasingly involved more data, a higher number of classes, and a more complicated class structure. Here, we only give a very short description of the tasks and refer to papers which describe the tasks in detail. In 2005, the annotation task comprised 9,000 training images and 1,000 test images. Each of the test images had to be assigned to one of the 57 classes from the training data (Deselaers et al., 2007). In 2006, the training and the test data from 2005 were published as development data for system tuning and a new set of 1,000 test images was added. The complexity of the task was increased by a higher number of classes: 116 instead of 57 classes were to be distinguished (Müller et al., 2007b). In 2007, the training data and the test data from 2006 were released as training and development set respectively, but instead of classification into 116 classes, the complete IRMA code (Lehmann et al., 2003), a multi-axial mono-hierarchical code had to be predicted, and an evaluation scheme taking into account the tree-like structure of all possible labels was used to evaluate the results (Müller et al., 2007a; Deselaers et al., 2008). The code is given on four axes, where each axis is represented by 3 or 4 characters, resuling in 13 characters in total. For each axis a class-hierarchy is defined and the aim is to classify images correctly according to each axis and to each depth in the hierarchy. Errors in the classification are weighted according to their depth in the hierarchy, where errors at a coarse level are weighted higher than errors

2005: class 1
2006: class 8
2007: code 1121-120-200-700

2005: class 12
2006: class 111
2007: code 1123-127-500-000

2005: class 7
2006: class 15
2007: code 1121-120-421-700

2005: class 6
2006: class 4
2007: code 1121-110-415-700

Fig. 1. Example images from the tasks with the classes from 2005, 2006 and the complete code from 2007.

at a fine level. In addition to simple classification according to the hierarchy, it is allowed to reject classification when too unsure about the correct class (i.e. predict a wild-card character) which is scored semi-wrong and thus has the weight of half an error. In most of our experiments, however, we did not use the hierarchy at all, but considered each unique code as a class of its own.

In Figure 1, example images that were used in 2005, 2006, and 2007 along with their classes and codes.

3 Methods

In the following, we describe the methods that we applied to the ImageCLEF medical image annotation tasks. First, we describe baseline methods using global image descriptors which can be compared very efficiently. Then, we present a non-linear image deformation model accounting for local image distortions. These methods are used in a nearest neighbour manner to annotate test images. Furthermore, two methods that extract local features from the

images and store these in histograms which are then classified using a log-linear discriminative model are presented. Finally, a basic approach to make use of the hierarchical class structure in the medical image annotation task of ImageCLEF 2007 is presented. All methods are compared and similarities and differences are analysed.

3.1 Baseline Methods: Global Image Descriptors

A descriptor that captures an image as a whole in relatively few numerical values is considered a global image descriptor. Global image descriptors have long been the state of the art in many image processing and analysis applications (Squire et al., 1999; Siggelkow et al., 2001; Faloutsos et al., 1994). Recently, and only possible due to the steadily growing computational power, methods that strongly build on local image descriptors are becoming more and more dominant (Fergus et al., 2003; Dorkó and Schmid, 2005).

The global image descriptors are compared using suitable distance functions in a nearest neighbour classifier. The decision rule of the nearest neighbour classifier is

$$x \mapsto r(x) = \arg\min_k \left\{ \min_{n=1\ldots N_k} d(x, x_{nk}) \right\}, \tag{1}$$

where x is the image descriptor to be classified, x_{nk} are the image descriptors of class k from the training data, N_k is the number of training image descriptors in class k and d is a suitable distance function.

3.1.1 Image Thumbnails.

The most straight-forward way to compare images is to scale them to a common size and compare them pixel-by-pixel using e.g. the Euclidean distance. Keysers et al. (2007) show that this method serves as a reasonable baseline for many tasks including optical character recognition and recognition of medical radiographs[1]. Here, we scale the images to a common size of 32×32 pixels. To account for the frequently occurring brightness differences in the images, a 16-bin grey-value histogram normalisation is applied to each image.

3.1.2 Tamura Texture Feature Histograms.

In Tamura et al. (1978), the authors propose six texture features corresponding to human visual perception: *coarseness*, *contrast*, *directionality*, *line-likeness*,

[1] Keysers et al. (2007) give results for the data of the ImageCLEF 2005 automatic medical image annotation tasks (cp. table 1).

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 (Deselaers, 2003) and compare these histograms using the Jeffrey divergence in the nearest neighbour decision rule which was shown to be a suitable way of comparing histograms by Puzicha et al. (1999). The Jeffrey divergence is defined as

$$d_{JD}(h, h') = \sum_{m=1}^{M} \left(h_m \log \frac{2h_m}{h_m + h'_m} + h'_m \log \frac{2h'_m}{h'_m + h_m} \right), \tag{2}$$

where h and h' denote the histograms being compared and h_m is the m-th bin of h.

Similar feature histograms were also used in the QBIC system by Faloutsos et al. (1994).

3.2 Deformation Models

The image thumbnail descriptors presented in section 3.1.1 can be seen as a reasonable baseline for many image recognition tasks. One obvious problem with this descriptor is that it does not take care of visual variability such as brightness variability and global and local deformations.

One way to account for image variability is to create invariant features or to normalise the images such that the differences disappear. Another possibility is to define image comparison measures that account for the variability.

Keysers et al. (2004b) present the tangent distance which is able to cope with different global transformations such as affine transformations and brightness changes.

The most common global difference between radiographs is the brightness which depends on the size of the subject and the strength of the x-rays. The brightness differences are taken care of by histogram normalisation as described above. For the medical radiographs, apart from brightness changes, global transformations are not the main source of variations in the images since the images are taken under normalised conditions, i.e. a radiograph of an abdomen usually shows nearly the same body region independent of the patient examined.

The main source of variation are small local displacements. Keysers et al. (2007) present different image comparison methods accounting for local deformations. One of these methods is the *image distortion model* (IDM), which is computationally the least complex of the proposed methods because it does not take the deformations of neighbouring pixels into account. The other methods presented are computationally too expensive to be applied to the tasks.

The IDM is an easily implemented method accounting for small local deformations of an image. Each pixel is aligned to the pixel with the smallest squared distance from its neighbourhood. These squared distances are summed up over the complete image to obtain a global dissimilarity measure. To compare a query image Q with a database image X (both of size $I \times J$ pixels), the distance d (cf. Eq. (1)) between these two images is calculated as

$$d_{\mathrm{IDM}}(Q,X) = \sum_{i=1}^{I}\sum_{j=1}^{J} \min_{\substack{i'=-w\ldots w \\ j'=-w\ldots w}} \left\{ d'\left(Q(i,j), X(i+i', j+j')\right)\right\}. \tag{3}$$

Here w is the warp range, i.e. the radius of the neighbourhood in which a pixel may be chosen for alignment, and d' is a pixel distance comparing the image pixels $Q(i,j)$ and $X(i+i', j+j')$.

The warp range is commonly chosen to be $w = 3$, which is about 10% of the image size in our setup and thus accounts for relatively large, but still local displacements in the images.

Keysers et al. (2007) present more experiments on the issue of choosing w and also show that choosing w arbitrarily large and incorporating a distortion penalty does not improve results but rather increases the computational complexity.

This method can be strongly improved by enhancing the pixel distance d' to compare sub-images (of border length v) instead of single pixels only:

$$d'\left(Q(i,j), X(i+i', j+j')\right) = \sum_{x=-v}^{v}\sum_{y=-v}^{v} \left(Q(i+x, j+y) - X(i+i'+x, j+j'+y)\right)^2. \tag{4}$$

Figure 2 gives an example how images are compared using the IDM for images from the same class (top), and images from different classes (bottom). It can be seen that the deformed image from the same class is much more similar to the test image than the corresponding one from a different class. Furthermore it can be observed that the deformation field for the same class is much smoother than the deformation field for the different class.

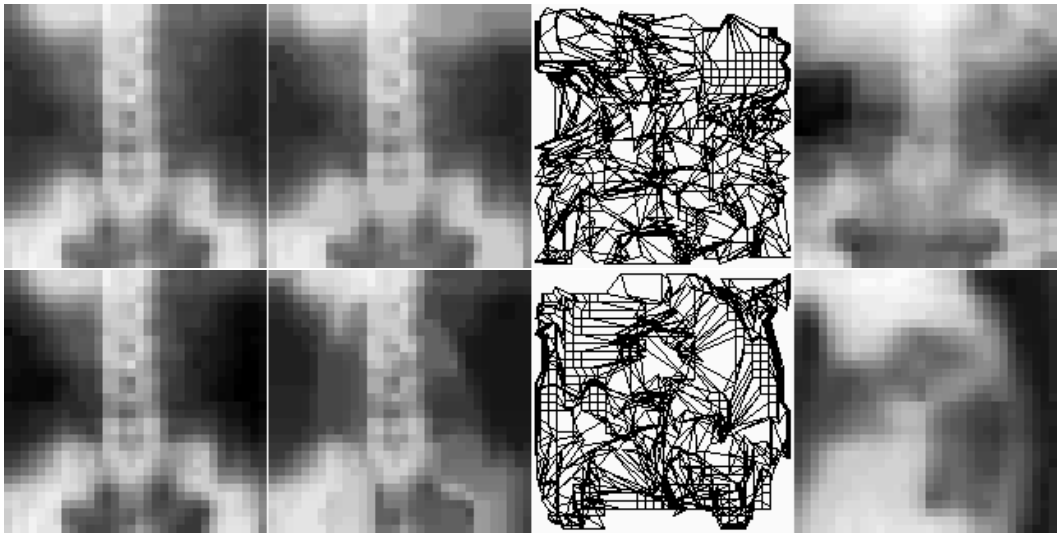

Fig. 2. Example images being compared using the image deformation model. first column: test image, second column: deformed database image, third column: deformation field, right column: database image. Top: images from the same class, bottom: images from different classes.

Further improvements are achieved by using derivatives instead of the images directly. Figure 3 gives a schematic overview how images are compared using the IDM with sub-images of local derivatives. Intuitively, the use of derivatives makes the IDM align edges to edges and homogeneous areas to homogeneous areas.

The IDM distance as described above is used in a nearest neighbour classifier (Eq. (1)) to determine the most similar training image for each test image and assign the corresponding label.

Recently, Springmann and Schuldt (2007) have presented an improvement of the IDM, lowering the computational costs significantly by reducing the number of pixels considered for highly dissimilar images.

3.3 Histogram of Patches Using a Trained Visual Vocabulary

A current trend in object recognition and detection which assumes that objects consist of parts which can be modelled independently is very common, which led to a wide variety of bag-of-features approaches (Dorkó and Schmid, 2004; Deselaers et al., 2005).

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

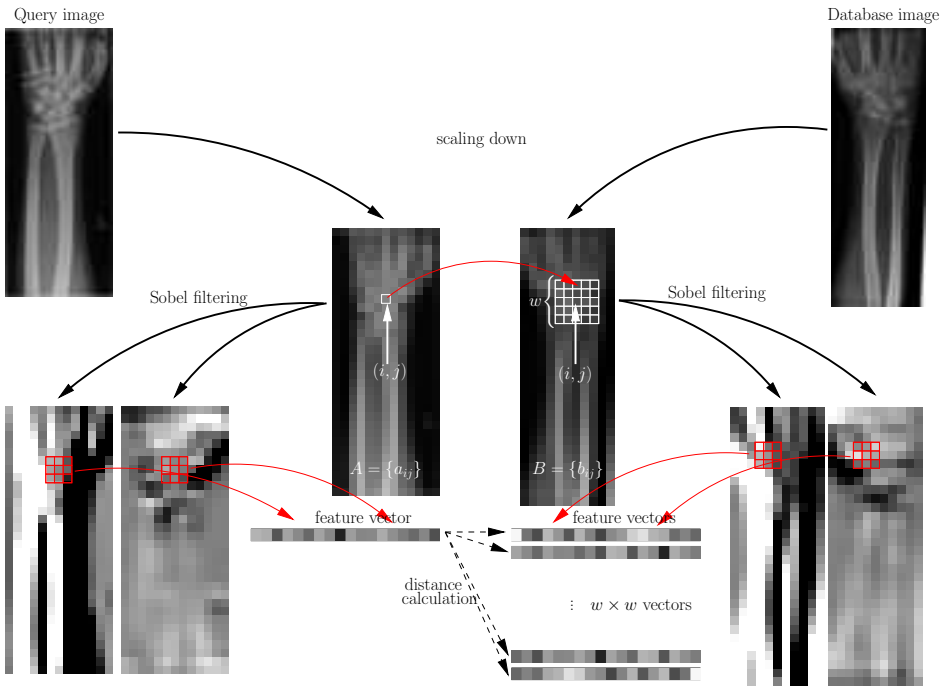

Fig. 3. Using sub-images of local gradients as features in the IDM. In a first step, the query and the database image are scaled down to a common height. Then the best match is determined for each pixel from position (i, j) in the query in a $w \times w$ neighbourhood of the corresponding pixel in the database image. To determine the best match, not only the grey value of the pixel but the values of local derivatives (Sobel filtered images) from a 3×3 neighbour hood are considered.

(1) In the first step, sub-images are extracted from all training images and the dimensionality is reduced to 40 dimensions using PCA transformation. The sub-images are extracted at interest points that were detected using the wavelet-based approach proposed by Loupias et al. (2000).
(2) In the second step, the sub-images of all training images are jointly clustered using the EM algorithm for Gaussian mixtures. The Gaussian mixture densities are iteratively split starting with one single density which is split to form two densities. Then, the densities are reestimated until convergence and then re-split until 2^n densities (n is the number of splits) are created. It has been observed that usually 2048 clusters or 11 splits are sufficient to obtain good classification accuracy.
(3) In the third step, all information about each sub-image is discarded except its closest cluster centre. Then, for each image, a histogram $h(X)$ over the cluster identifiers b of the respective patches x_l is created, thus effectively coding which "visual words" from the code-book occur in the image:

$$h_b(X) = \frac{1}{L_X} \sum_{l=1}^{L_x} \delta(b, b(x_l)),$$

where $h_b(X)$ denotes the b-th bin of the histogram representing image

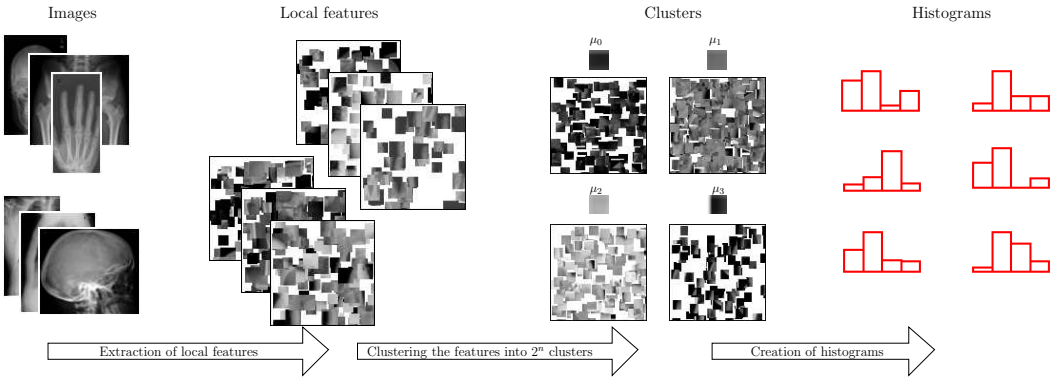

Fig. 4. Overview of the creation of a visual vocabulary and the creation of patch histograms. First, patches are extracted in all training images at interest points, then the patches from all training images are jointly clustered to form a Gaussian mixture density (we depict the means μ and the patches in the cluster) and finally the images are represented by histograms counting how many of the patches correspond to which cluster centres.

X, $b(x_l)$ is the identifier of the closest cluster centre of patch x_l and δ denotes the Kronecker delta.

These histograms are then classified using a log-linear model which was trained according to the maximum entropy approach (Deselaers et al., 2005).

The decision rule for this classifier is

$$h(X) \mapsto r(h(X)) = \arg\max_c \{p(c|h(X))\} \text{ with}$$
$$p(c|h(X)) = \frac{\exp\left(\alpha_c + \sum_{b=1}^{B} \lambda_{bc} h_b(X)\right)}{\sum_{c'=1}^{C} \exp\left(\alpha_{c'} + \sum_{b=1}^{B} \lambda_{bc'} h_b(X)\right)}.$$

The maximising model is unique and the problem is convex, thus it is possible to obtain the global optimum.

3.4 Histogram of Patches Using a General Visual Vocabulary

This approach is also 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 (Deselaers et al., 2005), 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.

3.4.1 Creation of histograms

The distribution of the patches extracted from an images is approximated using a histogram. In contrast to the previous section, here we extract patches in various scales from each position in the image. To reduce the necessary storage, the histograms are created without explicitly storing any feature vector. Thus, the creation of the histograms is a three step procedure: in the first step, the PCA transformation is determined as described above. In the second step, the mean and the variance of the transformed patches are calculated to determine a reasonable grid for the histograms. In the last step, the histograms themselves are created. For each of these steps, all training images are considered.

(1) In the first step, all possible patches in various sizes from all training images are extracted and their mean and the covariance matrix are estimated to determine the PCA transformation matrix.
(2) Given this PCA transformation matrix and the means, the mean μ_d and the variance σ_d^2 for each component d of the transformed vectors is calculated to determine the bin boundaries for the histograms. The bins for component d are uniformly distributed between $\mu_d - \alpha\sigma_d$ and $\mu_d + \alpha\sigma_d$.
(3) Then, we consider all dimensionality reduced patches from the training images and create one histogram per training image. This step is depicted in Figure 5. The processing is from left to right: first the patches are extracted, then PCA transformed, then the position of the patch is concatenated to the PCA transformed feature vector, and finally the vectors are inserted into the sparse histogram data structure.

As mentioned above, the patches are not explicitly stored in any of these steps as this would lead to immense memory requirements.

Informal experiments have shown that 6 to 8 dimensions for the PCA reduced vectors lead to the best results, and that $\alpha = 1.5$ is a good value to determine bin boundaries. Values exceeding the given range are clipped.

3.4.2 Spatial Information.

One serious issue with many part-based models is the incorporation of spatial information. To incorporate spatial information in our approach, we simply concatenate the extraction position to the PCA reduced feature vectors and thus simply add two further components to the histograms. These additional components can easily be handled by the histograms. As the range of values for each component is calculated individually and independently of the other components, no special processing of these additional components is required. One issue with the inclusion of the absolute patch extraction positions is that translation invariance, normally one of the major advantages of part-based

models, is partly lost. Still, currently it is unclear how to incorporate relative position information into the model presented here. It will be shown later that for the tasks considered here, either the translation invariance is not required, or translations are sufficiently represented in the training data.

Using this method, we create sparse histograms of 8^4 ($65536 = 2^{16}$) bins, i.e. joint histograms to store 8-dimensional data, where each dimension is split into 4 bins. These histograms can either be classified using the nearest neighbour rule with suitable histogram comparison measure, or a discriminative classifier can be applied. Here, we use a support vector machine with a histogram intersection kernel and the same discriminatively trained log-linear maximum entropy model which is used for the histograms described in Section 3.3.

For nearest neighbour classification, one problem is the sparseness of the data, and thus histogram comparison measures like the Jeffrey divergence (Eq. (2)) are not suitable for this data. On the other hand, the common histogram comparison measures which take into account neighbouring bins such as the Earth Movers Distance (EMD) (Rubner et al., 1998) are computationally too expensive to be applied to histograms with several thousand bins. Therefore, we propose the *Histogram Distortion Model* (HDM) which is inspired by the IDM.

Histogram Distortion Model. The HDM is inspired by the image distortion model (cp. section 3.2). The implementation of the HDM is straightforward for any bin-by-bin histogram comparison measure, as long as neighbourhoods are defined for the underlying histograms. Given a bin at position $p = (p_1, \dots p_D)$, we use the bin from position γ out of the neighbourhood $U(p)$ of p that minimises the resulting distance. The neighbourhood $U(p)$ is chosen such that one neighbouring bin in each direction of each dimension is considered, which, given the very coarse quantisation, accounts for relatively large changes. Here, we use it as an extension to the Jeffrey Divergence, i.e. we replace the distance function $d_{JD}(h(X), h(X'))$ in Eq. 1 with

$$d_{\text{JDDM}}(h(X), h(X')) = \sum_{p=1}^{P} \min_{\gamma \in U(p)} \left\{ h_p(X) \log \frac{2h_p(X)}{h_p(X) + h_\gamma(X')} + h_\gamma(X') \log \frac{2h_\gamma(X')}{h_\gamma(X') + h_p(X)} \right\}. \tag{5}$$

A related way to account for neighbouring bins in the comparison of histograms would be to smooth the histograms, which would require more memory as it would lead to non-sparse histograms.

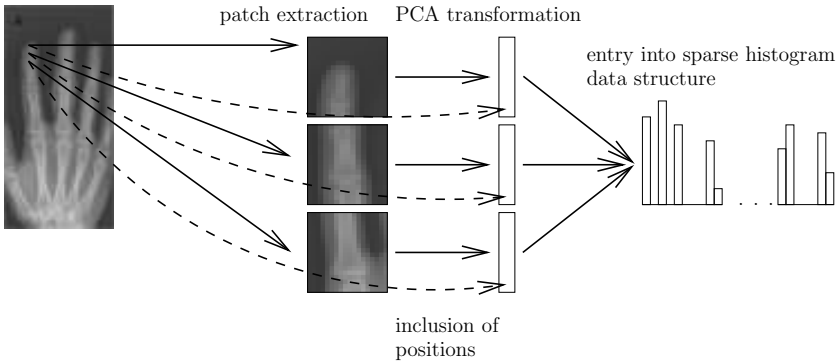

Fig. 5. Sparse histogram creation: first sub-images are extracted at every position, then the dimensionality of the patches is reduced with PCA. The position is appended to the PCA coefficients and then the histogram bin corresponding to the vector is increased.

Since neighbourhoods in the histograms have to be defined for the HDM to be applicable, the HDM can only be applied to the histograms using a general vocabulary. In these histograms, the neighbourhood is easily defined, because the whole feature space is covered with a rectangular grid. For the histograms over learnt vocabularies, the individual clusters are not created in any particular order and therefore a neighbourhood is not automatically created and thus the HDM cannot be directly applied. However, it would be possible to define a neighbourhood by considering the cluster-means for the individual clusters, but we do not expect a performance gain here, as sparsity is not an issue for these histograms.

3.5 Comparison of the Methods

The methods presented here are strongly connected in the following way. The most obvious one is the connection between the IDM and the simple thumbnails. Both methods use the same image representation but the IDM accounts for image deformations which clearly occur in radiographs due to different patients and slightly varying viewing angles. The allowed deformations can easily be restricted by choosing a small warp-range so that mainly image deformations which do not alter the classification of an image are modelled.

The two different patch histogram-based approaches are also strongly connected as they are based on the same assumption about the compositionality of objects to be recognised. They mainly differ in the way the vocabulary is created.

A not so obvious connection exists between the patch histogram-based methods and the IDM. Since the sparse histograms also include position (albeit strongly quantised) and the IDM includes local subwindows to determine the

best deformation the approaches are related to each other. In informal experiments, the IDM using PCA-coefficients of the neighbourhoods instead of gradients was evaluated and, as expected, the performance remains mainly unchanged. Furthermore, if the histograms had a much finer quantisation of the positions (and possibly also of the features), a reconstruction of the image would be possible, since then, each feature would have its own bin and the HDM in fact would become a variant of the IDM.

3.6 Hierarchy-Aware Classification

The above-described methods all are flat-classification schemes, i.e. they do not account for the hierarchical structure of the classes at all. In the following we describe how these approaches can be extended to account for the hierarchy.

3.6.1 Hierarchy-Aware Run-Combinations

One possibility to use the hierarchy is to create different flat-classification runs and combine these using the hierarchy, e.g. using three runs and putting a wild-card character at a position (and all succeeding positions) if not at least two of the three runs agree about a particular position. This approach can be tuned with respect to the number of runs being combined and the agreement necessary for a position not to be predicted with the wild-card.

3.6.2 Predicting Axes Independently

Another possibility to use the hierarchy for this task is to predict the four independent axes individually. This has the advantage that the individual training processes can be done more efficiently as the number of classes is lower and hopefully more reliable as the amount of training data is constant. The downside of this approach is that possible correlations between the axes are being lost and that combinations of invalid codes might be predicted. In principle, this approach can be extended toward using a classifier tree for each axis.

4 Experimental Results

Results from the ImageCLEF medical annotation tasks 2005, 2006, and 2007 are given in Table 1. The table lists the results for all methods presented in this paper and furthermore gives the best result for each year. In 2005 and

Table 1
Results for the different methods for the ImageCLEF tasks. Results from the official evaluation are printed in bold-face.

	ImageCLEF			
	error rate [%]			score
method	2005	2006	2007	2007
32x32 thumbnails	36.8	32.1	32.4	112.2
Tamura texture features	33.1	51.8	50.5	174.4
image distortion model (IDM)	**12.6**	**20.4**	21.6	61.7
patch histogram with learnt voc.	**13.9**	22.4	27.6	98.6
patch histogram with general voc.	9.3	**16.2**	**11.9**	**33.0**
patch histogram with general voc. (SVM)	10.0	**16.7**	–	–
best run in evaluation	**12.6**	**16.2**	**10.3**	**26.8**

2006, our methods performed best. Official results from the evaluation are shown in boldface. The best results in 2007 were from Tommasi et al. (2007) using a multi-cue kernel to fuse local and global image descriptors.

It is observed that the IDM, which was the best method in 2005, cannot compete with the discriminative method which performed best in 2006 and 2007. For comparison, we performed experiments with the histogram methods on the 2005 data. It can also be observed that the patch histograms with learnt visual vocabulary are slightly worse than the IDM. This is probably due to missing spatial information in this approach. The comparison experiments with the patch histograms with general vocabulary on the 2005 data show that they outperform the IDM in each year.

For the baseline methods it is observed that the Tamura texture features outperform the thumbnails on the 2005 data which consisted of 57 classes but is clearly beaten by the thumbnail images on the 2006 and 2007 data. A possible explanation for this effect is probably that the Tamura texture features are more invariant to certain changes in the images. Therefore, they capture the higher variability in the data of 2005 but are disadvantageous for the 2006 and 2007 tasks which have more classes and thus less intra-class variability.

In Table 2 we give additional results for the ImageCLEF 2005 task for the patch histograms with general visual vocabulary using different classifiers, with and without position information. It can be observed that position information always leads to improved results and that the maximum entropy method outperforms all other methods. Furthermore, it can be observed that the histogram distortion model based on the Jeffrey divergence outperforms the simple Jeffrey divergence in a nearest neighbour classifier. The support

Table 2
Impact of different settings on the performance of the sparse patch histograms on the ImageCLEF 2005 task.

method	ER [%]
sparse histograms (without position)	
+ nearest neighbour	13.0
+ HDM, nearest neighbour	12.5
+ maximum entropy classification	11.6
+ support vector machine	11.3
sparse histograms (with position)	
+ nearest neighbour	10.1
+ HDM, nearest neighbour	9.8
+ maximum entropy classification	9.3
+ support vector machine	10.0

vector machine performs slightly better than the maximum entropy method for the histograms without position information but slightly worse when position information is used. This might be due to slight over-fitting. The success of using absolute position information in the patch histograms can be explained by the very coarse quantisation of the data, the stored positions subdivide the image into only 16 regions. For the IDM method the relatively small displacements captured in the distance function is sufficient because the variability captured in the training data is sufficient to account for large displacements.

In 2007, we created two runs trying to exploit the hierarchy. The first run was a combination of four slightly different runs of the sparse histogram method differing only in the number of histogram bins and in the scaling of the original images. These four runs were combined such that the wild-card character was set for a position (and all succeeding positions) if not at least three of the runs agree about the position. This run was slightly better than the best of the four runs and thus was our best submission in 2007. The results for the four individual runs and the combined run are given in Table 3.

The second run trying to exploit the hierarchy used individual classifiers for the four axes. Unfortunately, this run could not achieve a competitive result, having a score of 44.6 and an error rate of 17.8%. The reason for the failing of this method is probably that the assumption that the four axes are independent is not valid. On the one hand, large parts of the code are not used at all, and on the other hand, some combinations of the code are not valid and, thus, a method that works on a per-axis basis can create codes that cannot be assigned to any image. For the data at hand, the axes one to four have 4, 26, 63, and 5 unique codes, respectively, which, in principle, can be combined to 32,760 different codes, but only 116 codes occur in the data.

Table 3
Hierarchy-aware combination of runs in ImageCLEF 2007. First four lines: individual runs, bottom line: combined run. All results from (Müller et al., 2007a).

run id	score	ER [%]
RWTHi6-SH65536-SC025-ME	33.0	11.9
RWTHi6-SH65536-SC05-ME	33.2	12.3
RWTHi6-SH4096-SC025-ME	34.6	12.7
RWTHi6-SH4096-SC05-ME	34.7	12.4
RWTHi6-4RUN-MV3	30.9	13.2

5 Conclusion

We presented three different approaches to automatic annotation of medical radiographs. The methods were applied to the automatic medical image annotation tasks of ImageCLEF 2005, 2006, and 2007 and achieved competitive results. It is observed that discriminatively trained methods clearly outperform other methods and that local image descriptors work better than global image descriptors.

The knowledge of the class hierarchy could be used to combine runs to obtain an improvement over single runs. The results of our approach which tries to make direct use of the hierarchy/independence of the axis, are worse than the ones which use a flat classification scheme because the independence assumption for the four axes seems to be invalid.

For the future we intend to combine image deformation models with discriminative training approaches to take advantage of both approaches.

Acknowledgements. The authors would like to thank Daniel Keysers, Christian Gollan, Andre Hegerath, Tobias Gass, Tobias Weyand for their contributions to this work. This work was partially funded by the DFG (Deutsche Forschungsgemeinschaft) under contract NE-572/6.

References

Deselaers, T., 2003. Features for image retrieval. Master's thesis, Human Language Technology and Pattern Recognition Group, RWTH Aachen University, Aachen, Germany.

Deselaers, T., Hegerath, A., Keysers, D., Ney, H., Sep. 2006. Sparse patch-histograms for object classification in cluttered images. In: DAGM 2006, Pattern Recognition, 27th DAGM Symposium. Vol. 4174 of Lecture Notes in Computer Science. Berlin, Germany, pp. 202–211.

Deselaers, T., Keysers, D., Ney, H., Jun. 2005. Discriminative training for object recognition using image patches. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR 05). Vol. 2. San Diego, CA, pp. 157–162.

Deselaers, T., Müller, H., Clough, P., Ney, H., Lehmann, T. M., Aug. 2007. The

CLEF 2005 automatic medical image annotation task. International Journal of Computer Vision 74 (1), 51–58.

Deselaers, T., Müller, H., Deserno, T. M., 2008. Automatic medical image annotation in ImageCLEF 2007: Overview, results, and discussion. Pattern Recognition Letters, in preparation.

Dorkó, G., Schmid, C., 2004. Object class recognition using discriminative local features. submitted to IEEE Transactions on Pattern Analysis and Machine Intelligence.

Dorkó, G., Schmid, C., Feb. 2005. Object class recognition using discriminative local features. Rapport de recherche RR-5497, INRIA - Rhone-Alpes.
URL http://lear.inrialpes.fr/pubs/2005/DS05a

Dreuw, P., Deselaers, T., Keysers, D., Ney, H., May 2006. Modeling image variability in appearance-based gesture recognition. In: European Conference on Computer Vision (ECCV 2006): 3rd Workshop on Statistical Methods in Multi-Image and Video Processing. Graz, Austria, pp. 7–18.

Duygulu, P., Barnard, K., de Freitas, N., Forsyth, D., May 2002. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In: European Conference on Computer Vision. Vol. 9. Copenhagen, Denmark, pp. 97–112.

Faloutsos, C., Barber, R., Flickner, M., Hafner, J., Niblack, W., Petkovic, D., Equitz, W., Jul. 1994. Efficient and effective querying by image content. Journal of Intelligent Information Systems 3 (3/4), 231–262.

Fergus, R., Perona, P., Zissermann, A., Jun. 2003. Object class recognition by unsupervised scale-invariant learning. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR 03). Blacksburg, VG, pp. 264–271.

Jeon, J., Manmatha, R., 2004. Using maximum entropy for automatic image annotation. In: Proceedings of the 3rd International Conference on Image and Video Retrieval. pp. 24–32.

Keysers, D., Deselaers, T., Gollan, C., Ney, H., Aug. 2007. Deformation models for image recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (8), 1422–1435.

Keysers, D., Gollan, C., Ney, H., Mar. 2004a. Classification of medical images using non-linear distortion models. In: Proc. BVM 2004, Bildverarbeitung für die Medizin. Berlin, Germany, pp. 366–370.

Keysers, D., Macherey, W., Ney, H., Dahmen, J., Feb. 2004b. Adaptation in statistical pattern recognition using tangent vectors. IEEE Transactions on Pattern Analysis and Machine Intelligence 26 (2), 269–274.

Lehmann, T. M., Schubert, H., Keysers, D., Kohnen, M., Wein, B. B., 2003. The irma code for unique classification of medical images. In: Proc. SPIE 2003. pp. 440–451.

Leibe, B., Schiele, B., Sep. 2003. Interleaved object categorization and segmentation. In: British Machine Vision Conference (BMVC'03). Norwich, UK.

Li, Y., Shapiro, L. G., Bilmes, J. A., Oct. 2005. A generative/discriminative learning algorithm for image classification. In: International Conference on Computer Vision (ICCV 2005). Beijing, China, pp. 1605–1612.

Loupias, E., Sebe, N., Bres, S., Jolion, J., Sep. 2000. Wavelet-based salient points for image retrieval. In: International Conference on Image Processing. Vol. 2.

Vancouver, Canada, pp. 518–521.

Mikolajczyk, K., Schmid, C., Jul. 2001. Indexing based on scale invariant interest points. In: International Conference on Computer Vision (ICCV 2001). Vancouver, Ca, pp. 525–531.

Müller, H., Deselaers, T., Kim, E., Kalpathy-Kramer, J., Deserno, T. M., Hersh, W., Sep. 2007a. Overview of the ImageCLEFmed 2007 medical retrieval and annotation tasks. In: Working notes of the CLEF 2007 Workshop. Budapest, Hungary.

Müller, H., Deselaers, T., Lehmann, T., Clough, P., Hersh, W., 2007b. 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. Vol. 4730 of LNCS. Alicante, Spain, pp. 595–608.

Opelt, A., Pinz, A., Fussenegger, M., Auer, P., Mar. 2006. Generic object recognition with boosting. IEEE Transaction on Pattern Analysis and Machine Intelligence (PAMI) 28 (3), 416–431.

Perronnin, F., Dance, C., Csurka, G., Bressan, M., May 2006. Adapted vocabularies for generic visual categorization. In: European Conference on Computer Vision (ECCV 2006). Graz, Austria.

Puzicha, J., Rubner, Y., Tomasi, C., Buhmann, J., Sep. 1999. Empirical evaluation of dissimilarity measures for color and texture. In: International Conference on Computer Vision (ICCV 1999). Vol. 2. Corfu, Greece, pp. 1165–1173.

Rubner, Y., Tomasi, C., Guibas, L. J., Jan. 1998. A metric for distributions with applications to image databases. In: International Conference on Computer Vision. Bombay, India, pp. 59–66.

Schiele, B., Crowley, J. L., Aug. 1996. Probabilistic object recognition using multidimensional receptive field histograms. In: International Conference on Pattern Recognition. Vienna, Austria.

Schmid, C., Mohr, R., 1997. Local grayvalue invariants for image retrieval. IEEE Transactions on Pattern Analysis & Machine Intelligence 19 (5), 530–534.

Shotton, J., Winn, J., Rother, C., Criminisi, A., May 2006. TextonBoost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In: ECCV 2006. Vol. 3951 of LNCS. Graz, Austria, pp. 1–15.

Siggelkow, S., Schael, M., Burkhardt, H., Sep. 2001. SIMBA — Search IMages By Appearance. In: DAGM 2001, Pattern Recognition, 23rd DAGM Symposium. Vol. 2191 of LNCS. Springer Verlag, Munich, Germany, pp. 9–17.

Simard, P., Le Cun, Y., Denker, J., 1993. Efficient pattern recognition using a new transformation distance. In: Advances in Neural Information Processing Systems. Vol. 5. pp. 50–58.

Springmann, M., Schuldt, H., Sep. 2007. Speeding up IDM without degradation of retrieval quality. In: Working Notes of the 2007 CLEF Workshop. Budapest, Hungary.

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

Tamura, H., Mori, S., Yamawaki, T., Jun. 1978. Textural features corresponding to

visual perception. IEEE Transaction on Systems, Man, and Cybernetics 8 (6), 460–472.

Tommasi, T., Orabona, F., Caputo, B., Sep. 2007. CLEF2007 Image Annotation Task: an SVM–based Cue Integration Approach. In: Working Notes of the 2007 CLEF Workshop. Budapest, Hungary.

Uchida, S., Sakoe, H., Aug. 1998. A monotonic and continuous two-dimensional warping based on dynamic programming. In: International Conference on Pattern Recognition (ICPR 1998). Vol. 1. pp. 521–524.

Viola, P., Jones, M., Jul. 2001. Robust real-time object detection. In: Second International Workshop on Statistical and Computational Theories of Vision – Modeling, Learning, Computing, and Sampling. Vancouver, Canada, pp. 1–25.