INTL JOURNAL OF ELECTRONICS AND TELECOMMUNICATIONS, 2024, VOL. 70, NO. 2, PP. 271–276
Manuscript received January 30, 2024; revised June, 2024. doi: 10.24425/ijet.2024.149541

Detection of human finger joints in ultrasound images: structure and optimization

Artur Bąk, Kamil Wereszczyński, Jakub Segen, Paweł Mielnik, Marcin Fojcik, and Marek Kulbacki

***Abstract*—Synovitis is the inflammation of a synovial membrane surrounding a joint. Its assessment is an important step in the diagnosis and treatment of rheumatoid arthritis. Joint detection is the first stage of an automated method of assessment of a degree of synovitis, from an Ultrasound (USG) image of a finger joint and its surrounding area. A joint detector consists of three parts: image preprocessing, feature extraction, and classification. Each part contains adjustable parameters that must be set experimentally to ensure the proper operation of the detector. Both the structure of a joint detector and a procedure for finding a near-optimal configuration of the adjustable parameters are described. The optimization process is based on two evaluation measures: Area Under the Receiver Operating Characteristic Curve (AUC) and False Positive Count (FPC). The optimization process decreases the number of pictures with multiple detections, which was the main point of works presented in this paper. This was achieved by increasing the number of components of the homogeneous mixed-SURF descriptor which has the greatest influence on the final result. Non-SURF descriptors achieve poorer classification results. Our research led to the creation of a better joint detector which could positively influence the final results of inflammation level classification.**

***Keywords*—SVM; USG; SURF; synovitis; feature extraction; classification**

I. Introduction

BOTH diagnosis and treatment of rheumatoid arthritis rely in part on determining the degree of synovitis, which is an inflammation of a synovial membrane that surrounds a joint. The assessment of synovitis is commonly done by a specialist examining ultrasound (USG) images of finger joints and surrounding areas [1]. An objective of the research conducted by the authors is the construction of a synovitis estimator for automated assessment of a degree of synovitis. The application of such an estimator is an automatic assist is a form of primary assistance in rheumatoid arthritis diagnosis by doctors. The main aim of the whole system is to speed up the work of radiologists preserving the level of its reliability. The presented work aims to optimize one of the modules of such a system: a joint detector.

The main system consists of several modules: low-level biological structure recognition (joint, skin, and bones that are depicted in Fig. 1), synovitis extraction, and marking of the inflammation level. This article presents the optimization process of a joint detector which is one of the low-level detectors and was described in [2]. As was mentioned in the cited paper, the original version of the joint detector returns more than one point recognized as a joint center in the case of ca. 30% of images. The Synovitis extraction process demands one joint center as the input. Therefore further processing for images that have more than one joint detection cannot proceed. To overcome this problem, the optimization of joint detection is necessary and should lead to a decrease in the number of multi-detections of the joint center.

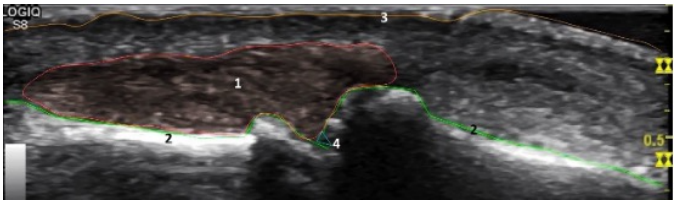


Fig. 1. An example USG image with marked biological structures: 1-synovitis, 2-bones, 3-skin, 4-joint area

In a previous work [2] we proposed different methods of efficient description of pixels used in the automatic search of joints in images. We have shown that using the same methods, but with different parameters, improves the result. In the current work, we present the results of the examination of this fact in more detail. We have proven that adding another version of the same method with different parameters reduces the number of multiple detections. We also present the results of an examination of some other methods, but they turned out to be less effective than those mentioned above. The investigations of the non-homogeneous concatenations of different methods also proved that their results were not satisfactory. Based on previous work, we also assumed that the last part of the detector (classification method) should not be changed. It has been shown, that the selected classification

Artur Bąk, Jakub Segen and Marek Kulbacki is with Polish-Japanese Academy of Information Technology, Warsaw, Poland (e-mail: {abak, js, kulbacki}@pjwstk.edu.pl).

Kamil Wereszczyński is with Silesian University of Technology, Institute of Informatics, Gliwice, Poland (e-mail: kamil.wereszczynski@polsl.pl).

Paweł Mielnik is with Department for Neurology, Rheumatology and Physical Medicine, Helse Førde, Førde, Norway (e-mail: mielnik.p@gmail.com).

Marcin Fojcik is with Western Norway University of Applied Sciences, Norway (e-mail: marcin.fojcik@hisf.no).

Marek Kulbacki is with DIVE IN AI, Wroclaw, Poland.

Regional Committee for Medical and Health Research Ethics, Region West, Norway has approved the study (ref. 2013/743). All participants signed an informed consent form.

method combined with proposed descriptors obtains the best effectiveness among many other tested classifiers.

To summarize, this paper presents the process of joint detector optimization. The main assumption is that the optimization process will lead to a reduction in the number of images with multiple detections obtained in previous work. First, we describe several methods and their combinations with partial results, and then we present the final result of the selected combination of methods.

The results will be shown in reference to the annotation made by domain experts using commonly known measures, which are described in section II. We assume at this point that if the automatically indicated joint center is not far from the center of the joint indicated by the domain expert, both the precision of joint detection and selection process are sufficient. The meaning of the ”far” concept is also introduced in section II.

II. Materials and methods

The joint detector consists of three parts: (a) image preprocessing, (b) feature extraction, and (c) pixel classification. Each of these parts can be configured with adjustable parameters. For example, one of such parameters is the type of feature extraction. There are many types of pixel descriptors like SURF [3], [4], and others. Another example of a parameter is the size of the window used for Gaussian filtering. The values of the parameters could significantly affect the results of the detector. The optimal value (e.g. best pixel descriptor or best Gaussian filter value) should be determined in the optimization process. The final version of the joint detector uses these selected parameter values as input information. We could not expect that the optimum values for each parameter will be achieved, because of their large amount. Therefore we expect sufficiently optimal parameter values. The decision if a given set of parameter values is sufficiently optimal is made according to achieved results. We are not able to check all possible values of all parameters. Therefore we introduce some measures for partial results. If we find that the partial results of one of the parts of the joint detector are reasonable, we will use these results for the optimization of the next part of the joint detector. In the current section, each of these three parts and measures of partial results evaluation are shortly described.

Finally, we take some promising sets of parameter values using partial results and compare the final joint detections with joint centers indicated by domain experts. This operation is made on the images that were not used in the optimization process. We assume that in all cases there will be sufficient detection, and the meaning of the ”sufficient” term will be described in subsection II-A. All sets of parameters for which the detector will not give a sufficient response will be assumed as not correct. On the other hand, we attempt to reduce the number of images that have more than one response. The measures used for the evaluation of this reduction are described in subsection II-F.

Regional Committee for Medical and Health Research Ethics, Region West, Norway has approved the study (ref. 2013/743). All participants signed an informed consent form.

A. Data sets

The training data set is a collection of annotated USG images of finger joints. The image annotations contain information added manually to an image by human experts - the medicine doctor who annotates the ultrasound images professionally and the persons who were trained by him to recognize the joint, bones, skin, and the area of synovitis in an ultrasound image. The bones and skin are marked in the image by lines, while a joint region and a synovitis area are denoted by outlines (closed curves), as shown in Fig. 1. The information contained in the annotations is used to train and optimize detectors and classifiers, as well as to test and evaluate the results.

A polygon marks the area of the joint in Fig. 1. The mean of its apexes is assumed as the indicated joint center. The mean of the distances between the indicated center of the joint and each of the apexes is called the indicated joint radius. A detection is considered sufficient if it lies within a circle centered at the indicated center of the joint. The radius of this circle is the multiplication of the indicated joint radius and ε, which is the parameter of the evaluation process.

For creating the joint detector a set of 190 ultrasound images was used. The images come from different sessions. The session is the set of USG pictures made in one examination. We assume that different sessions may have been performed on different patients. One session may contain a sequence of almost the same pictures. For our set of photos, we took only one photo from such a session. We divided this set of 190 images into two subsets: training (64 pictures) and evaluation (126 pictures). In the process of training classifiers, the cross-validation procedure involves further divisions of the training set. All results presented in this paper were obtained from the evaluation set and the pictures in this set were not used in the training process.

B. Image preprocessing

The first part of the joint detector is a series of image-processing operations. Several different operations were examined in the final classification context. The best final detection result was obtained for grayscaling and Gaussian smoothing. Gray-scaling is the process of changing the values of the RGB color image into the corresponding values of one channel color image. In the resulting image each pixel obtains a new value from the range [0,255] connected with the original image with function: G = 0.299R+0.587G+0.114B, where R is the value of the red channel, G - green channel, and B - blue one. The ultrasound images have the grayscale itself, but there are some additional color layers. Therefore the format of the file does not maintain this property. Gaussian smoothing is a mask filter with a Gaussian Kernel.

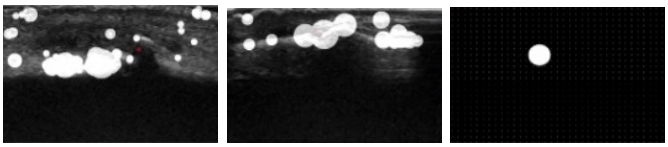
The decision process of joint detection determines if a pixel belongs to the joint area or not. This has to be processed for each pixel of the given USG image and makes the whole process long. That is why the count of pixels used in further analysis should be reduced. This operation will be called pixels preselection. For this purpose, several methods were examined (see Fig. 2). The first approach finds the pixels that are distinct

within their neighborhoods, for example, the positions where the determinant of the Hessian matrix has a local maximum. This method uses known feature detectors like SURF [3], BRISK [5], FAST [4], ORB [6]. The preselected pixels are located in the picture area where the detected density of points is high.

The other method that we examined is a simpler method that generates denser selections. In this method, every n-th pixel for every n-th row of the image is included in the preselected pixel set. As a result, n×n mesh of pixels is obtained. In the training phase, this mesh is extended by adding the pixels taken from the joint area. In the detection phase, only the mesh of pixels is used.

C. Feature extraction

The result of the image preprocessing phase is a modified image and a set of preselected pixels, which is used for feature extraction. In the feature extraction phase, feature vectors are computed for the preselected pixels by applying a chosen descriptor function to each pixel's neighborhood. The following descriptors were examined: SURF [3], ORB [6], BRISK [5] and FAST [4].



(a) (b) (c)

Fig. 2. Preselection of pixels using methods with local neighborhood detectors: (a) - false preselection using FAST detector: joint center (red circle) is outside of preselected area, (b) correct preselection using FAST detector, (c) correct, dense-like preselection

In a previous work [2] we proved, that using two concatenated SURF descriptors with different window sizes gives better classification results than using only one. In the current work, we investigated more complex versions of descriptor concatenation. Feature vectors created by concatenating descriptors of different types (non-homogeneous) and the same type descriptors with different parameters (homogeneous) were evaluated in the optimization of the feature extraction part. Several homogeneous and non-homogeneous mixtures of the descriptors mentioned above were examined. Consistent improvement was achieved with a homogeneous mixture of SURF descriptors.

D. Classification

Several well-known classifier types were examined as candidates for the classification part of the joint detector: Support Vector Machine (SVM) [7], Decision Trees (DT) in its specific version Classification And Regression Tree (CART) [8] and Nearest Neighbor (NN) [9].

A CART version of the Decision Trees classifier was used with the following parameter selection: surrogate splits, 10-fold built-in cross-validation, and pruned branches that were physically removed from the tree.

The Nearest Neighbor classifier was formed on a set of pixel descriptor clusters built from positive training examples obtained from the descriptors of pixels inside the joint area collected from all images in the training set. In the classification phase, if a pixel descriptor is close enough to one of the cluster centers, it is classified as "in joint"; otherwise it is classified as "outside joint".

E. Experiments

Each joint detector is composed of three components: image preprocessing, feature extraction, and classifier. Each component configuration is a candidate for the final version of the joint detector. Each joint detector candidate with specific values of parameters is described by one scenario in a scripting language called MEDUSA script that was developed for the synovitis estimator project [10]. The set of scenarios for one candidate is called a scenario template. The scenarios within a template differ only by their parameter values.

Within our experimental environment, a very large number of such scenarios can be generated (using scenario templates), executed, and evaluated using two measures: AUC (Area Under the Receiver Operating Characteristics Curve) [11] and FPC (False Positive Count), which are described in the next subsection. The best scenario found using this search process is used to configure the final joint detector and its parameters.

F. Evaluation process

As was mentioned in Section II-E, two measures were used: AUC and FPC. AUC is suitable for assessing the quality of a classifier. It was observed that the preprocessing phase has little impact on the AUC value. Therefore, in the first step of the evaluation process, the AUC value was computed for each scenario template.

As a result, the best classifier with a preprocessing component was chosen. While the feature extractor has a very small impact on AUC value, it has a significant influence on FPC. Therefore, a second step was introduced, which is to examine the influence of the feature extractor on FPC. In this step, the image preprocessing and classifier component settings obtained in the previous step remain unchanged. The final result of the evaluation process was the best joint detector components with the best parameter values.

1) Area Under the Receiver Operating Characteristics Curve (AUC) measure: AUC methods rely on a measure R which is a proportion of true positive rate (TPR) to false positive rate (FPR) :

$$TPR = \frac{t_p}{t_p + f_n}, FPR = \frac{f_p}{f_p + t_n}, R = \frac{TPR}{FPR} \quad (1)$$

For each image from the test set, a trained joint detector processes the image pixels assigning to each pixel a label: "in joint" or "outside joint" - this step is called detection. Each detection result is qualified as true or false by comparing it with the label computed from the annotations of the image by an expert. Each symbol used in Equation 1 represents the number of detection results belonging to a given category,

TABLE I
DESCRIPTION OF THE SYMBOLS IN EQUATION 1

		Detector	
		in joint	**outside joint**
Expert	**in joint**	true positive t_p	false negative f_n
	outside joint	false positive f_p	true negative t_n

TABLE II
VARIANTS OF THE JOINT DETECTOR (GRAY - GRAYSCALE, HIST.EQ - HISTOGRAM EQUALIZATION, BLUR - GAUSSIAN BLUR, SVM - SUPPORT VECTOR MACHINE, NN - NEAREST NEIGHBOR CLASSIFIER, DT - DECISION TREE CLASSIFIER, SURF - SPEEDED UP ROBUST FEATURES, AUC - AREA UNDER THE RECEIVER OPERATING CHARACTERISTICS CURVE)

Name	Preprocess. Comp.	Feature Extractor	Classifier	AUC
SVM1	gray,hist.eq,blur	2xSURF	SVM	0.966
SVM2+	gray	2xSURF	SVM	0.981
SVM3	gray,hist.eq.,blur	1xSURF	SVM	0.961
SVM4	gray	1xSURF	SVM	0.975
SVM5	gray,blur	2xSURF	SVM	0.978
SVM6+	gray,blur	1xSURF	SVM	0.981
NN1	gray,hist.eq.,blur	2xSURF	NN	0.911
NN2	gray	2xSURF	NN	0.928
NN3	gray,hist.eq.,blur	1xSURF	NN	0.912
NN4	gray	1xSURF	NN	0.925
DT1	gray,hist.eq.,blur	2xSURF	DT	0.710
DT2	gray	2xSURF	DT	0.749
DT3	gray,hist.eq.,blur	1xSURF	DT	0.889
DT4	gray	1xSURF	DT	0.887

where possible categories are: true positive, false positive, true negative, and false negative values. The meanings of these categories are explained in Table I.

Executing the classifier on the test set with a specified setting of the classifier parameters will produce an R value. The set of the R values for various parameter settings will create a curve called the Receiver Operating Curve (ROC). The area under the ROC is a quality measure for the classifier. One AUC value is obtained for each joint detector. The joint detector with the highest AUC value is considered the best.

G. False Positive Count (FPC)

In the case where two joint detector candidates have close AUC values and produce at least one true positive result for each test image, we further compare them using the False Positive Count (FPC), which is the number of separate clusters of false positive pixels per image. For each test image, we compute the FPC and collect the FPC values from the test set in the form of an FPC histogram. The percentages of images that have 0 FPC and those that have maximum FPC in the test set are secondary measures of detector quality. The better detector is the one with a higher 0 FPC. If both detectors have the same 0 FPC value, the one with the lower maximum FPC is better.

H. Methodology

This section details the steps and assumptions that lead to determining the target detector configuration.

1) A set of candidates for the best joint detector is found, given the partial results.
2) The exemplary variants are listed in Table II. Parameters for image pre-processing and feature extraction are fixed for each variant (e.g. sigma for Gaussian blur or window size for SURF). Since the current examination deals with the number of feature vectors, separate candidates are created for each number of components.
3) There are two separate image sets: the training set and the evaluation set.
4) For each classifier, based on the partial results, the parameter with the greatest impact on the final result is selected. Each candidate is trained on a training set of pictures with varying values of this parameter. For each value of this parameter, the trained joint detector generates results for the evaluation set.
5) Comparing the results on the evaluation set with annotation, the number of positive and false detections is calculated for each value of the changing parameter. The R value is also calculated for each parameter value (see subsection II-F1). All R values computed for one candidate create the ROC. The area under this ROC (AUC) is computed and used as a quality measure for the joint detector candidate.
6) The best joint detector candidates with the highest AUC value are selected. Additionally, other candidates are considered, but only those that differ from the best candidates in the number of SURF components.
7) FPC histograms are created for each candidate. The best joint detector is the one with the largest number of pictures that has FPC=0.

III. EXPERIMENT

This section discusses the partial and final results obtained during the experiments.

A. Partial results

Partial results are obtained in three phases, described separately below, to find promising sets of parameter values for the methods.

1) Preselection: Preselection of the pixels was made for SURF, ORB, FAST, ORB, and BRISK detectors. The results were very similar. Generally, this approach does not generate satisfactory results: the preselected area was too small and does not contain the joint center, or is too vast (ca. 50% of the image). Therefore the dense-like selection was used in further processing.

2) Feature extraction: Figure 3 shows the results obtained from the two descriptors: SURF [3] and ORB [6]. The i-th series in both graphs represents the SURF and ORB descriptor values of the same pixel (indicated joint center) obtained from the same image. The plot is made for 5 visually different pictures. It can be observed that the series plots for SURF are visually more similar to each other than those for ORB. This was also confirmed experimentally: for ORB [6], BRISK [5], and FAST [4] descriptor classification results were not satisfying: 95% of pixels were assigned to one class.

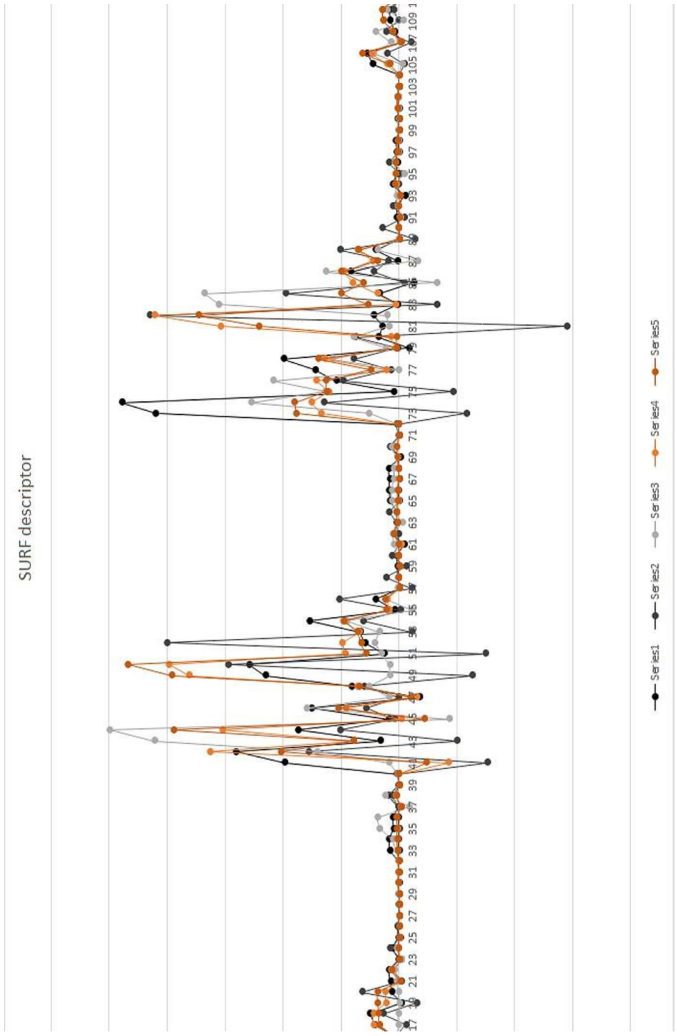


Fig. 3. The results obtained in the feature extraction step from the two descriptors: SURF [3] and ORB [6]

3) Classification: Table II lists 14 scenarios selected from several thousand. The selected scenarios produce the best results in each detector class when evaluated using classifier tests. There are 6 scenarios selected for the SVM, 4 for DT, and 4 for NN. The best results were obtained for SVM with Radial Basis Function Kernel: $K(x_i, x_j) = e^{-\gamma||x_i - x_j||^2}$ with $\gamma = 0.85$.

B. Final results

The final results were generated from AUC analysis and FPC analysis, and details are in the following subsections.

1) AUC analysis: AUC analysis described in Section II-F1 has been applied to thousands of joint detector candidates. The best 14 detector configurations and their AUC values are presented in Table II. The ROC plots for the best detectors based on different classifiers are shown in Figure 4. Detectors using the SVM classifier have much higher AUC values than others (see the last column AUC in Table II) and at the same time, the worst result of the SVM detector has a higher AUC than the best detector using another classifier (as shown in

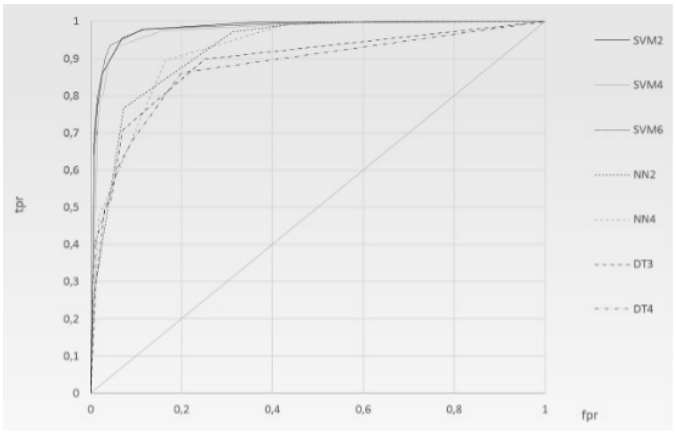


Fig. 4. ROC plots for 7 of all examined detectors. ROC - Receiver Operating Characteristic Curve, SVM - Support Vector Machine, NN - nearest neighbor classifier, DT - decision tree classifier, tpr - true positive rate, fpr - false positive rate

Table II). The two best detectors have the same AUC value of 0.981, which are named SVM2 and SVM6 marked by "+" in Table II. The SVM2 was chosen for the FPC analysis because it has a simpler preprocessing component.

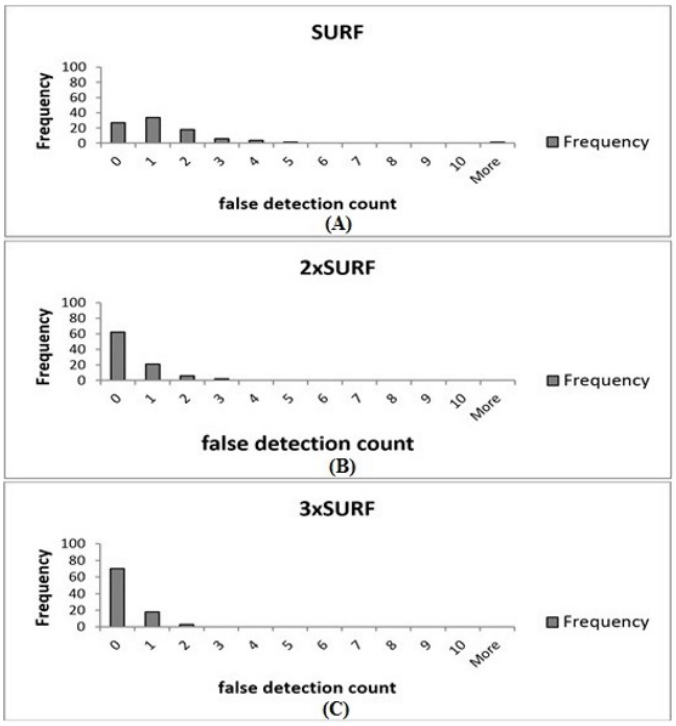


Fig. 5. ROC plots for 7 of all examined detectors. ROC - Receiver Operating Characteristic Curve, SVM - Support Vector Machine, NN - nearest neighbor classifier, DT - decision tree classifier, tpr - true positive rate, fpr - false positive rate

2) FPC analysis: Figure 5 shows improvements resulting from applying SVM2 multi-component SURF descriptors with varying sizes of the window on which the descriptor is computed. The plots SURF, 2xSURF, and 3xSURF display the effect on the FPC histogram calculated for consecutive

components being added to the feature vector. In the best case of 1 component SURF feature vector (window size=250 pixels), only about 24% of the images had 0 false detections and the maximum number of false detections was 5. Applying the second SURF component (window size=150 pixels) increased the frequency of images with 0 false detections to 61% and decreased the maximum number of false detections to 3. The feature vector with 3 SURF components and window sizes: 250, 150, and 80 pixels gave even better results: over 73% of images had 0 false detections and the maximum number of false detections was 2. Windows larger than 250 pixels were not used due to the high computational complexity of the SURF descriptor.

IV. Conclusion

The described experimental design procedure led to the formulation and implementation of a joint detector with the following properties: in all test images there is at least one true positive detection and the count of false detection clusters is minimized. Furthermore, in 73% of the test images, there are 0 false detections and the maximum number of false detections in each image does not exceed 3. Three useful observations were made from the experimental results: changing the image with filters except blur degrades the performance of the joint detector, combining SURF descriptors with different window sizes reduces the number of false detections, and the non-SURF local descriptors were inferior to SURF in computing the neighborhood similarity at characteristic points in USG images. The optimized joint detector produces surprisingly good results considering the visual variability of a joint in USG images.

The main assumption was proven true. The optimization process decreases the number of pictures with multiple detections from a level of 72% in previous work [Wereszczynski, 2014] to 27% in current works. We also proved that increasing the number of components of homogeneous mixed-SURF descriptors also decreases the count of pictures with multiple detections from 76% in the case of one descriptor, to 27% in the case of 3 components. The maximum number of false detection is also decreased from 5 to 2. Due to the high computational complexity, experiments with larger numbers of SURF components have not been performed. Such experiments may be performed in the future using CUDA processors.

We have proven that non-homogeneous mixtures of descriptors give very poor results compared to homogeneous SURF descriptors. The reason appears to be not homogeneity by itself, but the fact that the different descriptors tested, produced very poor results. To confirm this statement, further research should be carried out for non-homogeneous descriptors.

In summary, our research led to the creation of a better joint detector which should have a positive influence on the final results of inflammation level classification.

References

[1] M. ØStergaard and M. Szkudlarek, "Ultrasonography: A valid method for assessing rheumatoid arthritis?" *Arthritis & Rheumatism*, vol. 52, no. 3, pp. 681–686, 2005. [Online]. Available: https://doi.org/10.1002/art.20940

[2] K. Wereszczyński, J. Segen, M. Kulbacki, P. Mielnik, M. Fojcik, and K. Wojciechowski, "Identifying a joint in medical ultrasound images using trained classifiers," in *Computer Vision and Graphics*. Cham: Springer International Publishing, 2014, pp. 626–635. [Online]. Available: https://doi.org/10.1007/978-3-319-11331-9_75

[3] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," in *Computer Vision – ECCV 2006*, A. Leonardis, H. Bischof, and A. Pinz, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 404–417. [Online]. Available: https://doi.org/10.1007/11744023_32

[4] E. Rosten and T. Drummond, "Machine learning for high-speed corner detection," in *Computer Vision – ECCV 2006*, A. Leonardis, H. Bischof, and A. Pinz, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 430–443. [Online]. Available: https://doi.org/10.1007/11744023_34

[5] S. Leutenegger, M. Chli, and R. Y. Siegwart, "Brisk: Binary robust invariant scalable keypoints," in *2011 International Conference on Computer Vision*, 2011, pp. 2548–2555. [Online]. Available: https://doi.org/10.1109/ICCV.2011.6126542

[6] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: An efficient alternative to sift or surf," in *2011 International Conference on Computer Vision*, 2011, pp. 2564–2571. [Online]. Available: https://doi.org/10.1109/ICCV.2011.6126544

[7] C.-C. CHANG, "Libsvm : a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011. [Online]. Available: https://cir.nii.ac.jp/crid/1574231874006333696

[8] L. Breiman, *Classification and regression trees*. Routledge, 2017. [Online]. Available: https://doi.org/10.1201/9781315139470

[9] M. Muja and D. G. Lowe, "Fast approximate nearest neighbors with automatic algorithm configuration." *VISAPP (1)*, vol. 2, no. 331-340, p. 2, 2009.

[10] "Automated assessment of joint synovitis activity from medical ultrasound and power doppler examinations using image processing and machine learning methods." [Online]. Available: http://eeagrants.org/project-portal/project/PL12-0015

[11] D. J. Hand and R. J. Till, "A simple generalisation of the area under the roc curve for multiple class classification problems," *Machine Learning*, vol. 45, no. 2, pp. 171–186, Nov 2001. [Online]. Available: https://doi.org/10.1023/A:1010920819831