Image Classification using Cluster-Cooccurrence Matrices of Local Relational Features

Lokesh Setia, Alexandra Teynor, Alaa Halawani and Hans Burkhardt
Albert-Ludwigs-University Freiburg
79110 Freiburg im Breisgau, Germany
{setia, teynor, halawani, burkhardt} @informatik.uni-freiburg.de

ABSTRACT

Image classification systems have received a recent boost from methods using local features generated over interest points, delivering higher robustness against partial occlusion and cluttered backgrounds. We propose in this paper to use relational features calculated over multiple directions and scales around these interest points. Furthermore, a very important design issue is the choice of similarity measure to compare the bags of local feature vectors generated by each image, for which we propose a novel approach by computing image similarity using cluster co-occurrence matrices of local features. Excellent results are achieved for a widely used medical image classification task, and ideas to generalize to other tasks are discussed.

Categories and Subject Descriptors: H.3 [Information Storage and Retrieval]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; I.5 [Pattern Recognition]: I.5.3 Clustering; I.5.4 Applications;

General Terms: Algorithms, Design

Keywords: Classification, Image Analysis, Image Annotation, Local Features

1. INTRODUCTION

Digital recordings of all kinds of visual data are stored in huge databases, consider e.g. image libraries at publishing companies, digital art galleries or medical image archives at hospitals. In order to retrieve specific images in these databases again, powerful automatic tools working on the image content are desirable. Content based image retrieval (CBIR) methods are already capable of finding visually similar images (e.g. from a color and/or texture point of view) quite reliably. However, users want to retrieve images based on semantic entities present in them. A traditional approach is to use keywords for the description of the image content, as was already used for printed image collections. Manual annotation of the sheer mass of images is usually not feasible any more, even worse error prone, since different people tend to judge the content of an image differently.

Also in areas where the content of the image is undisputed, as e.g. in medical X-ray images, we need automatic procedures to assign keywords, since sometimes the images were never labeled or simply mislabeled. As an example, a study carried out by the University Hospital in Aachen [12] revealed over 15% errors in just one specific tag alone, which was manually assigned to X-Ray images taken during normal clinical routine. Unless corrected, these images cannot be found again with keyword search alone.

The assignment of keywords to an image can be seen as a classification problem, especially if the pool of keywords is fixed and only certain combinations of keywords are allowed [26]. In some applications, the possibility to be able to assign multiple classes to a single image has been considered [19]. These applications call for effective image analysis algorithms. Recent image analysis methods often employ local information extracted around so called *interest points*. The benefits are easy to see: they can deal with partial occlusion and cluttered backgrounds better than globally computed features. In this work, we present a method for classification of images based on gradient-like relational features computed around interest points. Apart from the features themselves, also their spatial relation to each other is considered in order to improve recognition performance.

Related Work

The research area of general image classification has always been very active. Recently the focus is drawn to methods using local information extracted in various ways from patches around the above mentioned interest points. In this paper, we extend the relational features introduced originally for texture analysis to interest-point based general image classification. Their main advantages are increased robustness to illumination changes, and the ease at which local information at various scales can be captured. Current object classification systems also differ in the way the local information of the patches is combined: some use only the feature vectors extracted at the points, others also take their (normalised) position information into account. Examples for effective object classification systems neglecting the spatial layout of the interest points are by Csurka et al [6] or Deselaers et al [9]. They use histograms of patch cluster memberships as features and for classification, SVM and discriminative training respectively is used. An early example for the combination of local appearance and positional information was by Weber et al [27], further developed by Fergus et al [11] and Fei-Fei et al [10]. They introduced a so called "constella-

tion model", i.e. specific local image features in a probabilistic spatial arrangement, to decide whether a certain object is present in a scene or not. The positional information is also used by Agarwal et al [1]. They classify sub-windows in images using binary vectors coding the occurrences and spatial relations of local features. In order to incorporate the spatial position of interest points, we propose the use of co-occurrence relationships derived from cluster memberships of relational features calculated over different distances and orientations. Gray-value co-occurrence matrices were originally introduced for texture classification by Davis et al [7], these were extended to color correlograms by Huang et al [15]. Belongie et al [4] used constellations of such correlograms for contour matching, which were extended to use local features by Amores et al [2] with encouraging results. They used constellations of $d+2$ dimensional correlograms calculated for each member of a sparse set of points, by linearly quantizing the d dimensions of the local feature vectors along with log-polar spatial quantization. However, joint local information of multiple feature vectors was not used. In our case, we propose a multi-dimensional co-occurrence matrix capturing the statistical properties of the joint distribution of cluster membership indices derived from local features.

Outline

The paper is structured in a top-down approach to make it more accessible. In Section 2 we first present in short all the steps belonging to the proposed method, followed by details on the individual stages. Experiments performed on a radiograph database are described in Section 3, along with a comparison with previously published results. We then provide a thorough discussion on various interesting aspects of the algorithm in Section 4, followed by a conclusion in Section 5.

2. DESCRIPTION OF THE ALGORITHM

The major steps involved can be summarized as follows. For all training images, do the following:

1. **Preprocessing** Convert image to grayscale if needed, normalize grayvalues between 0 and 1.

2. **Interest Points** Apply an interest point detector (in our case, the Lupias Salient-Point Detector). Sort the obtained saliency map, and take the N_s points with the highest saliency values for further computation.

3. **Local Relational Feature Generation** Evaluate a number of relational functions $\Re(x, y, r_1, r_2, \phi, n)$. Each function gives for each interest point, a sub feature vector of length n. These are concatenated to get a local feature vector for each interest point.

4. **Clustering** Take a random subset of local feature vectors from all training images. Cluster these feature vectors in N_c clusters according to some optimization criteria. Save the cluster centers for later use with test images.

5. **Cluster Co-occurrence Matrix** The nearest cluster is calculated for all local feature vectors of the image. The complete local feature vectors are discarded, and the only retained information is the index of the nearest cluster. Consider all possible salient-point pairs. A cluster co-occurrence matrix of size $N_c \times N_c$ is generated sector-wise (i.e. over radial and angle ranges), yielding a 4-D feature matrix. This 4-D matrix is flattened and used as the final feature vector for the image for use with an independent classifier.

6. **For the Test Images** The steps 1, 2, 3 and 5 are repeated for the test images. The cluster centers used in step 4 are used as required.

7. **Classification** Multi-Class Support Vector Machines (SVMs) are trained using the features obtained in step 5 for the training examples. Various parameters are tuned using cross-validation. The best parameters are selected and used to train a multi-class SVM using all training examples. This is used to classify the given test images.

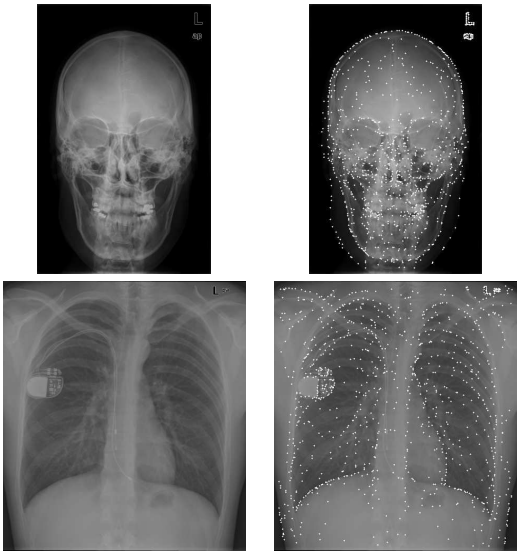

Figure 1: 1000 Interest points with the highest saliancy value for each of the two images shown on the left. Although some interest points are found in the non-discriminative parts of the images (for example, the man-made object embedded in the chest or the text in the top-right corner), local methods are still very robust to partial matching.

2.1 Interest Point Detection

The method used here is the salient point extraction algorithm introduced by Loupias and Sebe [20]. We have decided to use this salient point detector as it was found that it has more information content and better repeatability compared with the well-known Harris detector [24]. Our informal tests affirmed these statements atleast for the images taken from the database used in this paper.

The assumption is that image points, where high variations occur, represent important information in the image (areas of high relevance) and are thus extracted. One can study the variations that are present in an image using the wavelet analysis which allows for multiresolution representation of a signal (image). The algorithm starts from the coarsest resolution after representing the image in the wavelet

domain, always going back one step to a finer resolution, choosing from the set of available points the one with the highest wavelet coeffient at that level. This is applied until one ends up with picking one coefficient at level 1 (level 0 represents the original image). This coefficient represents a number of points in the original image. Among these points, the point with the maximum gradient is chosen and is given a value representing its saliency. This saliency value is equal to the sum of the absolute value of the wavelet coefficients along the whole track:

$$s = \sum_{i=1}^{l} |c_i| \tag{1}$$

The above scenario is repeated for every wavelet coefficient that exceeds a certain threshold, τ, in order to avoid computation time by not investigating small wavelet coefficients. We end up with a matrix (which we call the *"Saliency map"* here) representing the saliencies of the image pixels. The saliency map is then sorted and a fixed number of salient points (N_s) per image is taken in this work. An alternative strategy is to fix a threshold, and select all points having a saliency above this threshold. The pixels very near to the image boundary (upto 6 pixels in our case) are not considered candidates for an interest point, as the local features cannot be accurately calculated there without introducing artifacts. The detected interest points for two sample images from the used database are shown in Figure 1

2.2 Relational Features

Relational features are motivated from the use of relational kernels in texture classification, introduced by Schael in [23] based on the Local Binary Pattern (LBP) texture features [22] which map the relation between a center pixel and the pixels in its neighborhood into a binary pattern.

Local Binary Pattern features are invariant against monotonic grayscale transformations. They eliminate the effect of illumination by comparing the value of a center pixel with the values of the pixels in its neighborhood. Then the sign of the difference is considered instead of the value itself. If the value of a neighboring pixel is greater than or equal the value of the center pixel, then the difference is mapped to the value 1, else it is set to 0. Applying this to all pixels in a circular neighborhood of the center pixel, we end up with a binary pattern which can be transformed into a unique number as follows [22]:

$$LBP = \sum_{i=0}^{n-1} s\left(v_i - v_c\right) 2^i, \quad \text{where} \tag{2}$$

$$s\left(x\right) = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0, \end{cases} \tag{3}$$

where v_i and v_c are the grayvalues at a neighboring pixel and at the center pixel, respectively, and n gives the number of the pixels in the circular neighborhood of the center pixel. Since the signed difference $(v_i - v_c)$ is considered, the effect of grayscale shifts is totally eliminated. Invariance against scaling of the grayscale is achieved by the s operator as the sign of the difference is mapped to 0 or 1.

It is obvious that the main disadvantage of these features is the discontinuity of the LBP operator (the s function), which makes them sensitive to noise; a small disturbance in the image may cause a big deviation of the feature. To overcome this problem, Schael [23] has introduced an operator which extends the step function in Equation 3 to a ramp function giving values in the range of $[0, 1]$:

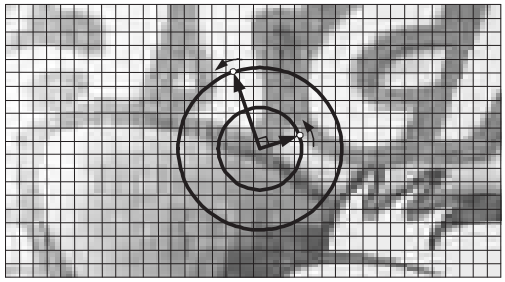

Figure 2: Calculation of a set of relational features. A feature is formed by applying the relational function to the gray-value difference of the pixels lying on specific distance and phase to the salient point in question (i.e. center of the circles)

$$\mathrm{rel}\left(x\right) = \begin{cases} 1 & \text{if } x < -\epsilon \\ \frac{\epsilon - x}{2\epsilon} & \text{if } -\epsilon \leq x \leq \epsilon \\ 0 & \text{if } \epsilon < x \end{cases}, \tag{4}$$

where ϵ is a threshold parameter. This way, the features are much more robust against noise, but we also sacrifice the 100% invariancy to monotonic grayscale transformations (although the features are still robust to these transformations). If ϵ is set to zero, the rel function will reduce to the simple LBP operator s.

Based on the relational operator defined in Equation 4, we define a relational function $\Re(x, y, r_1, r_2, \phi, n) \mapsto \mathbb{R}^n$, calculated on a salient point x, y of the image $\mathbf{I}$ as center. To simplify notation, let the individual output values of the function be given by

$$R_k = [\Re(x, y, r_1, r_2, \phi, n)]_k, \quad k = 1, \ldots, n$$

Then,

$$R_k = \mathrm{rel}(\mathbf{I}(x_2, y_2) - \mathbf{I}(x_1, y_1)),$$

where

$$(x_1, y_1) = (x + r_1 \cos(k \cdot 2\pi/n), y + r_1 \sin(k \cdot 2\pi/n)),$$

and

$$(x_2, y_2) = (x + r_2 \cos(k \cdot 2\pi/n + \phi), y + r_2 \sin(k \cdot 2\pi/n + \phi))$$

The process is illustrated in Figure 2. Bilinear interpolation is used for points not lying exactly on the image grid. Based on different combinations of r_1, r_2 and ϕ, local information at different scales and orientations can be captured. In this work, we use 3 sets of parameters, $(0, 5, 0)$, $(3, 6, \pi/2)$ and $(2, 3, \pi)$, each with $n = 12$. The 3 subvectors are concatenated to yield a local feature vector of length 36 at each salient point. It is of interest to note, that in applications where rotation invariance is desired, a subvector can simply be summed up to yield a rotation invariant descriptor.

The ensemble of local feature vectors extracted from all training images are clustered as explained in the next section. To remain computationally feasible, the process is carried out on 18000 randomly chosen local feature vectors.

2.3 Clustering

Clustering can be understood as a grouping of similar objects. For vectorial data, this process has also been extensively studied in the branch of vector quantization. The main benefit of clustering is that of compaction (and thus implicitly abstraction) of data. In this respect, clustering can be more effective than a uniformly spaced histogram whose bins have been derived more or less independently of the training data.

There are many possible approaches to clustering, in this work we use one of the most common algorithm, the *k-means clustering algorithm*. K-means is an iterative algorithm which minimizes the sum, over all clusters, of the within-cluster sums of point-to-cluster-centroid distances. The k in k-means stands for the number of desired clusters and is an input to the algorithm. Other decisions to be made include an appropriate distance measure, which we simply take to be Euclidean, and the choice of initial clusters, which we take to be randomly chosen local feature vectors.

The number of clusters is denoted in this paper by N_c and must be selected carefully as the size of the final feature vector increases quadratically with N_c.

2.4 Cluster Co-occurrence Matrix

An open research problem with local features has been on how to incorporate in the image similarity definition, both the similarity between local feature vectors as well as the spatial orientation of the salient points where the local feature vectors were extracted. In [8], for example, it was found that even simply appending the raw (x, y) coordinates of the salient point to its local feature vector improves classification performance for tasks in which translation invariance is not required.

In this paper, we propose the use of cluster co-occurrence matrices (CCM), which can be interpreted as the joint probability of two kinds of local regions to occur at a specific distance and orientation to each other. Thus, this type of CCM is invariant to translation. If required, the CCMs can be made rotation invariant simply by averaging the orientation information. To generate a CCM, the local feature vectors are first clustered using the cluster centroids derived in the previous section. The local feature vectors are no longer needed, they are thus disposed off, and only the assigned cluster indices as well as the location where they were extracted are retained. This process is illustratively shown in Figure 3. The process of generating a CCM from a cluster image can be described as follows:

Let $\mathbf{s}$ be the location (x, y) of a salient point, and $I(\mathbf{s})$ denote the cluster index assigned to the local relational feature vector extracted at $\mathbf{s}$. We define a vector defining the bin boundaries for the distance quantization in the co-occurrence matrix, $\mathbf{D} = (D_1,\, D_2,\, \ldots\, D_{N_d+1})^T$, and another for the angle quantization, $\mathbf{A} = (A_1,\, A_2,\, \ldots\, A_{N_\measuredangle+1})^T$, where N_d and $N_\measuredangle$ are the number of quantization bins used for the radial and angular direction, respectively. The set of all salient point pairs having cluster indices c_1 and c_2 respectively, and located at a specific spatial orientation to each other is then given as:

$$\begin{aligned}\mathbf{S}(c_1, c_2, d, a) \;=\; \{\, (\mathbf{s}_1, \mathbf{s}_2) \;\mid\; & I(\mathbf{s}_1) = c_1 \\ & \wedge\, I(\mathbf{s}_2) = c_2 \\ & \wedge\, D_d < ||\mathbf{s}_1 - \mathbf{s}_2||_2 < D_{d+1} \\ & \wedge\, A_a < \measuredangle(\mathbf{s}_1, \mathbf{s}_2) < A_{a+1}\,\}\end{aligned}$$

The indices run $c_1 = 1, \ldots, N_c$, $c_2 = 1, \ldots, N_c$, $d = 1, \ldots, N_d$, and $a = 1, \ldots, N_\measuredangle$. Furthermore, $\measuredangle(\mathbf{s}_1, \mathbf{s}_2)$ is the angle in the range $[0, 2\pi)$ made by the vector $\mathbf{s}_2 - \mathbf{s}_1$ with the x-axis. The co-occurrence matrix $\mathbf{M}$ consists of the cardinalities of the above sets.

$$\mathbf{M}(c_1, c_2, d, a) = |\,\mathbf{S}(c_1, c_2, d, a)\,|$$

The resulting CCM can be interpreted either as a single 4-D array, or as a series of 2-D CCM matrices, one each for a specific ring sector. To give an idea about the values which can be used in practice, the distance bin boundaries selected in this paper are for example,

$$\mathbf{D} \;=\; (\,0,\; 15,\; 30,\; \ldots,\; 150\,)^T$$

measured in pixels (in comparison, the larger dimension of images in the used database was always 512), and the angle bin boundaries are for example,

$$\mathbf{A} \;=\; (\,0,\; \pi/4,\; \pi/2,\; 3\pi/4,\; \pi\,)^T$$

measured in radians. It can be seen that only the $[0, \pi)$ angle range needs to be covered, as each salient point pair $(\mathbf{s}_1, \mathbf{s}_2)$ would otherwise be counted twice in the matrix (with cluster bins swapped, and in an angle bin which is at an angle π radians from the other). A fuzzy accumulator can also be used to generate the matrix $\mathbf{M}$, but was not investigated in this work.

2.5 Classification

We use Support Vector Machines (SVM) to classify the images based on the above generated feature vectors. SVMs are binary functions (i.e. meant for distinguishing two classes), which find optimal seperating hyperplanes (OSH) for given training data.

In general, the classes may not be linearly seperable in the original feature space. In this case, a feature vector $\mathbf{x}_i$ can be transformed to another (usually higher-dimensional) space as $\phi(\mathbf{x}_i)$ using the mapping ϕ. With the kernel trick [25], it is possible to work in the transformed space without ever calculating the map $\phi(\mathbf{x}_i)$ explicitly. This can be achieved by defining a kernel function $k(\mathbf{x}_i, \mathbf{x}_j) = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle$ as the algorithm needs access only to scalar products between vectors, and not to the actual vectors themselves. Determination of a kernel function appropriate for a given problem remains an open research problem. The most commonly used functions are Linear, Polynomial and Gaussian kernels, as shown in Table 1. General RBF kernels of the form $k(\mathbf{x}, \mathbf{y}) = \exp(-\gamma(d(\mathbf{x}, \mathbf{y}))^2)$ where d is an appropriate distance metric are also popular. In [5], it was shown that kernels based on the L_1 norm are very suitable for accumulator-like features, which is the case with cluster co-occurrence matrices. In this work, we use the histogram intersection kernel defined in Table 1 which is provably positive definite [3].

SVMs in their original form are binary classifiers. Many extensions have been proposed for the multi-class case [14].

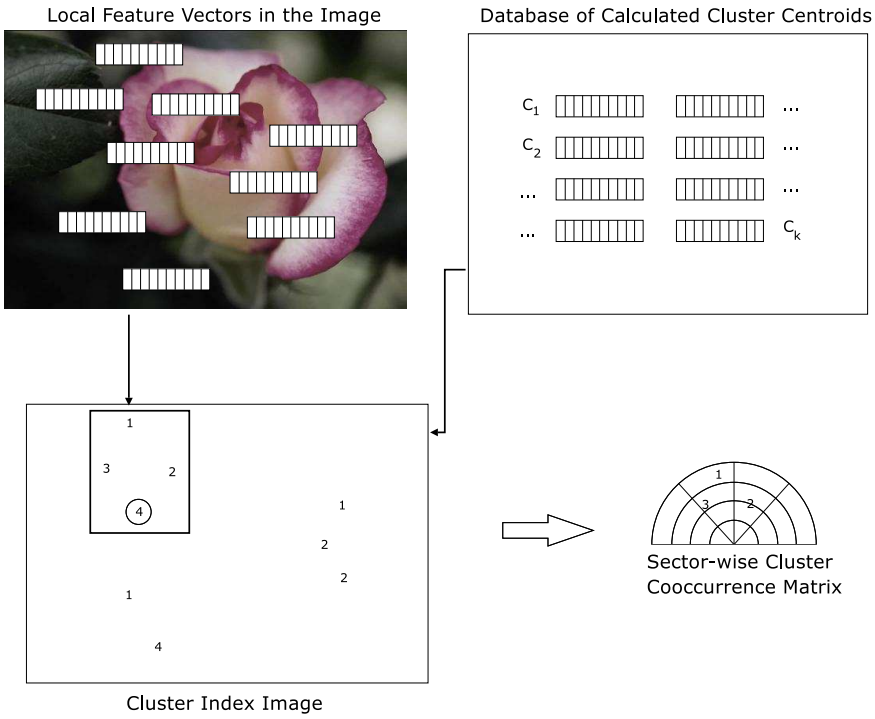

Figure 3: Schematic diagram depicting how the final features are reached. A cluster index image is formed using the local feature vectors around the salient points. For each sector of the ring, a cluster co-occurrence matrix is formed by considering all pairs of salient points whose orientation is the same as that of the sector with respect to the center of the semi-circle. Taking the point with cluster index 4 in the boxed region as an example, the three other interest points would be considered in the co-occurrence matrix.

Table 1: SVM kernels

Kernel	$k(\mathbf{x}, \mathbf{y})$
Linear	$\langle \mathbf{x}, \mathbf{y} \rangle$
Polynomial	$(\gamma(\mathbf{x}_i \cdot \mathbf{x}_j) + \text{coef0})^d,\ \gamma > 0$
RBF	$\exp(-\gamma \|\mathbf{x} - \mathbf{y}\|^2),\ \gamma > 0$
Histogram Intersection	$\sum_{i=1}^{n} \min(x_i, y_i)$

We use here the so-called `one-vs-rest` approach. A binary classifier is trained for each class, with the remaining classes grouped into a single class. A test object is subjected to all trained SVMs, and is assigned the class for which the most positive decision function output is achieved. The LibSVMTL implementation[1] is used for the experiments.

3. EXPERIMENTS AND RESULTS

We test our algorithm on the publicly available[2] IRMA 2005 Radiograph database. The database consists of 9.000 fully classified radiographs taken randomly from medical routine at the Aachen University Hospital, Germany. A further 1000 images are available as test radiographs which have to be classified in one of the 57 pre-defined categories. The categories differ from each other either on account of difference along one of the four axis: image modality, body orientation, body region examined and biological system examined [18].

Some characteristics exhibited by the database are:

- Relatively high number of classes (in total 57, see Figure 4 for examples)
- Overall high intra-class variability (see Figure 5).
- Asymmetric a-priori distribution of classes (some classes have over 1000 training images, while some have less than 20 training images per class).
- Small to medium position variability.
- Mostly upright images, but a few arbitrarily rotated outliers.
- Medium to drastic brightness and contrast variability.

The database was extensively tested during the ImageCLEF 2005 Medical Annotation Task [3]. A total of 12 groups participated in the campaign. The aim of the benchmark is to find out how well current techniques can identify image modality, body orientation, body region, and biological system examined based on the image content. The results of the classification step can be used for multilingual image annotations as well as for DICOM header corrections. The best

[1]http://lmb.informatik.uni-freiburg.de/lmbsoft/libsvmtl/
[2]http://irma-project.org
[3]http://www-i6.informatik.rwth-aachen.de/~deselaers/imageclef05annotation.html

Figure 4: Sample Images from the IRMA05 Database. One good-quality image each from 10 fairly distinct categories are depicted.

Method	Group	Error Rate (%)
<u>This work</u>		
Cluster-Coocurrence Matrices w/ Rel. Features	Uni Freiburg	
- $20 \times 20 \times 10 \times 4$ matrix, 1000 salient points		**8.1**
- $20 \times 20 \times 10 \times 4$ matrix, 600 salient points		8.9
- $15 \times 15 \times 10 \times 4$ matrix, 1000 salient points		9.1
<u>Previous Best</u>		
Sparse Histograms w/ Position[8]	RWTH-Aachen	
- using maximum entropy classification		9.3
- using support vector machine		10.0
<u>ImageCLEF 2005 Benchmark, Top-7 results</u>		
Image Distortion Model[16]	RWTH Aachen	12.6
Image Distortion Model & Texture	IRMA-Group	13.3
Patch-Based Classifier (Maximum Entropy)	RWTH Aachen	13.9
Patch-Based Classifier (Boosting)	Uni-Liège	14.1
Image Distortion Model & Texture	IRMA-Group	14.6
Patch-Based Classifier[21] (Decision Trees)	Uni-Liège	14.7
GNU Image Finding Tool (GIFT)	Uni Geneva	20.6
<u>Baseline Results</u>		
32×32 images as feature, 1-NN/L_2 classification	-	36.8

Table 2: Results for the IRMA 05 Database. The comparison results are taken from the ImageCLEF 2005 Benchmark, and from a recent improvement we are aware of. The complete ImageCLEF results can be viewed at the URL *http://www-i6.informatik.rwth-aachen.de/~deselaers/imageclef05_aat_results.html*.

results obtained in the 2005 benchmark was 12.6 % (measured as error rate for 1000 test images), which was improved later on by Deselaers et al [8] to 9.3 %. The method proposed here already reaches a best of 8.1 %. Our results and a selection of published results are shown in Table 2. It should be mentioned that due to the large number of parameters (see discussion below), a joint optimization was not feasible, and each parameter was only individually optimized during cross-validation.

Our group participated in the 2006 ImageCLEF Medical Annotation Task by applying the proposed method to the newly available database without further modification. The

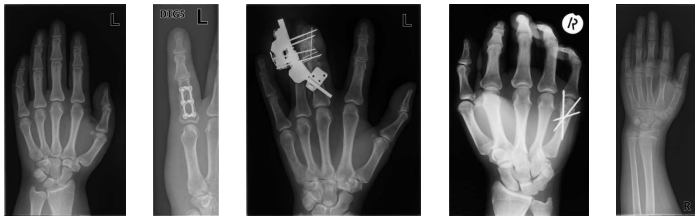

Figure 5: Intra class variability: Five images from the class described as "*x-ray, plain radiography, coronal, upper extremity (arm), hand, musculosceletal system*". As can be observed, the variability includes brightness changes, partial occlusion, and translation, apart from some inevitable human to human variablility.

new database contains X-ray images which are more finely granuled in 116 classes. Our method achieved the second best results from 27 submissions. The complete results along with the task description can be viewed on the web[4], and are not repeated here for brevity.

4. DISCUSSION

Relational Features vs. Gray-value features

Gray-value patches as features over interest points have been used in various works. In [9], robustness towards additive illumination changes is achieved by performing a PCA transformation, and simply discarding the first coefficient. While in many cases this does hold true, it is possible that important discriminative information is also lost in the process, as the PCA is performed over the ensemble containing patches from all classes. On the other hand, the features proposed in this work are implicitly invariant towards additive illumination changes, and due to the clamping performed by the `rel` operator, also robust towards other kinds of monotonic illumination changes. Furthermore, the features can be elegantly adjusted to work in different scenarios. For example,

- To achieve robustness against desired transformations, so called virtual samples are often generated from training data, by subjecting the training examples to small transformations. In case robustness to small rotations is needed, the existing relational features for each (r_1, r_2) pair can simply be circularly shifted by a small displacement (say one unit shift), to achieve the desired effect.
- On the other hand, if complete invariance towards rotation is desired, the existing relational feature subvector for each (r_1, r_2, ϕ) combination need only be summed up to get a rotation-invariant descriptor.

Scale Robustness

Incorporating robustness to scale changes is trickier but possible. First of all we propose to use an interest point detector which can deliver a scale ζ for every interest point $\mathbf{s}$ (for example, the Difference-of-Gaussian detector as used in SIFT). The relational features can be adapted to the new scale information simply by mapping the parameter set (r_1, r_2, ϕ) to $(\zeta r_1, \zeta r_2, \phi)$. However, the scale information must still be incorporated in the radial quantization performed during the calculation of the co-occurrence matrix. This can be done by redefining the distance measure between two salient points as $d(\mathbf{s}_1, \mathbf{s}_2) = ||\mathbf{s}_2 - \mathbf{s}_1||/\tilde{\zeta}$ where $\tilde{\zeta}$ is appropriately derived from the scale information of the two salient points, e.g. $\tilde{\zeta} = \sqrt{\zeta_1 \zeta_2}$, under the assumption that the scale changes are smooth over the image.

[4]http://www-i6.informatik.rwth-aachen.de/~deselaers/imageclef06/medaat-results.html

Tunable Parameters

An important issue to discuss is the number of tunable parameters or the choices that one has to make at each feature extraction or classification step. Some of these issues are to be tackled by any image classification algorithm, while some are unique to our approach.

Number of Salient Points, N_s

One option is to select a threshold and take all points which have a higher saliency value in the saliency map. We however take a fixed number of salient points per image by sorting the saliency map. For this particular database, N_s between 500 and 1000 gives best results.

Number of Clusters, N_c

This parameter is often choosen heuristically and is required as an input to most clustering algorithms. We found that our algorithm is not very sensitive to small changes to N_c. Values of $N_c < 15$ leads to a scenario that the feature differences are not satisfactorily modelled. N_c between 15 and 30 gives good results, and increasing N_c even further does not lead to any improvement. Since the size of the co-occurrence matrix increases quadratically with N_c, we choose $N_c = 20$ for most experiments.

Number of bins for spatial quantization, N_d *and* $N_\measuredangle$

For the radial direction, we conducted experiments to determine the radial range needs to be covered, and choose the range upto 150 pixels, which is about a third of the image size. This was quantized in $N_d = 10$ bins, which was only coarsely determined through experimentation. For angle, we showed before that only the $[0, \pi)$ range needs to be covered. We experimented with 3, 4 and 6 bins in this angle range, finally selecting 4 evenly sized angular bins.

Length of Final Feature Vector

As stated earlier, the dimensions of the final cluster co-occurence matrix are given by $N_c \times N_c \times N_d \times N_\measuredangle$. For the final values chosen after cross-validation, this results in a 16000-dimensional feature vector per image. It is clear that the information contained in the vector is sparse and should be compressible. In this work, we do not perform any further feature selection or extraction steps, but still provide

some general ideas, in case it becomes necessary for much larger databases. A seemingly paradoxical advantage gained by the high dimensions could be that simple functions sets (e.g. linear) can be used for classification, perhaps avoiding overfitting.

Principal Component Analysis (PCA)

PCA is a commonly used technique for dimensionality reduction. The basic idea behind PCA is to find another orthogonal coordinate system such that each subsequent dimension explains as much of the remaining variance in the given data. For correlated data, it is not uncommon to find that most of the variance (say, above 99 %) is contained in a small fraction of the number of final dimensions. However, care should still be taken while selecting the number of retained PCA dimensions, as the low amplitude information contained in the higher PCA coefficients might still be critical for discrimination between different classes.

Feature Selection

We include this paragraph for the sake of completeness. There exist various feature selection strategies in the machine learning literature. The interested reader can refer to [17] for an overview of some of the popular alternatives. The main distinction is between so-called *filter* methods, which compute a ranking for the features without taking the inducer (classifier) into account, and the *wrapper* methods, which search in the set of subsets of features for the optimum subset for the specific inducer. However, the high number of classes in this case can make it difficult to perform feature selection, as different features might be discriminative for different classes.

Further Feature Extraction

Gray-value co-occurrence matrices have been used extensively as texture descriptors in the image processing literature. The high-dimensionality has been tackled there by using instead of the complete co-occurrence matrix, some statistical properties derived from it. For example, Haralick et al [13] proposed 10 statistical properties which can be derived from a 2-D GLCM, namely Entropy, Energy (Angular Second Moment), Contrast, Homogeneity, SumMean (Mean), Variance, Correlation, Maximum Probability, Inverse Difference Moment, and Cluster Tendency. As stated before, our 4-D CCM can be interpreted as a series of 2-D CCMs. It is however, important to note that only 3 of the above given properties, namely Entropy, Energy and Maximum Probability can be adapted to use in a CCM. This is because, unlike a GLCM, in which neighbouring rows or columns indicate similar regions (gray value), the clusters in CCMs cannot be ordered linearly, and thus properties such as contrast ($\sum_{i=1}^{N_c}\sum_{j=1}^{N_c}(i-j)^2C(i,j)$, where C is a 2-D CCM) are simply meaningless in our context. In the future, the possibility of extracting further statistical information from a CCM would be considered.

Computation Time

The experiments were carried out on an AMD Opteron 2.4 GHz machine running Debian Linux. Relational feature calculation per image takes less than 0.5 s for an image of size approximately 512×512. K-means clustering with 20 clusters takes about 3 minutes. It takes slightly less than a second to compute one 4-D co-occurrence matrix. Inspite of the large feature vector, SVM training with 9000 feature vectors of dimension 16000 takes less than 2 hours, most of the time being spent in calculating the kernel matrix. Final classification of 1000 image vectors takes about 10 minutes, or about $0.5s$ per image.

5. CONCLUSIONS

A content based image classification system was presented in this paper. The main novelties are the use of local relational features for illumination-robust general-purpose image classification, and the introduction of cluster co-occurrence matrices which incorporate elegantly the spatial information of the interest points in the matching process. The method obtains without much tuning the best results published so far for a publicly available medical radiograph database. Still, the main concerns are the many possibilities to choose various parameters, and the high dimensionality of the final feature vector. We discussed these and other issues in detail and provided alternatives where necessary.

6. ACKNOWLEDGMENTS

We would like to thank Dr. TM Lehmann, Dept. of Medical Informatics, RWTH Aachen, Germany for making the IRMA 05 Database available for research purposes. This work was partially supported by the MUSCLE Network of Excellence[5] through contract #507752.

7. REFERENCES

[1] S. Agarwal, A. Awan, and D. Roth. Learning to detect objects in images via a sparse, part-based representation. *PAMI, IEEE Transactions on*, 26(11):1475–1490, November 2004.

[2] J. Amores, N. Sebe, and P. Radeva. Efficient object-class recognition by boosting contextual information. In *IbPRIA (1)*, pages 28–35, 2005.

[3] A. Barla, F. Odone, and A. Verri. Histogram intersection kernel for image classification. In *ICIP (3)*, pages 513–516, 2003.

[4] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts, 2001.

[5] O. Chapelle, P. Haffner, and V. Vapnik. Svms for histogram-based image classification, 1999.

[6] G. Csurka, L. Dance, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *ECCV Workshop on statistical Learning in computer vision*, pages 59–74, 2004.

[7] L. Davis, S. Johns, and J. Aggarwal. Texture analysis using generalized co-occurrence matrices. *PAMI*, 3:251–259, 1979.

[8] 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*, Lecture Notes in Computer Science, page accepted for publication, Berlin, Germany, September 2006.

[9] T. Deselaers, D. Keysers, and H. Ney. Discriminative training for object recognition using image patches. In *Proceedings of the International Conference on CVPR*, volume 2, pages 157–162, San Diego, CA, June 2005.

[5] http://muscle-noe.org

[10] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples an incremental bayesian approach tested on 101 object categories. In *Workshop on Generative-Model Based Vision*, Washington, DC, June 2004.

[11] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *Proceedings of the IEEE Conference on CVPR*, volume 2, pages 264–27, Madison, Wisconsin, 2003.

[12] M. O. Gueld, M. Kohnen, D. Keysers, H. Schubert, B. B. Wein, J. Bredno, and T. M. Lehmann. Quality of DICOM header information for image categorization. In *Proc. SPIE Vol. 4685, p. 280-287, Medical Imaging 2002:* , pages 280–287, May 2002.

[13] Haralick, Shanmugam, and Dinstein. Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-3(6):610–621, 1973.

[14] C. Hsu and C. Lin. A comparison of methods for multi-class support vector machines, 2001.

[15] J. Huang, S. Kumar, M. Mitra, W. Zhu, and R. Zabih. Image indexing using color correlograms, 1997.

[16] D. Keysers, C. Gollan, and H. Ney. Classification of medical images using non-linear distortion models. In *Bildverarbeitung für die Medizin*, pages 366–370, 2004.

[17] R. Kohavi and G. H. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97(1-2):273–324, 1997.

[18] T. M. Lehmann, H. Schubert, D. Keysers, M. Kohnen, and B. B. Wein. The IRMA code for unique classification of medical images. *Medical Imaging 2003: Proceedings of the SPIE, Volume 5033, pp. 440-451 (2003).*, pages 440–451, May 2003.

[19] J. Li and J. Z. Wang. Automatic linguistic indexing of pictures by a statistical modeling approach. *IEEE Transactions Pattern Analysis and Machine Intelligence*, 25(9):1075–1088, 2003.

[20] E. Loupias and N. Sebe. Wavelet-based salient points for image retrieval, 1999.

[21] R. Marée, P. Geurts, J. Piater, and L. Wehenkel. Biomedical image classification with random subwindows and decision trees. In *Proc. ICCV workshop on Computer Vision for Biomedical Image Applications (CVIBA 2005)*, volume 3765 of *LNCS*, pages 220–229. Springer-Verlag, oct 2005.

[22] T. Ojala, M. Pietikäinen, and T. Mäenpää. Gray Scale and Rotation Invariant Texture Classification with Local Binary Patterns. In *Proc. Sixth European Conference on Computer Vision*, pages 404–420, Dublin, Ireland, 2000.

[23] M. Schael. *Methoden zur Konstruktion invarianter Merkmale für die Texturanalyse*. PhD thesis, Albert-Ludwigs-Universität, Freiburg, June 2005.

[24] N. Sebe and M. S. Lew. Comparing Salient Point Detectors. *Pattern Recognition Letters*, 24(1-3):89–96, January 2003.

[25] V. N. Vapnik. *The Nature of Statistical Learning Theory (Information Science and Statistics)*. Springer, 1999.

[26] J. Vogel. *Semantic Scene Modeling and Retrieval*. PhD thesis, ETH Zurich, October 2004.

[27] M. Weber, M. Welling, and P. Perona. Unsupervised learning of models for recognition. In *In Proceedings of the 6th European Conference of Computer Vision*, Dublin, Ireland, 2000.