Computer-aided interpretation approach for optical tomographic images

Christian D. Klose*
Columbia University
Department of Biomedical Engineering
New York, New York 10027

Alexander D. Klose
Columbia University
Department of Radiology
New York, New York 10027

Uwe J. Netz*
Charité - Universitätsmedizin Berlin
Institut für Medizinische Physik und Lasermedizin
14195 Berlin, Germany

Alexander K. Scheel*
Georg-August-Universität Göttingen
Department of Medicine
Nephrology and Rheumatology
37075 Göttingen, Germany

Jürgen Beuthan
Charité - Universitätsmedizin
Institut für Medizinische Physik und Lasermedizin
14195 Berlin, Germany

Andreas H. Hielscher
Columbia University
Department of Biomedical Engineering
Department of Radiology
Department of Electrical Engineering
New York, New York 10027

Abstract. A computer-aided interpretation approach is proposed to detect rheumatic arthritis (RA) in human finger joints using optical tomographic images. The image interpretation method employs a classification algorithm that makes use of a so-called self-organizing mapping scheme to classify fingers as either affected or unaffected by RA. Unlike in previous studies, this allows for combining multiple image features, such as minimum and maximum values of the absorption coefficient for identifying affected and not affected joints. Classification performances obtained by the proposed method were evaluated in terms of sensitivity, specificity, Youden index, and mutual information. Different methods (i.e., clinical diagnostics, ultrasound imaging, magnet resonance imaging, and inspection of optical tomographic images), were used to produce ground truth benchmarks to determine the performance of image interpretations. Using data from 100 finger joints, findings suggest that some parameter combinations lead to higher sensitivities, while others to higher specificities when compared to single parameter classifications employed in previous studies. Maximum performances are reached when combining the minimum/maximum ratio of the absorption coefficient and image variance. In this case, sensitivities and specificities over 0.9 can be achieved. These values are much higher than values obtained when only single parameter classifications were used, where sensitivities and specificities remained well below 0.8.
© *2010 Society of Photo-Optical Instrumentation Engineers.* [DOI: 10.1117/1.3516705]

1 Introduction

Recently work in the field of diffuse optical tomography (DOT) has progressed from purely theoretical studies and bench-top experiments to first clinical trials that explore the utility in breast cancer diagnosis,[1–3] brain imaging,[4–6] and arthritis detection.[7–13] While substantial advances have been made in building clinically useful instruments, and developing an image reconstruction algorithms, much less effort has been spend on developing image analysis tools. Other medical imaging fields such as magnet resonance imaging (MRI), computer tomographic imaging (CT), and ultrasound (US) imaging frequently make use of advanced image analysis methods that enhance sensitivity and specificity in many cases. For example computer-aided diagnostics (CAD) systems have been successfully employed in areas such as mammography,[14] chest CT (Refs. 15 and 16), and brain imaging.[17,18] In biomedical optics, CAD has only been applied in two studies related to optical coherence tomography (OCT), which explored its utility in esophagial and cervical cancer.[19,20] To the best of our knowledge, no studies have been presented where CAD was employed in the analysis of DOT images.

In this paper a CAD system is introduced to enhance the analysis of sagittal laser optical tomographic (SLOT) images obtained from proximal interphalangeal (PIP) joint of patient with rheumatoid arthritis. These SLOT images display spatially varying absorption coefficient μ_a and scattering coefficient μ_s across the joint. Previous studies evaluated the potential of using features such as minimal and maximal absorption or scattering coefficients or their ratios in a region of interest, to distinguish between affected and not affected joints.[11] Using the minimal absorption coefficient as an image feature, sensitivities and

*As of 2009-08-01, C.D. Klose is with Think GeoHazard, One Columbus Place #N34C, New York, NY 10019. Furthermore, as of 2008-10-01, U.J. Netz is with Laser- und Medizin-Technologie Berlin (LMTB), Fabeckstraße 60-62, 14195 Berlin, Germany; and A.K. Scheel is now with the Johann Wolfgang Goethe Universität, Dept. of Rheumatology, Theodor Stern Kai 7, 60590 Frankfurt, Germany.

Address all correspondence to: Andreas Hielscher, Columbia University, Department of Biomedical Engineering, 500 West 120th St, New York, New York, 10027. Tel: 212-854-5080; Fax: 212-854-8725. E-mail: ahh2004@columbia.edu.
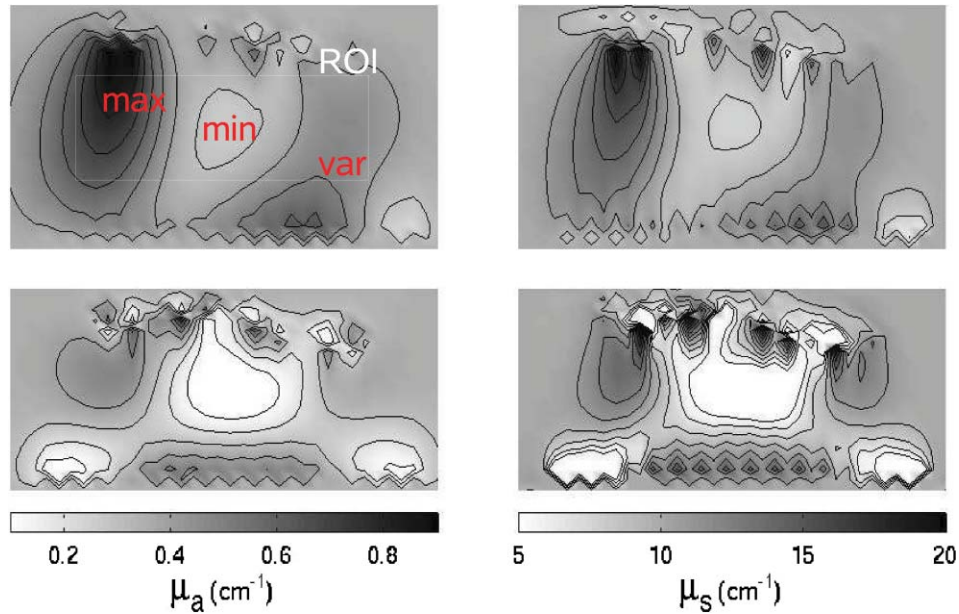
Fig. 1 Example of SLOT images. Shown are 2×3.5-cm sagittal cross section through the PIP joints of middle fingers. The tip of the finger is to the right and base of the finger is to the left. The joint cavity is in the center of the image. The images display the absorption coefficient μ_a and the scattering coefficient μ_s, e.g., for not affected finger (lower row) and RA affected finger (upper row). Parameters of minimum and maximum values can be extracted within an area of interest ROI. These parameters, in turn, can be used for multiparameter classification and interpretation.

specificities of 0.71 could be achieved in identifying affected joints assuming that ultrasound can be considered as an imaging gold standard that provides ground truth.

This study goes beyond previous analyses in several ways. First, it includes the combination of multiple features (e.g., minimum absorption coefficient and ratio of maximum and minimum absorption coefficient) in addition to evaluating classification performances using only a single feature. We also add image features previously not considered, such as the variance of optical properties in the images. Furthermore, instead of using only US as gold standard that provides the ground truth, we evaluate the classification performance using additional ground truth derived from MRI, clinical evaluation, and visual inspection of optical tomographic imaging itself. Finally, a larger data set was used, which includes optical tomographic images of 100 finger joints. These images were obtained with a SLOT system that showed better SNRs and long-term measurement stability than systems used in previous studies.[11]

To deal with the problem of multiparameter classification we employ in this work a machine learning tool that explores, sorts through, and interprets tomographic image data.[21] Classifications were performed by an interpretation system based on self-organizing mapping (SOM). The classification technique was originally developed as a physical-mathematical model to mimic the human's visual system.[22, 23] Indeed, this method has been used in the past in other scientific fields for similar classification problems.[21, 24, 25] It has shown to produce significant better results than approaches based on discriminant analysis and logistic regression.[26, 27] To see if this holds true in our case, we also performed a discriminant analysis and compared the results with the SOM machine learning approach.

In the reminder of the paper, we first describe in detail the data used for the analysis. This is followed by a detailed description of the machine-learning-based classification approach applied in

this paper. Subsequently, the results obtained with this approach are presented and discussed.

2 Optical Tomography Data

Data sets were analyzed resulting from tomographic reconstructions of SLOT images to determine best image interpretation results. An example of a SLOT image is shown in Fig. 1. These images were generated by measuring the transmitted light intensities along the central axis of the index, middle, and ring fingers on the left and right hands. The light source was a laser with wavelength $\lambda = 675$ nm, which was focused to $\approx$1-mm spot on 11 different position on the back of each finger. For each position the transmitted light intensities were measured with a Si photo diode. This transmission data became input to a model-based iterative image reconstruction code that used the equation of radiative transfer as light propagation model. For a more detailed description of the experimental setup and the image reconstruction code see Refs. 10, 28, and 29.

In total, 100 optical tomographic images of human finger joints were used in this study. A region of interest (ROI) was defined within each image to prepare the images for CAD analysis. Data were eliminated in the first 4 mm on the top and bottom of each image and 7 mm on the left and right. In this way, the chosen ROI did not contain potential image artifacts, which are often encountered near source and detector positions (image boundaries). Within the ROI, different parameters of the absorption coefficients where extracted including the smallest value min($\cdot$), maximum value max($\cdot$), mutual ratios [e.g., min($\cdot$)/max($\cdot$)], and statistical variance var($\cdot$). All extracted image features were combinatorially combined. Thus, each image was characterized by a n-dimensional feature vector $\boldsymbol{x}_n$ consisting of a set of n image features. These feature vectors became input to a machine learning tool classifying each
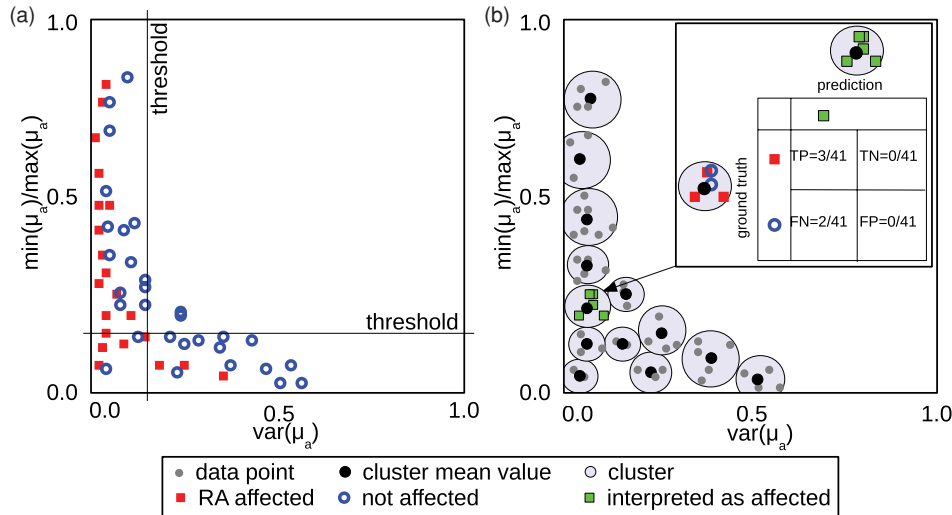
Fig. 2 Schematic illustration of the clustering problem and the validation of the clustering results: (a) data distribution in a 2-Dimensional feature space and the assignment of affected and not affected finger joints based on the US-derived ground truth. Single parameter classifications using parameter thresholds lead to misinterpretations. This can be reduced when using multiparameter classifications. (b) A SOM neural network separates the same data set into disjoint subsets (clusters). Each cluster assigns the cluster members to a certain class (here: affected). Assignments depend on the probability threshold p_t, which changes the interpretation/prediction outcome of the classification with respect to the benchmark.

image as a finger affected or not affected by rheumatic arthritis (RA).

To perform this analysis, one requires a ground-truth benchmark, that identifies each patient as affected or not affected by RA. In previous works US images were considered to provide such ground truth.[11] Many researchers consider magnetic resonance images of finger joints as the most accurate indicator for RA. However, no studies have been presented that MRI is indeed the most accurate ground truth. This would require longitudinal studies spanning many years of follow up to establish the predictive and prognostic value of each imaging method. A study like this has not yet been performed. Therefore, we decided to report on the performance of our CAD system for different ground truths, derived from different sources, including MRI, US, clinical evaluation (CE), and optical inspection of SLOT images. For each modality experts scored images and data on a four-point scale: 0 for definitely no synovitis, 1 for probably no synovitis, 2 for possibly synovitis, and 3 for definitely synovitis. Subsequently each finger was labeled by only two different classes: not affected class $c_0 = \{$definitely and probably no synovitis$\}$ and affected class $c_1 = \{$possibly and probably synovitis$\}$ (Figs. 1 and 2). Feature vectors $x_{\{n,c\}}$ containing various optical parameters were labeled accordingly and the performance of the CAD system was evaluated using various performance measures, including sensitivity and specificity.

3 Methodology

3.1 *Machine-Learning-Based Classification Method*

In the most general case, the goal of any medical classification scheme is to determine ranges of diagnostic parameters for which a person is found to be healthy or afflicted by a certain disease. In this study, we look for features in the optical tomographic images that can be used to determine whether a patient is affected by RA or not affected. In previous works, only single features were considered. For example, we found that patients

with min $(\mu_a) > 0.272$ cm^{-1} should be considered affected by RA, while patients with min$(\mu_a) < 0.272$ cm^{-1} should be classified as healthy.[11] However, as mentioned in the introduction, using this criteria, we only achieved a sensitivity and specificity of about 0.71.

If multiple features are used, as in the study at hand, the simple cutoff value $[\min(\mu_a)] = 0.272$ cm$^{-1}]$ has to be replaced by a more complex rule. If n features are considered, a classification scheme seeks an $(n-1)$-dimensional hyperplane that separates the n-dimensional space into two regions: one region characteristic for patients affected by RA, and another region that characterizes unaffected people.

In the approach presented here, a separation into two regions is achieved in two steps. First, an SOM algorithm is used to cluster the given training data of feature vectors x_i of affected and healthy fingers. The main variable of this algorithm is the cluster size. Therefore, if k feature vectors are presented and the cluster size is q, the algorithm returns about k/q clusters that each contain q feature vectors (also called members). The members of each cluster are typically located close to each other in the n-dimensional feature space [see, for example, the clusters in Fig. 2(b)]. At this point, clusters may contain feature vectors belonging to images of only healthy, only affected, or a mixture of affected and healthy joints. Details of the SOM-based clustering scheme can be found in the appendix (Sec. 6).

In the next step, all members belonging to a given cluster are assigned to either the affected class or the not-affected class, depending on a threshold p_t that is set the same for all clusters. Therefore, if at least $p_t\%$ of all members of a given cluster are affected, that cluster will be assigned to the affected class. If less than $p_t\%$ of all members of a given cluster are affected, that cluster will be assigned to the nonaffected class. At the end of this process the n-dimensional feature space has been divided into clusters that represent features of unaffected and affected patients. If now a new feature vector, derived from an optical tomographic image of a finger of a new patient, is considered,

this new feature vector will "fall" into one of these clusters and the patient will be declared affected or unaffected by RA, depending on the status of that cluster.

The following example illustrates how this classification approach works. Figure 2(a) shows the distribution of a 2-D feature vector, with the variance (μ_a) as first component (x axis) and $\min(\mu_a)/\max(\mu_a)$ as the second component (y axis). The red squares identify the feature vectors belonging to images of affected joints, while the blue circles identify the feature vectors belonging to images of not-affected joints, as determined, in this case, by using the US-derived ground truth. One can see that if one would attempt to classify a feature vector (representing an image) as affected or not affected using only threshold values for either $\text{var}(\mu_a)$ or $\min(\mu_a)/\max(\mu_a)$, it would lead to a large number of misclassifications. For example, postulating that all fingers with $\text{var}(\mu_a) < 0.2$ are affected would lead to three false negatives [the three red squares to the right of the threshold in Fig. 2(a)] and 16 false positives [the 16 blue circles to the left of the threshold in Fig. 2(a)]. Similarly postulating that all fingers with $\min(\mu_a)/\max(\mu_a) < 0.2$ are affected, this would lead to 13 false negatives [red squares above the threshold in Fig. 2(a)] and 14 false positives [blue circles below the threshold in Fig. 2(a)]. Therefore, classifications on only one feature would be highly flawed.

As stated, the SOM method is used to partition a given data set x_n into subregions or clusters [Fig. 2(b)]. In this particular example the cluster number was set to 13. Thus, each feature vector belongs to a subregion that includes on average five data points.

In the next step, a given cluster is either assigned to the affected class or not-affected class depending on a threshold p_t that is set the same for all clusters. In the given example [see inset of Fig. 2(b)] the cluster is populated with three data points representing affected joints (red squares) and two data points representing not affected joints (blue circles). Therefore, choosing a frequency threshold of, for example, $p_t > 50\%$ will result in assigning the cluster to the affected class, since 60% of its members are indeed affected. On the other hand, if $p_t > 80\%$ is chosen, the cluster will be assigned to the unaffected class, since less than 80% of its members are actually affected.

Choosing different cluster size q and threshold p_t will result in different separations of this 2-D feature spaces into regions typical for affected and unaffected joints. These differences in separations will lead to different classification results. In general, classification performance increases as the cluster size gets smaller (which is equivalent to the number of subgroups gets larger) until an optimum is reached when misclassification is at a minimum. Thus, a too small number of clusters leads to "underfitting" of a given data set, whereas a too large number may lead to a data "overfitting." In both cases, misclassification increases.[21,30]

In this work, cross-validation is employed to determine the optimum cluster size of maximum classification performance with respect to a random sampling of the data points. The number of clusters is varied from 1 to 100 and p_t from 0 to 100. For each combination the classification performance was evaluated by using a leave-q-out approach.[30] Therefore, the given n-dimensional data manifold of $L = 100$ realizations was randomly split into q disjoint subsets (e.g., $q = 10$). The accuracy was determined when performing SOM classification with $q-1$

of q subsets (learning step) and applying it to the remaining 1 of q subsets (validation step). By leaving one subset out, the procedure was conducted q times and the mean and standard deviation of various performance measures (see Sec. 3.3) were calculated.

The meta-algorithm shown below provides a summary of the classification and interpretation procedure of any given n-dimensional data manifold of image features, e.g., drawn from tomographic images. More details of the SOM clustering algorithm are given in the appendix (Sec. 6).

> Set dimensionality n feature space X_n
> Set cluster number l
> Set target classes for "ground truth"-benchmarks $\{c_0, c_1\}$
> Set all j "ground truth"-benchmarks (MRI, CE, US, SLOT)
> Generate an ensemble of L SOM with inital and final learning rates
> BEGIN Loop for each SOM(l) $m = 1 \rightarrow L$
>   Partition all x_n into l subgroups w_n
>   BEGIN Loop for each $p_t = 0 \rightarrow 100$
>   Calculate $S_e(\cdot)$
>   Calculate $S_p(\cdot)$
>   Calculate $J(\cdot)$
>   Calculate $I(\cdot)$
>   END Loop
> END Loop
> Determine the maximum argument of J or I as best interpreted classification result

The objective in the remainder of this paper is to analyze a given data with respect to (1) dimensionality of the feature space n, (2) structure of the SOM (e.g., cluster number l, learning rates, etc.), and (3) ground truth benchmark with target classes (e.g., $\{\hat{c}_0, \hat{c}_1\}$).

3.2 Discriminant Analysis

Discriminant analysis (DA) was also applied to quantify classification performances with a more traditional statistical analysis method.[31] Equivalent to the SOM machine learning approach, the goal of the discriminant analysis is to separate/predict group members (e.g., RA-affected and not affected) from a set of predictors [e.g., $\min(\mu_a)$, $\max(\mu_a)$, min/max, and $\text{var}(\mu_a)$]. For this purpose, discriminant function scores and statistical significances are estimated to determine the best linear combination of predictors. A discriminant function is predicted for a case from the sum of the series of predictors, which are, in turn, weighted by a coefficient. Thus, each discriminant function is based on a set of coefficients. Performance measures as described below can be used to quantify the quality of classifications/predictions (see Sec. 3.3). In this study, the JMP software package was used to perform the DA. More details on the DA can be found, for example, in Ref. 31.

3.3 Performance Measures

To quantify the classification performance the following four measures were considered. First, the sensitivity S_e and the speci-
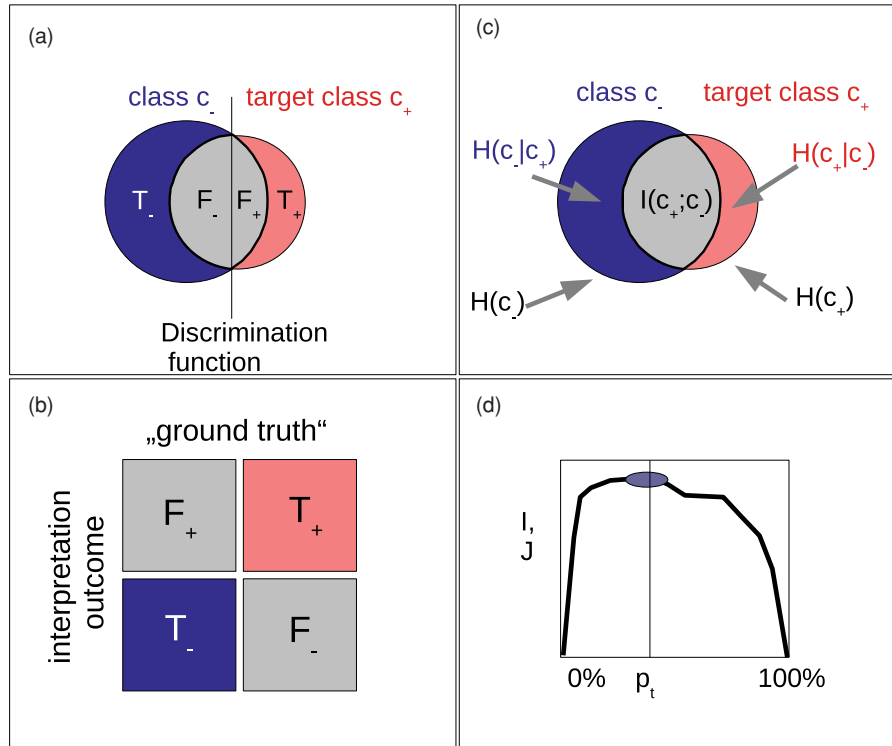
Fig. 3 Classification performance: (a) and (b) measuring true/false positive values T_+/F_+, true/false negative values T_-/F_- of a classification when compared to a ground truth benchmark [see Figs. 2(b)] and (d) and (c) measuring the mutual information I between the interpreted/predicted class labels c of the feature values $\boldsymbol{x}$ and a ground truth benchmark, (d) performance measures Youden index $J(T_+, F_+, T_-, F_-)$ and mutual information $I(T_+, F_+, T_-, F_-)$ change as a function of the frequency threshold p_t (Figs. 10 and 11 and Table 2 later in the paper).

ficity S_p are defined as[32]

$$S_e = \frac{T_+}{(T_+ + F_-)}, \qquad (1)$$

$$S_p = 1 - \frac{F_+}{(F_+ + T_-)}, \qquad (2)$$

where true positive values $T_+ = \sum t_+$, true negative values $T_- = \sum t_-$, false positive values $F_+ = \sum f_+$, false negative values $F_- = \sum f_-$ [Figs. 3(a) and 3(b)]. Therefore $S_e = [0,1]$ is the relative number of all $\boldsymbol{x}_n$ vectors that are truly identified (t) as the target class ($+$) with respect to the ground truth; and $S_p = [0,1]$ is the relative number of all $\boldsymbol{x}_n$-vectors that were falsely identified (f) as the target class ($+$). Using the sensitivity and specificity one can calculate the, third measure, the Youden index[33] $J = S_e + S_p - 1$.

Furthermore, by varying p_t from 0 to 100%, ROCs were generated and analyzed as they are frequently used in the characterization of medical classification schemes. If p_t is set to 0, all images will be qualified as not affected leading to a sensitivity of 0, however, specificity is 1. If p_t is set to 100%, the specificity will be 0. Intermediate p_t values lead to intermediate S_e and S_p values usually maximizing the J for a given pair. It should be pointed out that this approach differs from classical ROC analysis, which is typically applied to only on observable parameter,[11] e.g., $\min(\mu_a)$. By varying the threshold of μ_a for which a patient is considered affected S_e and S_p can be calculated and ROC curves generated. By introducing p_t as threshold, we effectively extended the ROC analysis to multiple parameter interpretation in the frame work of SOM neural networks.

The fourth and more generalized performance measure[34] is the mutual information $I[C(\boldsymbol{x}, \boldsymbol{w}); \hat{C}]$:

$$H(C|\hat{C}) = H(C, \hat{C}) - H(\hat{C})$$

$$I(C; \hat{C}) = H(C) - H(C|\hat{C})$$

$$= \sum_{c \in C} \sum_{\hat{c} \in \hat{C}} p(c, \hat{c}) \log \left[\frac{p(c, \hat{c})}{p(c)\, p(\hat{c})} \right]. \qquad (3)$$

Note $I[C(\boldsymbol{x}, \boldsymbol{w}); \hat{C}]$ expresses the similarity between the amount of data vectors labeled as class $\hat{C}$ of the ground truth benchmark and the interpreted/predicted data vectors labeled as class $C(\boldsymbol{x}, \boldsymbol{w})$, which were estimated by the the SOM neural network. Also $I(\cdot)$ is 1 when the class labels of all interpreted data vectors match with the labels of the ground truth.

4 Results and Discussion

We start our analysis by plotting the distributions of the four single parameters $\max(\mu_a)$, $\min(\mu_a)$, $\min(\mu_a)/\max(\mu_a)$, and $\mathrm{var}(\mu_a)$ for unaffected and affected finger joints as identified with the four different ground truth, derived from CE, MRI, US, and SLOT (Figs. 4 to 7). The machine intelligent classification was performed entirely independent from the visual inspections of the CE, MRI, US, and SLOT data. Thus, researchers did not have any knowledge of the outcomes from other methods.

The green triangles indicate the mean and standard deviation of the data distribution. Looking at these figures, we can observe several things. First, we notice that the distributions for affected
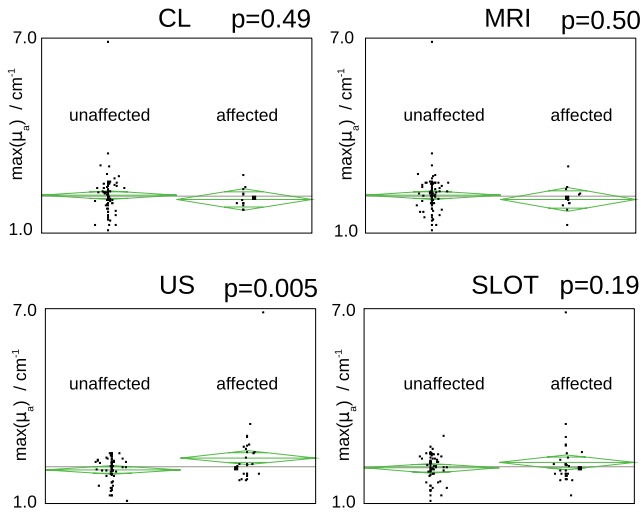
Fig. 4 Statistical distributions of the maximum absorption coefficient $\max(\mu_a)$ with respect to RA-affected and unaffected finger groups and ground truth derived from CL, MRI, US, and SLOT; p values resulting from analysis of variance (ANOVA) that are less than 0.05 indicate statistical significant differences between both groups.

Fig. 6 Statistical distributions of the minimum/maximum-ratio of the absorption coefficient $\min(\mu_a)/\max(\mu_a)$ with respect to RA-affected and unaffected finger groups and ground truth derived from CL, MRI, US, and SLOT; p-values resulting from ANOVA that are less than 0.05 indicate statistical significant differences between both groups.

and unaffected fingers for all four parameters are very similar given US and SLOT as ground truth. This indicates that SLOT and US will show similar classification results. The distributions for MRI as ground truth, resemble closer the distribution found for CE.

Furthermore, we observe that the distributions for $\max(\mu_a)$ are very similar for affected and unaffected finger joints across all four ground truths (Fig. 4). This indicates that $\max(\mu_a)$ is a very poor classifier. The plots for $\min(\mu_a)$, $\min(\mu_a)/\max(\mu_a)$, and $\text{var}(\mu_a)$ (Figs. 5 to 7) show much larger differences between the mean values of affected and unaffected groups, and yield ANOVA p values $<5\%$. (All sample were Box-

Cox-transformed into normal distributions in order to perform ANOVA testing.) The parameter $\text{var}(\mu_a)$ is a measure of the variation in μ_a in the images. A healthy joint typically shows higher variation in this parameter, as the almost clear synovial fluid has a very small μ_a value compared to adjacent bones, cartilage, and other tissues. In a patient with RA, the μ_a value of the synovial fluid increases and overall, μ_a of all tissues becomes more simlar, hence $\text{var}(\mu_a)$ decreases. Differences in the ratio of $\min(\mu_a)/\max(\mu_a)$ between healthy and affected finger joints can be explained in a similar fashion. In a healthy joint, this ratio should be small, since $\min(\mu_a)$ is small in the synovial fluid. In a joint affected by RA $\min(\mu_a)$ is increased and closer
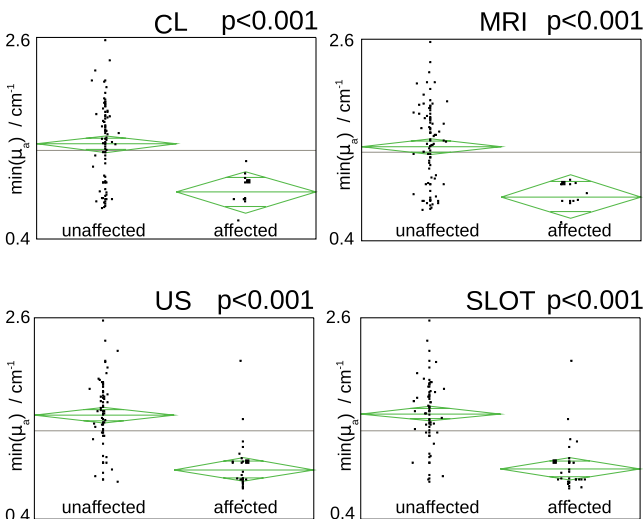

Fig. 5 Statistical distributions of the minimum absorption coefficient $\min(\mu_a)$ with respect to RA-affected and unaffected finger groups and ground truth derived from CL, MRI, US, and SLOT; p-values resulting from ANOVA that are less than 0.05 indicate statistical significant differences between both groups.
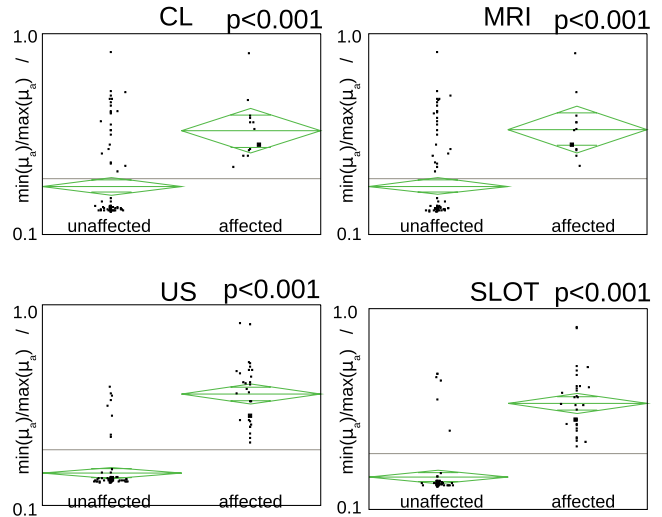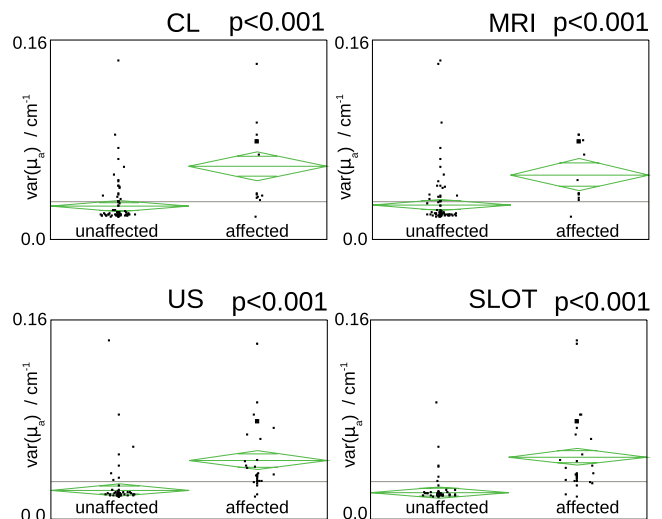

Fig. 7 Statistical distributions of the image variances of the absorption coefficient $\text{var}(\mu_a)$ with respect to RA-affected and unaffected finger groups and ground truth derived from CL, MRI, US, and SLOT; p-values resulting from ANOVA that are less than 0.05 indicate statistical significant differences between both groups.

Table 1 Results of the traditional discriminant analysis with respect to different parameter combinations and the MRI and US ground truth.

Data vector $x = \{\}$	MRI			US		
	S_e	S_p	J	S_e	S_e	J
$\max(\mu_a)$	0.12	0.88	0.00	0.75	0.34	0.09
$\min(\mu_a)$	0.21	0.92	0.13	0.80	0.47	0.27
$\min(\mu_a)/\max(\mu_a)$	0.17	0.93	0.10	0.81	0.40	0.20
$\mathrm{var}(\mu_a)$	0.33	0.93	0.26	0.81	0.70	0.51
$\max(\mu_a),\min(\mu_a)$	0.22	0.93	0.15	0.82	0.50	0.32
$\max(\mu_a),\min(\mu_a)/\max(\mu_a)$	0.21	0.92	0.13	0.80	0.47	0.27
$\max(\mu_a),\mathrm{var}(\mu_a)$	0.32	0.93	0.24	0.87	0.61	0.47
$\min(\mu_a),\min(\mu_a)/\max(\mu_a)$	0.19	0.92	0.12	0.85	0.49	0.34
$\min(\mu_a),\mathrm{var}(\mu_a)$	0.38	0.93	0.30	0.81	0.74	0.55
$\min(\mu_a)/\max(\mu_a),\mathrm{var}(\mu_a)$	0.32	0.93	0.24	0.85	0.62	0.47
$\min(\mu_a),\max(\mu_a),\min(\mu_a)/\max(\mu_a)$	0.21	0.92	0.13	0.83	0.50	0.33
$\min(\mu_a),\max(\mu_a),\mathrm{var}(\mu_a)$	0.32	0.93	0.24	0.85	0.59	0.44
$\min(\mu_a),\min(\mu_a)/\max(\mu_a),\mathrm{var}(\mu_a)$	0.32	0.93	0.24	0.87	0.60	0.47
$\max(\mu_a),\min(\mu_a)/\max(\mu_a),\mathrm{var}(\mu_a)$	0.42	0.97	0.39	0.87	0.64	0.51
$\min(\mu_a),\max(\mu_a),\min(\mu_a)/\max(\mu_a),\mathrm{var}(\mu_a)$	0.44	0.97	0.41	0.43	0.97	0.40

The table shows the sensitivity S_e, specificity S_p, and Youden index $J = S_e + S_p - 1$.

to $\max(\mu_a)$, hence the ration becomes larger. However, results of the DA, which are summarized in Table 1, show that good classification into groups of affected and unaffected fingers will be difficult using these single parameters.

Table 1 shows the classification results in terms of sensitivity, specificity and Youden index J for single and multi-parameters using MRI and US to produce the ground truth. Results for CE and SLOT are similar but were omitted here for clarity. We see that J-values for the single parameters are comparatively low, except for $\mathrm{var}(\mu_a)$. $\max(\mu_a)$ yields the lowest S_e, S_p, and J values. Using MRI to determine ground truth, the highest J value ($J = 0.41$) is achieved when all four parameters are combined. If US is used to determine the ground truth the highest J value ($J = 0.55$) is achieved with a combination of only two parameters, $\mathrm{var}(\mu_a)$ and $\min(\mu_a)$. Notable is also that in general S_p is very high (> 0.9) and S_e very low (< 0.44), when MRI is used to determine the ground truth. With US as ground truth, these roles seem to be reversed, therefore S_e is in general higher (≈ 0.8) than S_p (≈ 0.5 to ≈ 0.6).

The main hypothesis of this study is that a machine learning approach that makes use of SOM methods applied to multiparameter analysis will yield better classification with respect to RA than currently available methods. To demonstrate this the SOM-network was trained with 100 input n-dimensional data vectors, with respect to a cross-validation.

Figures 8 and 9 show the estimated classification and prediction performances, J and I, of the SOM method with respect to different sets of optical parameters for the four different ground truths (derived from CE, MRI, US, and SLOT). Displayed are the changes of J and I as a function of the frequency thresholds p_t for 11 different parameter configuration. The error bars in these figures represent the prediction uncertainties (standard deviations), which result from the varying cluster size and the cross-validation methods described in the previous section. To arrive at these particular error bars, the computer-aided algorithm varied the cluster size for each parameter combination. For example, when combining $\mathrm{var}(\mu_a)$ with $\min(\mu_a)/\max(\mu_a)$, Fig. 10 shows the classification performances J and I with respect to US as ground truth and varying cluster sizes. Initially, J and I improve as the number of clusters is increased. However, once the number of clusters reaches 25 (four feature vectors per cluster) J and I are almost constant, approximately equal to 0.94 and 0.91, respectively. Therefore, optimal interpretations of the classification results are performed with a SOM neural network architecture of ~ 25 subgroups (clusters).

Figures 8 and 9 show that for all ground truths we can find optical parameter combinations and frequency thresholds p_t for which the Youden index J is larger than 0.75. This is substantially higher than the highest value obtained with the DA approach ($J = 0.55$, see Table 1). In general, using the
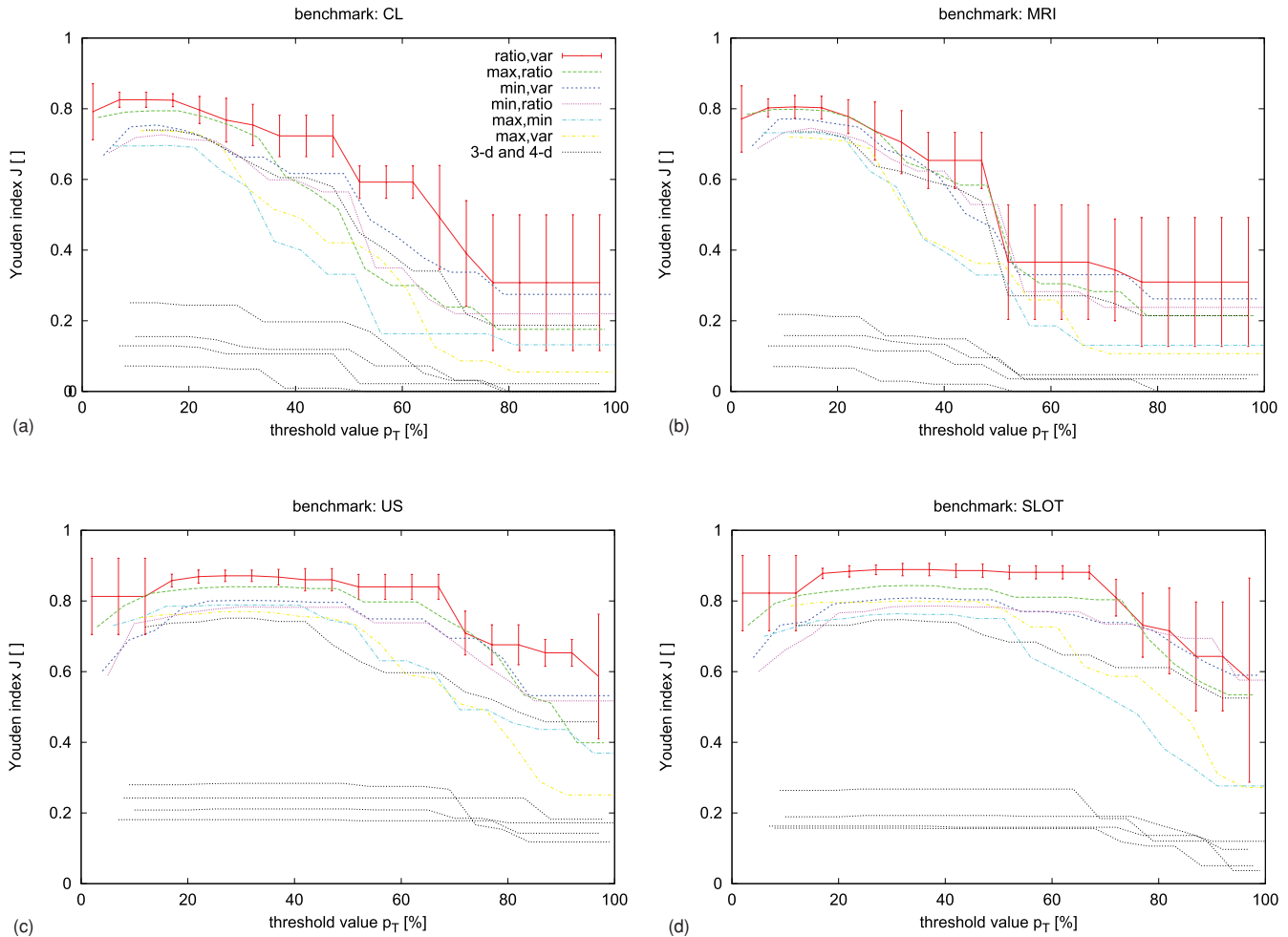
Fig. 8 Youden index J as function of the threshold value p_t. Results show different image feature combinations with respect to 4 different ground truth benchmarks (a) to (d). Feature combinations based on two features show higher J values than based on three and four features (see key). Error bars are given only for the most reliable features. They result from uncertainties due to different SOM neural network sizes and the cross validation.

machine-learning/SOM approach the highest J values are obtained when combing $\mathrm{var}(\mu_a)$ with the $\min(\mu_a)/\max(\mu_a)$ ratio. Here p_t is 13% and J is 0.81 when using MRI-derived ground truth. The corresponding values for sensitivity and specificity are $S_e = 0.96$ and $S_p = 0.85$ (see Table 2). With US-derived ground truth, and $p_t = 31\%$, these values increase to $J = 0.87$ ($S_e = 0.96$ and $Sp = 0.91$). Figure 11 shows the related ROC curves. Also shown in Fig. 11 are the ROC curves using SLOT and US as ground truth and the curve reported by Scheel et al.[11] Scheel's analysis, which yielded $J = 0.41$ ($S_e = 0.71$ and $S_p = 0.71$) relied on a single parameter [$\min(\mu_a)$] for classification.

Looking at Figures 8 and 9, we furthermore find that parameter combinations of only two features [e.g., $\min(\mu_a)/\max(\mu_a)$ and $\mathrm{var}(\mu_a)$ or $\max(\mu_a)$ and $\mathrm{var}(\mu_a)$] lead to higher accuracy measures (J and I) than three of four feature combinations (shown in gray). The reasons for that behavior are not entirely clear. However, using the DA approach with US as ground truth we found a similar result. Here a two-parameter combination [$\max(\mu_a)$ plus $\mathrm{var}(\mu_a)$] gave the best J value.

Furthermore, curves generated with US-derived ground truth look similar to curves generated with SLOT-derived ground

truth. In both cases, the largest J values are reached in the range of $20\% < p_t < 70\%$. The associated S_e and S_p values are all larger than 0.85 in this range. For values of $p_t > 70$ the Youden index falls off. This similarity, which we already observed when looking at the distribution of the single parameters (Figs. 6 to 8), suggest that US and SLOT are similar in the assessment of RA in finger joints.

5 Conclusion

Optical tomographic imaging is increasingly applied in clinical studies concerning the detection of various diseases such as breast cancer, arthritis, or brain hemorraghes. While substantial progress has been made with respect to imaging instrumentation and optical tomographic image reconstruction schemes, relatively little effort has been expended on image analysis schemes that extract useful features from tomographic images and help classifying a patient as free or afflicted by a certain disease.

This study presents the first attempt in the field of optical tomography to use advanced CAD methods. In particular we employ an unsupervised interpretation system based on SOMs
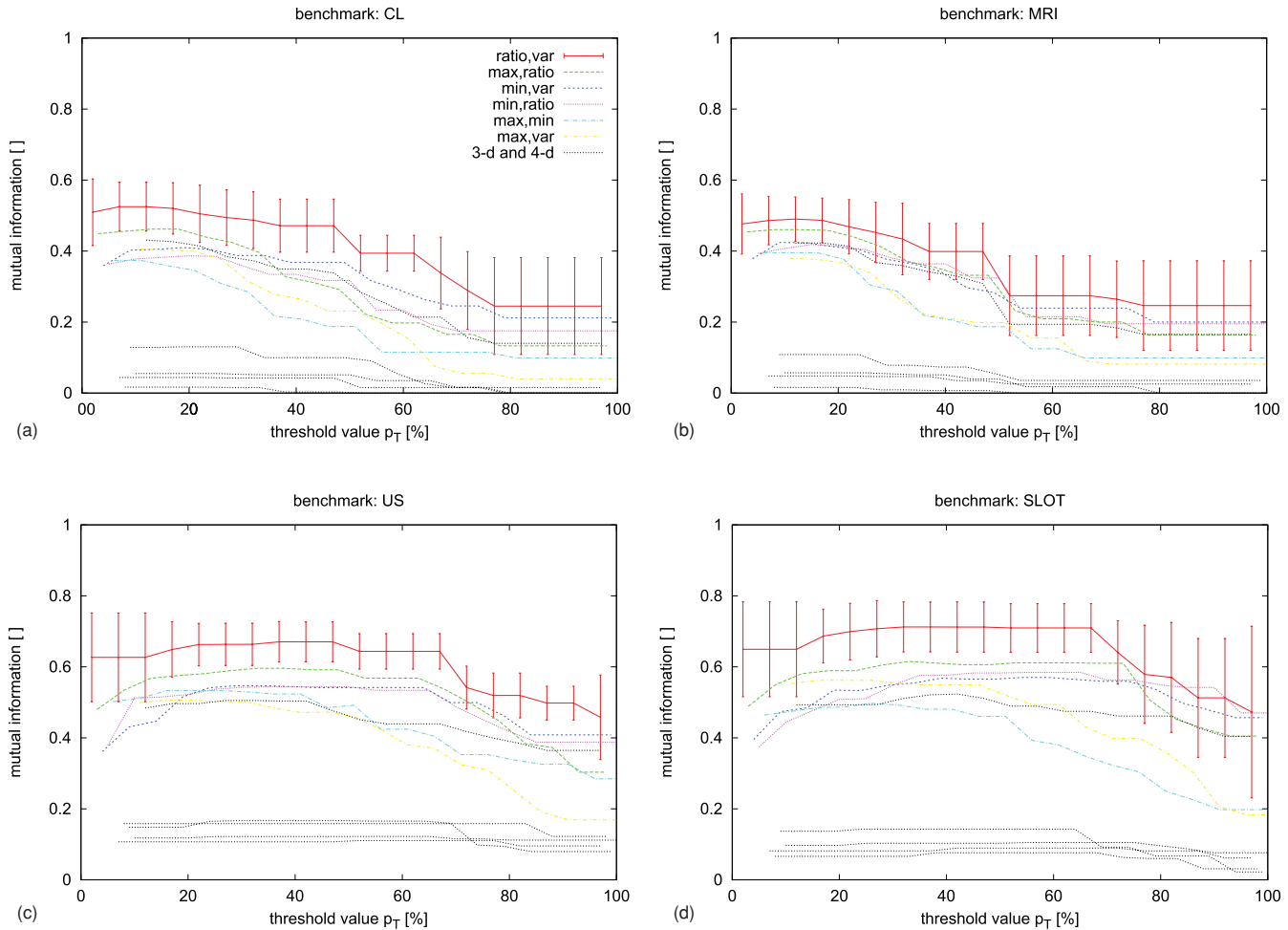
Fig. 9 Mutual information *I* as interpretation accuracy to identifying arthritis affected finger joints. Results show different image feature combinations with respect to ground truth benchmarks (a) to (d). Similar to Fig. 8, feature combinations based on two features show higher *I* values that base on three and four features (see key). Error bars are given only for the most reliable features. They result from uncertainties due to different SOM neural network sizes and the cross validation.
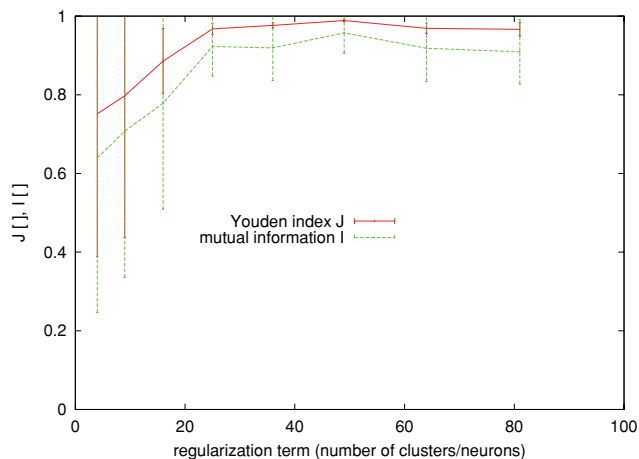

Fig. 10 Interpretation accuracies *J* and *I* for the most reliable image feature combination of $\min(\mu_a)/\max(\mu_a)$ ratio and variance(μ_a). The maximum performance is reached with >25 clusters. Results and error bars are based on the SOM neural network size (number of clusters or neurons) and on a 90 to 10% cross validation.

to distinguish between finger joints affected and not affected by RA. Different parameters (e.g., smallest and largest absorption and scattering coefficient and respective ratios) drawn from SLOT images became input to the CAD algorithm, and Youden index, specificity and sensitivity, and mutual information were used as classification performance measures. The performance measures were calculated for four different ground truth (generated by MRI, US, CE, and optical inspection of SLOT images) and compared to results of conventional statistical analysis methods, such as discriminant analysis.

Specificity and sensitivity of 0.85 and 0.96, respectively, could be achieved, when combining the ratio of the minimal and maximal absorption coefficient and the variance in an image, assuming MRI provides a good ground truth. If US is chosen to provide the ground truth, we get $S_p = 0.91$ and $S_e = 0.96$. These values are considerably higher than values obtained with single-parameter analysis reported earlier, or best case scenarios obtained with a discriminant analysis approach. The specificity and sensitivity levels that were reached with this proposed image classification approach make sagittal optical tomographic

Table 2 Results of the machine-learning-based classification of all two-parameter combinations that have shown best classification performances including sensifity S_e, specificity S_p, the resulting Youden index $J = S_e + S_p - 1$, the mutual information I, and the area under the curve (AUC).

Ground Truth	Data Vector $x = \{\ \}$	S_e	S_p	J	$p_t(J)$ (%)	I (%)	$p_t(I)$	AUC
CE	$\max(\mu_a)$, $\min(\mu_a)$	0.99	0.71	0.70	14	0.37	8	0.44
MRI		1.00	0.73	0.73	9	0.40	9	0.44
US		0.97	0.82	0.79	27	0.53	24	0.75
SLOT		0.95	0.81	0.76	34	0.5	23	0.73
CE	$\max(\mu_a)$, $\min(\mu_a)/\max(\mu_a)$	0.97	0.82	0.79	13	0.46	13	0.54
MRI		0.99	0.81	0.80	13	0.46	11	0.55
US		0.95	0.89	0.84	36	0.60	34	0.81
SLOT		0.92	0.92	0.84	33	0.62	32	0.84
CE	$\max(\mu_a)$, $\mathrm{var}(\mu_a)$	0.99	0.75	0.74	12	0.41	13	0.46
MRI		0.99	0.73	0.72	12	0.38	9	0.44
US		0.97	0.80	0.77	29	0.51	17	0.72
SLOT		0.96	0.84	0.80	29	0.56	21	0.77
CE	$\min(\mu_a)$, $\min(\mu_a)/\max(\mu_a)$	0.91	0.82	0.73	14	0.39	20	0.53
MRI		0.91	0.83	0.74	13	0.42	13	0.53
US		0.91	0.87	0.78	37	0.54	37	0.78
SLOT		0.82	0.97	0.79	35	0.58	48	0.82
CE	$\min(\mu_a)$, $\mathrm{var}(\mu_a)$	0.96	0.79	0.75	12	0.41	15	0.58
MRI		0.96	0.81	0.77	12	0.42	12	0.55
US		0.89	0.91	0.80	36	0.55	30	0.80
SLOT		0.89	0.92	0.81	32	0.57	46	0.82
CE	$\min(\mu_a)/\max(\mu_a)$, $\mathrm{var}(\mu_a)$	0.98	0.85	0.83	15	0.53	10	0.65
MRI		0.96	0.85	0.81	13	0.49	13	0.60
US		0.96	0.91	0.87	31	0.67	42	0.86
SLOT		0.94	0.95	0.89	38	0.71	46	0.88

J and I are also characterized by average frequency thresholds $p_t(J)$ and $p_t(I)$, which represent the predicted amount of all RA affected finger joints.

imaging an attractive tool for the evaluation of arthritis in finger joints. Larger clinical trials are now under way to further explore the clinical usefulness of this medical imaging procedure.

6 Appendix: Self Organizing Mapping

As outlined in the Sec. 3.1, a machine learning algorithm based on SOM, was employed as part of an automated unsupervised interpretation system. In particular, we use the SOM method to cluster data derived from the optical tomographic images. For the given case, each image is represented by a feature vector

x whose components are given either by $\min(\mu_a)$, $\max(\mu_a)$, $\min(\mu_a)/\max(\mu_a)$, $\mathrm{var}(\mu_a)$, or a subset of these four parameters. Therefore, depending on what combination of features is considered for the clustering, x either has two, three, or four dimensions. Given that in this studies 100 finger images were available, 100 feature vectors were derived that were separated into l clusters, where l took on values between 2 to 80.

To understand how the clustering was performed, one needs to understand the basic structure of a SOM network. A SOM is structured in two layers: an input layer and a Kohonen layer [Fig. 12(a)]. The input layer is a one-on-one representation of
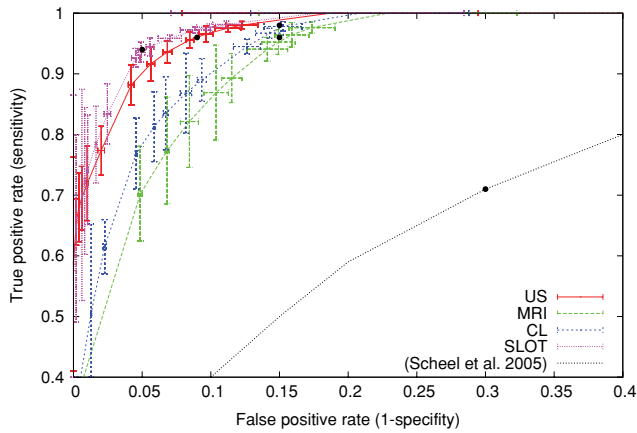
Fig. 11 Sensitivity-specificity curves (ROC curves) illustrating image interpretation results based on the combination of $\min(\mu_a)/\max(\mu_a)$ ratio and variance(μ_a). The curves show the best classification performances using J (black dots). The error bars of the curves result from cross validations. The results are compared to all ground truth benchmarks the best single-parameter classifications reported by Scheel et al.[11]

each given feature vector. Therefore, the number of neurons in the input layer equals the dimensions of the feature vector. If all four features [$\min(\mu_a)$, $\max(\mu_a)$, $\min(\mu_a)/\max(\mu_a)$, and $\text{var}(\mu_a)$] are considered at the same time, the input layers has four neurons, one for each feature [Fig. 12(a)]. The Kohonen layer represents a structure with a single 2-D map (lattice) consisting of neurons arranged in rows and columns. For the given example, each neuron in the Kohonen layer represents one cluster. Therefore, if 100 feature vectors need to be distributed into, e.g., 25 clusters, the Kohonen layer will have 25 or 5×5 neurons. Each neuron of the Kohonen layer is fixed and is fully linked with all neurons of the input layer. The links are described by weights w_k, given by

$$w_k = \{w_{(k,\text{feature}_1)}, w_{(k,\text{feature}_2)}, \ldots, w_{(k,\text{feature}_n)}\}^T. \quad (4)$$

Here n is the dimension of the feature vector x, with $n = 2, 3$, or 4; k is the index to a specific neuron, representing a specific cluster in the Kohonen layer; with $k = 1, 2, \ldots, m, \ldots, l$, where

l is number of all Kohonen neurons. The weights can take on values between 0 and 1. If n features are considered and l clusters are desired, the number of weights is $n \times l$.

Clustering is now achieved in the following way:

1. First the weights w_k are initialized, i.e., by assigning random values.

2. All feature vectors x_i are presented to the network and the Euclidean distance $\|x_i - w_k\|$ between each x_i and each w_k is calculated. Note, all original feature vectors x_i^o drawn from an image must be scaled by the standard mean x_{SM} and normalized by the standard deviation x_{SD} to make all features within a feature vector comparable: $x_i = \frac{x_i^o - x_{SM}}{x_{SD}}$.

3. The index j of the Kohonen neuron, whose weight w_k is the closest to vector x_i is determined by

$$j(x_i) = \arg\min_k \|x_i - w_k\| \quad | \quad k = 1, 2, \ldots, m, \ldots, l. \quad (5)$$

4. Given theses "winner" neurons, new weights are calculated for the entire network according to

$$w_k(t+1) = w_k(t) + \eta(t) \ h_{k,j(x)}(t) \ [x_i(t) - w_k(t)], \quad (6)$$

where $\eta(t)$ is the learning-rate parameter during the calculation step t, and $h_{k,j(x)}(t)$ is the neighborhood function centered around the winning neuron $j(x_i)$. The neighborhood function is given by

$$h_{k,j(x)} = \exp\left(-\frac{d_{k,j(x)}^2}{2\sigma^2}\right). \quad (7)$$

This Gaussian function depends on the lateral neuron distance d and the effective width, which is a variable of the network that determines how many neighboring neurons become modified.

5. After the first update of the weights, the next learning cycle starts by again presenting all feature vectors to the SOM network and repeating steps 2 through 4 etc. The learning rate is reduced in each cycle according to $\eta(t+1) = (1 - t/t_F)\eta(t)$.
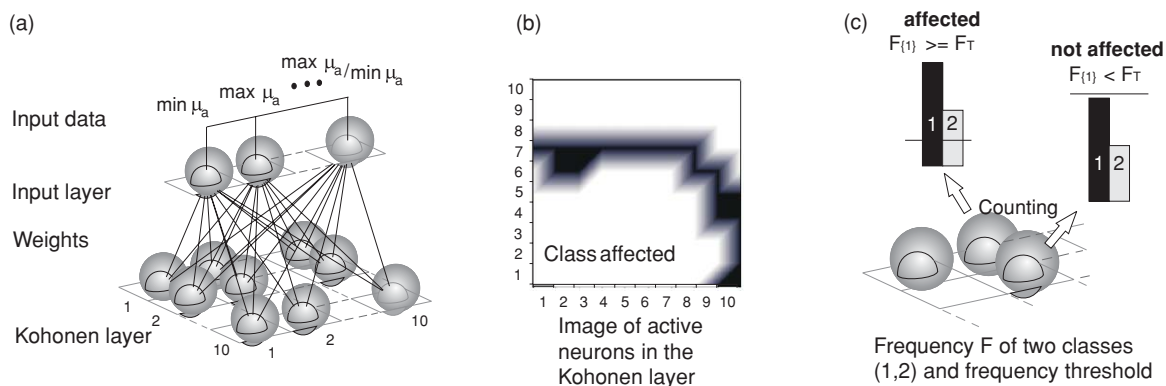

Fig. 12 Scheme for multiparameter classifications based on SOM: (a) structure of a SOM neural network, (b) image of active neurons representing the class "affected with rheumatic arthritis" within the Kohonen layer after discrimination of the given input data and (c) frequency determination and final classification of the classes "affected" (black) and "not affected" (gray) with respect to a probability threshold p_t.

Table 3 SOM algorithm parameters.

Represented feature space	$\mathbb{R}^n$	$\{\min(\mu_a), \max(\mu_a), \ldots,$ $\min(\mu_a)/\max(\mu_a), var(\mu_a)\}^T$
Number of input neurons	2,3 or 4	
Number of feature/ input vectors	100	data drawn from images
Number of Kohonen neurons	2,…,80	(e.g., 4×4 lattice)
Number of weights	up to 4×80	
Number of iterations t_F	10,000	
Initial learning rate $\eta(t=0)$	0.5	
Final learning rate $\eta(t_F)$	0.01	
Initial neighbourhood size $\sigma(t=0)$	up to 10	depends on the structure
Final neighbourhood size $\sigma(t_F)$	1	decreasing every 2000 iterations

This process is repeated until all the weights converge to stable values, meaning $\Delta w_k(t+1) = \|w_k(t+1) - w_k(t)\|$ is smaller than a preset value, or a preset number of learning cycles (iterations) t_F have been completed.

After training, a feature vector presented to the trained network, will excite exactly one neuron that represents one cluster. Hence each feature vector "belongs" to exactly one cluster. For example, if feature vectors are chosen with two components [e.g. $\min(\mu_a)/\max(\mu_a)$ and $var(\mu_a)$] and 13 neurons populate the Kohonen layer, each of the 100 feature vectors would excite one of these 13 neurons. The 100 data points would have been divided into 13 clusters as shown in Fig. 2(b), the example discussed in the main text. Similar Kohonen layer neurons correspond to similar feature vectors of the given input space. The structure and network parameters for the SOM algorithm used in this work are as shown in Table 3. Further details about SOM and the general learning process on tomographic image data can be found in Refs. 21 and 35.

References

1. B. J. Tromberg, B. W. Pogue, K. D. Paulsen, A. G. Yodh, D. A. Boas, and A. E. Cerussi, "Assessing the future of diffuse optical imaging technologies for breast cancer management," *Med. Phys.* **35**(6) 2443–2451 (2008).
2. A. Karellas and S. Vedantham, "Breast cancer imaging: a perspective for the next decade," *Med. Phys.* **35**(11), 4878–4897 (2008).
3. Q. Q. Fang, S. A. Carp, J. Selb, G. Boverman, Q. Zhang, D. B. Kopans, R. H. Moore, E. L. Miller, D. H. Brooks, and D. A. Boas, "Combined optical imaging and mammography of the healthy breast: optical contrast derived from breast structure and compression," *IEEE Trans. Med. Imaging* **28**(1), 30–42 (2009).
4. A. V. Meduedev, J. Kainerstorfer, S. V. Borisov, R. L. Barbour, and J. VanMeter, "Event-related fast optical signal in a rapid object recognition task: improving detection by the independent component analysis," *Brain Res.* **1236**, 145–158 (2008).
5. T. J. Huppert, S. G. Diamond, and D. A. Boas, "Direct estimation of evoked hemoglobin changes by multimodality fusion imaging," *J. Biomed. Opt.* **13**(5), 195–201 (2008).
6. A. K. Dunn, T. Bolay, M. A. Moskowitz, and D. A. Boas, "Dynamic imaging of cerebral blood flow using laser speckle," *J. Cereb. Blood Flow Metabol.* **21**(3), 195–201 (2001).
7. A. D. Klose, V. Prapavat, O. Minet, J. Beuthan, and J. G. Mueller, "RA-diagnostics applying optical tomography in frequency domain," *Optical and Imaging Techniques for Biomonitoring*, H. J. Foth, R. Marchesini, and H. Podbielska, Eds. *Proc. SPIE* **3196**, 194–204 (1998).
8. A. D. Klose, A. H. Hielscher, K. M. Hanson, J. Beuthan "Two and three-dimensional optical tomography of a finger joint model for diagnostic of rheumatoid arthritis," *Proc. SPIE* **3566**, 15160 (1998).
9. A. D. Klose, "Optical tomography based on the equation of radiative transfer," PhD Thesis, Freie Universität Berlin (2002); http://www.diss.fu-berlin.de/2002/135/indexe.html.
10. A. H. Hielscher, A. D. Klose, A. Scheel, B. Moa-Anderson, M. Backhaus, U. Netz, and J. Beuthan, "Sagittal laser optical tomography for imaging of rheumatoid finger joints," *Phy. Med. and Biol.* **49**(7), 1147–1163 (2004).
11. A. K. Scheel, M. Backhaus, A. D. Klose, B. Moa-Anderson, U. J. Netz, K.-G. A. Hermann, J. Beuthan, G. A. Mller, G. R. Burmester, and A. H. Hielscher, "First clinical evaluation of sagittal laser optical tomography for detection of synovitis in arthritic finger joints," *Ann. Rheum. Dis.* **64**, 239–245 (2005).
12. Q. Z. Zhang and H. B. Jiang, "Three-dimensional diffuse optical imaging of hand joints: System description and phantom studies," *Opt. Lasers eng.* **43**(11), 1237–1251 (2005).
13. J. M. Lasker, C. J. Fong, D.T. Ginat, E. Dwyer, and A. H. Hielscher, "Dynamic optical imaging of vascular and metabolic reactivity in rheumatoid joints," *J. Biomed. Opt* **12**(5), 052001 (2007).
14. L. A. Meinel, A. H. Stolpen, K. S. Berbaum, L. L. Fajardo, J. M. Reinhardt, "Breast MRI lesion classification: improved performance of human readers with a backpropagation neural network computer-aided diagnosis (CAD) system," *J. Magn. Reson. Imaging* **25**(1), 89–95 (2007).
15. K. Awai, K. Murao, A. Ozawa, M. Komi, H. Hayakawa, S. Hori, and Y. Nishimura, "Pulmonary nodules at chest CT: effect of computer-aided diagnosis on radiologists," *Detect. Perform.* **230**, 347–352 (2004).
16. C. M. Chen, Y. H. Chou, K. C. Han, G. S. Hung, C. M. Tiu, H. J. Chiou, and S. Y. Chiou, "Breast lesions on sonograms: computer-aided diagnosis with nearly setting-independent features and artificial neural networks," *Radiology* **226**, 504–514 (2003).
17. P. G. Spetsieris, Y. Ma, V. Dhawan, J. R. Moeller, and D. Eidelberg, "Highly-automated computer-aided diagnosis of neurological disorders using functional brain imaging," *Proc. SPIE* **6144**, 61445M (2006).
18. K. Doi, "Computer-aided diagnosis in medical imaging: historical review, current status and future potential," *Comput. Med. Imaging and Graph.* **31**(4–5), 198–211 (2007).
19. X. Qi, M. Sivak, G. Insberg, J. E. Willis, and A. M. Rollins, "Computer-aided diagnostics of dysplasia in Barrett's esophagus using endoscopic optical tomography," *J. Biomed. Opt.* **11**(4), 044010 (2006).
20. F. Bazant-Hegemark, N. Stone, M. D. Read, K. McCarthy, and R. K. Wang, "Optical coherence tomography (OCT) imaging and computer aided diagnosis of human cervical tissue specimens," *Proc. SPIE* **6627**, 66270F (2007).
21. C. D. Klose, Self-organising maps for geoscientific data analysis: geological interpretation of multi-dimensional geophysical data," *Computat. Geosci.* **10**(3), 265–277 (2006).

22. T. Kohonen, "Self-organizing formation of topologically correct feature maps," *Biol. Cyb.* **43**(1), 59–69 (1982).
23. T. Kohonen, *Self-Organizing Maps*, 3rd ed. Springer, Berlin (2001).
24. T. W. Nattkemper and A. Wismüller, "Tumor feature visualization with unsupervised learning," *Med. Image Anal.* **9**, 344351 (2005).
25. A. Pascual-Montano, K. H. Taylor, H. Winkler, R. D. Pascual-Marqui, and J.-M. Carazo, "Quantitative self-organizing maps for clustering electron tomograms," *J. Struct. Biol.* **138**, 114122 (2002).
26. R. Schönweiler, P. Wübbelt, R. Tolloczko, C. Rose, and M. Ptok, "Classification of passive auditory event-related potentials using discriminant analysis and self-organizing feature Maps," *Audiol. Neurotol.* **5**, 69–82 (2000).
27. R. W. Veltri, M. Chaudhari, M. C. Miller, E. C. Poole, G. J. O'Dowd, and A. W. Partin, "Comparison of logistic regression and neural net modeling for prediction of prostate cancer pathologic stage," *Clin. Chem.* **48**(10), 1828–1834 (2002).
28. A. D. Klose, U. Netz, J. Beuthan, and A. H. Hielscher, "Optical tomography using the time-independent equation of radiative transfer. Part I: Forward model," *J. Quant. Spectrosc. Radiat. Transf.* **72**(5), 691–713 (2002).
29. A. D. Klose and A. H. Hielscher, "Optical tomography using the time-independent equation of radiative transfer. Part II: Inverse model," *J. Quant. Spectrosc. Radiat. Transf.* **72**(5), 715–732 (2002).
30. S. Haykin, *Neural Networks—A Comprehensive Foundation*, 2nd ed., Englewood Cliffs, NJ, Prentice Hall (1999).
31. B. G. Tabachnick and L. S. Fidell, *Using Multivariate Statistics*, Boston, Pearson Education Inc. (2007).
32. T. D. Wickens, *Elementary Signal Detection Theory*, New York: Oxford University Press (2002).
33. W. J. Youden, "Index rating for diagnostic tests," *Cancer* **3**, 32–35 (1950).
34. D. J. C. McKay, *Information Theory, Inference, and Learning Algorithms*, 7th ed. (Cambridge University Press, New York, 2004).
35. C. D. Klose, A. D. Klose, J. Beuthan, and A. Hielscher, "Multi-parameter classifications of optical tomographic images," *J. Biomed. Opt.* **13**(5), 050503 (2008).