





# Content Based Image Retrieval in Digital Pathology Using Speeded Up Robust Features

A. Kallipolitis  and I. Maglogiannis <sup>(✉)</sup> 

Department of Digital Systems, University of Piraeus, Piraeus, Greece  
nasscall@yahoo.gr, imaglo@unipi.gr

**Abstract.** The recent expand in the utilization of Whole Slide scanners in Digital Pathology gave birth to a production of massive amount of data and the need of integration of Digital Pathology Systems (DPS's) into modern Laboratory Information Systems (LIS's). In this context, the problem of automatically retrieving a particular image from a large set of digital images that contains similar medical visual content has gained fruitful ground. This work investigates the fast and consistent properties of the Speeded-Up Robust Features (SURF) algorithm in order to search in the content of a digital pathology image, detect and find similarities for content-based image retrieval. An important aspect of this work is the diversity of Whole Slide Scanners. The proposed methodology that involves the process of the comparison of digital pathology images, mostly WSI, with the use of the SURF algorithm was proved robust to various condition changes.

**Keywords:** Digital pathology · Laboratory information system  
Content based image retrieval (CBIR) · Speeded Up Robust Features (SURF)  
Whole Slide Imaging (WSI)

## 1 Introduction

Digital Pathology was meant to solve many problems that physicians in this domain had faced in the earlier years, mainly associated with the management and preservation of tissue samples, the inability of conducting tele medical consultations and the lack of advanced computer based systems for diagnosis, analysis and education. Nevertheless, the transition to a new digital era of pathology brought up many new challenges. A vast amount of data is created every second as the digital image produced by a glass slide has a typical size of 3 GB along with its metadata. Analyzing such images in order to recognize patterns and similarities against images found in medical books and atlases have been proven tedious and time-consuming tasks for pathologists. There is an immense need for automated and fast book, database and storage systems screening. Furthermore, the variety of whole slide scanners vendors led to the building of a new “Babel” tower, where each DPS speaks a different language as far as hardware, operating systems, formats of digital images and communicating protocols are concerned [9]. To address these two challenges, the paper describes a system that, at first stage, is able to translate the different formats of WSI and, at second, detects similarities in pathology images. The system can also import medical books in pdf format and

extract the images found in them for storage and analysis along with whole slide images. A general overview of the system is shown in Fig. 1. Inputs to content based image retrieval system can be whole slides images, images extracted from digital pathology books (literature) and local storage or digital pathology databases. One of these images represents the query image and the others the test images. These inputs, digital images to their whole, are imported in the system, processed and the output of the process is the set of images that bear the greatest resemblance to the query image.

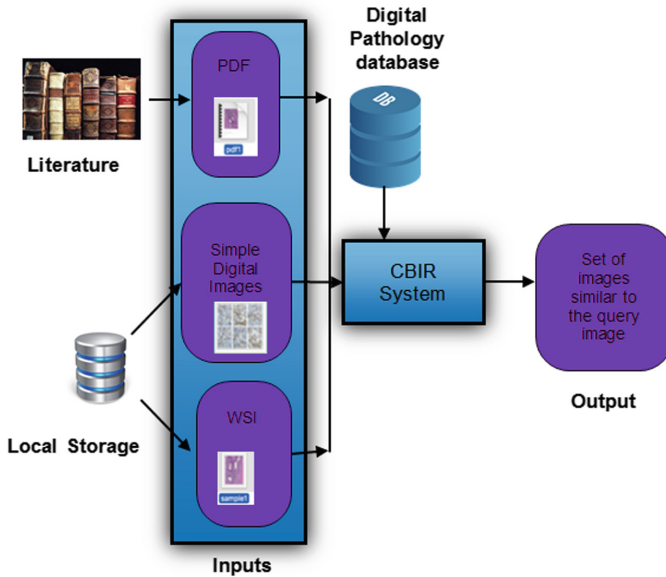


Fig. 1. General view of image retrieval system

The remainder of this paper is structured in 5 sections, as follows: Sect. 2 presents the related work, while Sect. 3 describes the proposed methodology for image retrieval. Section 4 describes the system in practice and Sect. 5 reports the experiments conducted and the corresponding results. Finally, Sect. 6 concludes the paper.

## 2 Related Work

The discipline of pathology has well accepted its “digital” character, since several years of evolution transforming the conventional LISs (Laboratory Information Systems) to their modern digital pathology version. This new version embodies several technologies such as Relational Database Management Systems (RDBMSs), Structured Query Language (SQL), eXtensible Markup Language (XML) and, most significantly, WSI (Whole Slide Imaging) [1]. The process of WSI is based on Whole Slide Scanners that produce Whole Slide Images and, more than often, participate in DPSs (Digital Pathology Systems), whose integration/cooperation within the LIS’s have improved the

functionality of the later in means of diagnostic assistance, educational tools and quality control [2]. Through DPSs and Whole Slide Scanners, glass slides that are used in light microscopes by pathologists to observe and examine specimens taken from surgery, are transformed into Whole Slide Images (digital slides). The digital slides can be read by viewing software in order to navigate into a vast range of standard magnifications of sectors of the specimen. However, the variety of manufacturers that are involved in the production of different whole slide scanners poses a complicity factor in the process of reading and viewing such images. This complicity factor is explained by the fact that each digital slide, which is produced by a brand scanner, has a certain format, different from all the other formats produced by other scanners. This lack of uniformity between formats of digital slides that are produced by different manufacturers will continue to exist despite the recent efforts for the creation of standards that will pose certain patterns and regulations in the process of WSI [3].

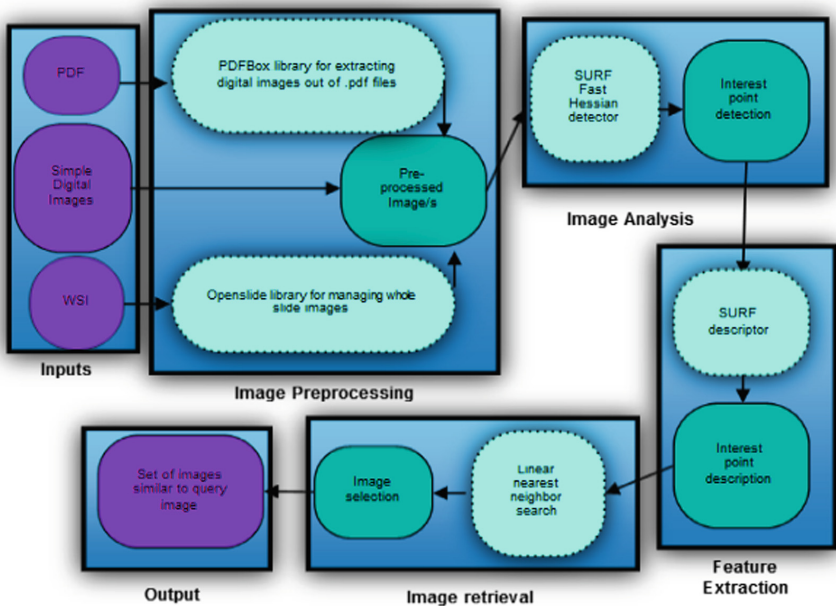
Apart from the difficult task of handling heterogeneous whole slide images, this paper proposes a solution to the problem of retrieving images from a dataset of digital images that are similar to a query image. Once the first technical task of deploying a universal digital image viewer is completed successfully, most formats of digital slides are available for preview, analysis, interpretation and handling. In the vast bank of data that digital images can provide, similarities are tracked and found between a query image and a dataset of images for the scope of retrieving the most similar image/s. For the purpose of detecting specific areas of interest where similarities can be found, a local feature representation, using the SURF algorithm [5], of the image is proposed due to the algorithm's promising properties concerning speed and robustness to variations. With the exploitation of SURF, interest points are, first, detected in an image and, then, the image is described by multiple 64d vectors, created in each interest point. Consequently, the image from the dataset that bears the most similarities with the query image is selected and retrieved. Similarities are defined by the score that is nominated to its image of the dataset when their interest points are compared with the interest points of the query image, not by the "naive" one-to-one comparison but by the implementation of the nearest neighbor algorithm with kd-trees.

The procedure of retrieving the image with the most similarities to the query image (best score) is explained in CBIR (Content Based Image Retrieval) methodology. For many years, image retrieval was based on text found in the meta-data attached to digital images. Text based image retrieval had faced several challenges from the early 70s up to today, mainly, concerning human effort to add annotations to images and inconsistencies between textual meta data information which describes visual content. These unsolved challenges lead to the advance of CBIR in the 80s, which has stepped forward to gain excessive terrain as a preferred method of solving image retrieval problems. In CBIR, images are indexed by their visual content automatically without the need of human intervention. CBIR can be performed in images that are represented either globally, or locally. As explained in [7], local feature representations can provide the localized information of interest from a small region of the whole slide image, whereas global feature representations fail to contribute to image retrieval due to the fact that the produced global signature depicts the whole image and not the small area of interest. In reference to relevant research work that has been conducted in CBIR and Digital Pathology, Mehta et al. in [6], retrieves sub-images from whole slide images using

scale-invariant feature extraction algorithm SIFT (Scale Invariant Feature Transform). However, SURF algorithm has been proven to be fast, more accurate and reliable than SIFT. Velmurugan et al. in [8], used a combination of SURF and Color Moments to retrieve similar images of general context. The basic assumption to combine Color Moments in order improve the results of SURF algorithm lies on the fact that SURF is applied on gray scale images. In [9], Govindaraju and Ramesh Kumar propose a novel CBIR system using SURF and Bag of Words to detect resemblances in medical images. Results are compared with the use of both SIFT and SURF algorithms with the latter to prove better.

### 3 Methodology

The method, which is proposed in this paper to solve the image retrieval problem, consists of four (4) stages (illustrated in Fig. 2), as follows: (i) Image preprocessing, (ii) Image analysis, (iii) Feature extraction and (iv) Image retrieval.



**Fig. 2.** Workflow diagram of the proposed methodology (Magenta boxes correspond to input/output, light green to processing steps and dark green refers to intermediate objects) (Color figure online)

Starting from the input of our system, they can be digital images downloaded from the Internet, selected regions of interest from whole slide images or images extracted from digital pathology atlases. One of these images represents the query image and all the others comprise the dataset from which the system will detect the image with the

best similarity to the query image (output of the system). The following paragraphs discuss briefly the 4 stages.

- (i) Image pre-processing. In this first stage the system creates the input images whether they take part in the search dataset, or they constitute the query image. It is important to note that not all images are ready to be analyzed by the system, as they may be contained in a PDF file or they need to be extracted from a whole slide image (being a part of it). In the case of handling.pdf files, the initial.pdf file is being processed by the system and all images are extracted in a folder by means of a java library called PDFBOX. Should a whole slide image be used, another java library called OPENSLIDE, reviewed in [10], offers its functionality in order for the whole slide images to be read, previewed, and processed. If the input image is a simple digital image the system's task is trivial. In the end of this stage all digital images are ready to be analyzed, as they enter the following stage of image analysis.
- (ii) Image analysis. Each digital image is being analyzed and the interest points are detected using the fast Hessian detector of SURF algorithm. The detector function is divided in two basic steps:

1. Tracking of interest points by means of the determinant of the approximated Hessian Matrix. This is accomplished by the use of filter boxes and integral images.
2. Detection of interest points by means of the non-maximum suppression technique.

In the end of this stage all interest points from all images are detected.

- (iii) Feature extraction. Since all interest points are detected, the descriptor of SURF algorithm creates a 64-dimension vector that, uniquely describes each robust interest point of all images. The descriptor function is summarized in two basic steps, involving the orientation assignment to each interest point using Haar wavelets responses and the vector computation by adding Haar wavelets responses in horizontal and vertical axis. In the end of this stage each image has interest points that are described by 64-dimension vectors in a unique and robust manner.
- (iv) Image retrieval. In the last stage the images with the highest score are selected. The score is calculated by adding the number of matches between the interest points of the query image and the candidate image. A match is found by calculating the Euclidean distance between vectors of the descriptors. This process is not accomplished in a naive way (one-by-one exhausting search), but by means of linear nearest neighbor search to avoid large computational cost. The results of the abovementioned methodology are influenced by the parameters of the SURF algorithm. In the following lines a laconic definition of the parameters of the parameters octaves and threshold is given in order to provide an understanding of their role in the implementation of SURF.

**Hessian threshold:** The minimum value of the determinant of the Fast Hessian Matrix for a feature point should be selected. This parameter is closely connected to the repeatability of the algorithm, which means that by lowering the threshold the outcome is weaker feature points with less repeatability.

**Octaves:** A series of filter response maps obtained by convolving the same input image with a filter of increasing size. The Gaussian pyramid of the scale space is divided into octaves, which in turn are divided into layers. The increase of octaves detects larger features but costs in time.

**Layers:** The next level of division of the scale space.

**Initial step:** The initial sampling step, which is doubled for each next octave. This value determines how many pixels separate each pixel in the given pyramid octave.

In the next Section we discuss the usage scenarios of the implemented system.

## 4 The System in Practice

An application was development in java programming language utilizing the libraries Openslide ([www.openslide.org](http://www.openslide.org)), ImageJ Surf ([www.labun.com/imagej-surf](http://www.labun.com/imagej-surf)), and PDFBOX ([www.pdfbox.apache.org](http://www.pdfbox.apache.org)). The GUI of the application, as depicted in Fig. 3, is divided in two panels, the screen panel and the control panel. Four basics buttons are provided on the upper section of the main menu, as follows:

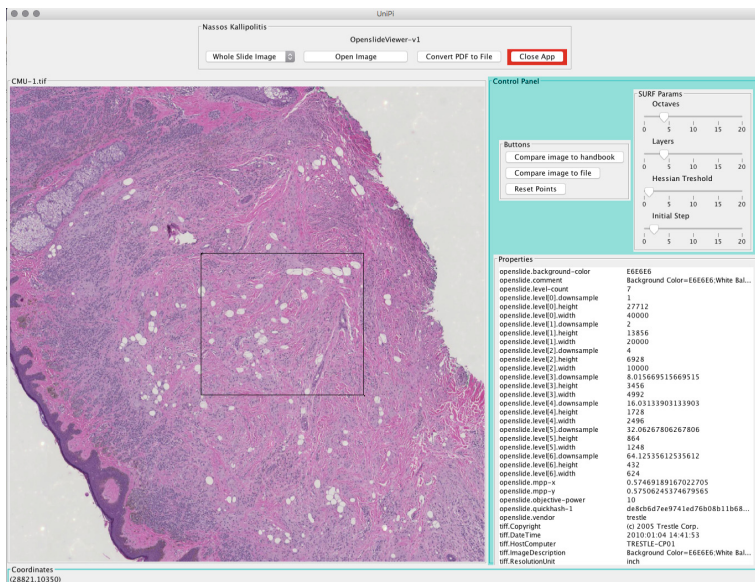


Fig. 3. Main menu of the application

- Simple Image/WSI
- Open Image
- Convert PDF to file
- Close App

The functionality of each button is briefly described below.

- a. Simple Image/WSI. By pressing this button the type of the query image is selected. The user can choose between two options: Simple digital image or whole slide image.
- b. Open Image. Selected query image is read and opened in order to be viewed and processed (for WSI).
- c. Convert PDF to file. A digital pathology atlas can be selected and converted in a folder of digital images.
- d. Close App. Self-explained.

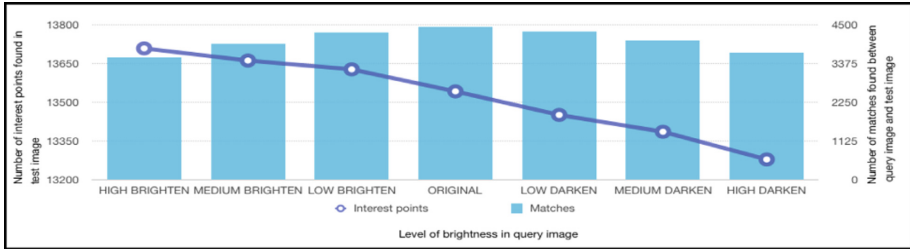
Once a query image is selected, the requested image is visible in the screen panel section, as shown in Fig. 3. In this case the query image is a whole slide image produced by a Trestle whole slide scanner and its format is single-file pyramidal tiled TIFF (tagged image file format). In the control panel additional functionality appears by means of two new panels, SURF parameters and properties, and three buttons: compare image to handbook, compare image to folder and reset points. By using the SURF parameters panels the user can specify the parameters of the SURF algorithm for the detection and description of interest points. In the properties field the user can view the meta-data that are stored in a whole slide image in reference to the image attributes. The buttons serve the purpose of choosing the dataset of digital images that will be compared to the query image. By choosing dataset of images and query image, the process of image retrieval begins and in the end the result appears in the screen and the image/s with the most similarities to the query image is/are shown.

## 5 Experimental Results

A series of experiments is conducted by comparing a query image with another sample image. The images can be simple images (.jpg, .png, .bmp), whole slide images or images extracted from medical atlases in.pdf format. The query image is altered applying different brightness, rotation and scale transformations. Each time the SURF algorithm is applied and a number of matches between the query image and the image dataset are detected. Basic criteria of understanding the level of influence posed by the transformations is the number of interest points detected in each transformed image (test image) and the number of matches found between the query image and each test image.

The first set of conducted experiment uses the query image and a set of six variations of the query image from the brightest to the darkest one. Results are shown in Fig. 4.

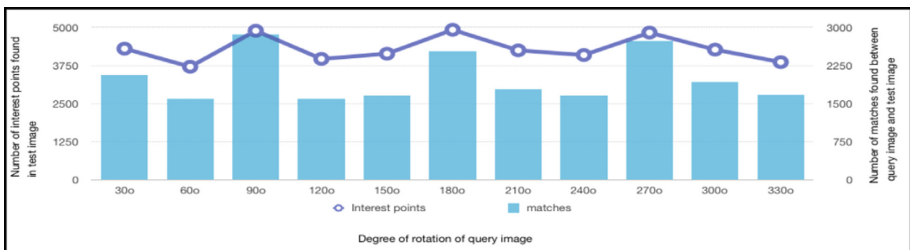
The number of interest points increases as the image gets brighter and decreases as the image gets darken. However, the increase of the number of interest points does not



**Fig. 4.** Graph of matches found between query image and test image and interest points found in test image as the brightness changes

necessarily mean increase of the numbers of matches between the query and test image. The influence of the variations in brightness is slightly stronger when the pathology image gets darken (22% less matches) in respect to the brightest image (19% less matches).

The second set of experiments refers to rotation transformations. Results are illustrated in Fig. 5.



**Fig. 5.** Graph of matches found between query image and test image and interest points found in test image as the rotation changes

The worst case for the rotation transformations occurs at  $120^\circ$  (37% less matches), which is more intense than the worst case for brightness transformations. Transformations related with the scale up and down of the query image are also tried out to check the effect of these transformations to the function of SURF algorithm. The results are shown in Fig. 6.

As it is projected by the graph (Fig. 6) the effect of minimizing the query image is devastating more than any other transformation performed (97% less matches). Maximizing the image has a smoother impact to the algorithm with a 33% decrease on matches. Apart from the experiments performed with reference to the transformations of the query image, tests were conducted with the different values of the following parameters octaves, hessian threshold, explained earlier in Sect. 3, are assigned to check the influence of these variations.

The results for variations of the octaves and the threshold parameter are shown in Figs. 7 and 8.



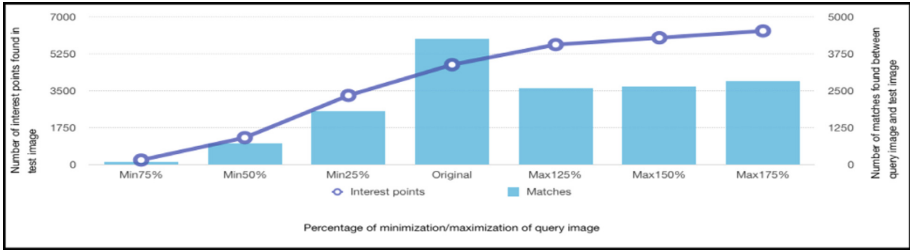


Fig. 6. Graph of matches found between query image and test image and interest points found in test images as the scale changes

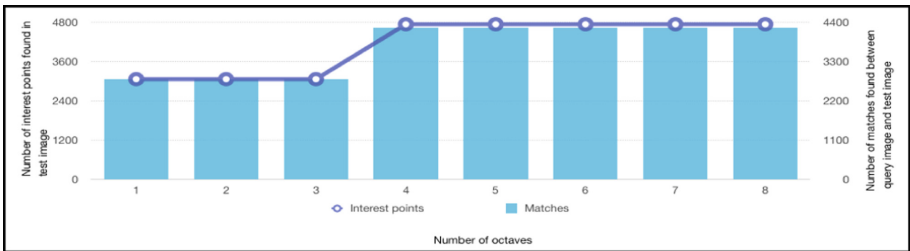


Fig. 7. Graph of matches found between query image and test image and interest points found in the query image as the octave parameter changes

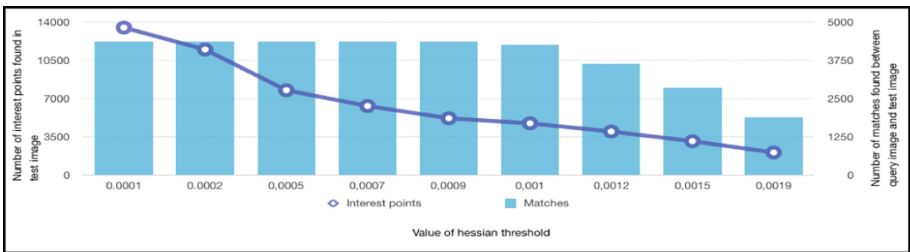


Fig. 8. Graph of matches found between query image and test image and interest points found in the query image as the threshold parameter changes

In Fig. 7 it is highlighted that the most matches are provided for four octaves and more, whereas in Fig. 8 the best results are given for a 0.00009 value of the parameter threshold, since this value ensures the most matches for the less interest points found.

One last experiment is conducted on 4 groups of similar images, which depict the effect of three different treatments (drug, radiation, drug and radiation) on cancer cells. The dataset, which is comprised by .jpg images, is described in [10]. This experiment proves the effectiveness of SURF algorithm on retrieving the most similar image/s even when the examined dataset is comprised by images that bear great resemblance. Using the application, the query image, selected from one the groups (control, drug, radiation,

drug and radiation), is compared to each image of the dataset. Once the image with the greater similarity (best image) to the query image has been found, the query image is classified in the group of origin of the best image with success 67%, as illustrated in the confusion matrix below (Table 1).

**Table 1.** Confusion matrix of classification of 24 images in four classes (control, drug, radiation, drug n radiation)

Predicted class	Actual class			
	Control	Drug	Radiation	Drug and radiation
Control	5	0	0	1
Drug	0	4	1	1
Radiation	0	2	3	1
Drug and radiation	0	1	1	4

## 6 Conclusion

The retrieval of digital pathology images in references to similar images found in medical literature, storage systems and medical databases is the main goal of the work presented in this paper. The proposed methodology deploys a speeded up robust feature extraction technique along with a viewer that reads whole slide images. The experiments conducted confirm that the SURF algorithm fulfills the process efficiently and fast, mainly due to the small computational effort (integral images, box filters). To give a rough estimation of speed, the system compares 573 images (with the sized of 2.12 GB) extracted from a medical atlas called “Surgical Pathology of the Gastro-nomical System” in 5 min. The specific processing time refers to a desktop computer with a 3.06 GHz Intel “Core 2 Duo” Processor (E7600) and 4 GB RAM. Moreover, the robustness of SURF algorithm is verified to all brightness, rotation and scale up variations of the query image apart from the scale down transformation where the efficiency of the system is proven relatively low. Even in the worst case scenario (37% less matches) the final outcome of the system is not compromised, because the remaining matches are adequate for the system to correctly classify the query image in most cases. The initial goal of the system was the retrieval of specific images that bear resemblance to others. Nevertheless, results extracted from experiments, which are performed on cancer cells digital images, imply the capability of the system to classify an image to a certain class. A suggestion for future work would be to apply the Visual Bag of Words technique to create a single vector from each image and use the results for machine learning classification. Concluding, this work might be considered as a significant adjunct tool for pathologists in their everyday work, assisting them in searching atlases, electronic books and other electronic resources for similar cases to the one that they have in front of them. Additional evaluation with real users is required to validate this assumption.

## References

1. Park, S.L., Pantarowitz, L., Sharma, G., Parwani, A.V.: Anatomic pathology laboratory information systems. *Adv. Anatomic Pathol. Rev.* **19**(2), 81–96 (2012). <https://doi.org/10.1097/pap.0b013e318248b787>
2. Ellin, J., Haskitz, A., Premraj, P., Shields, K., Smith, M., Stratman, C., Wrenn, M.: Interoperability between Anatomic Laboratory Information Systems and Digital Pathology Systems. Digital Pathology Association. <http://www.digitalpathologyassociation.org>
3. Singh, R., Chubb, L., Pantarowitz, L., Parwani, A.: Standardization in digital pathology: Supplement 145 of the DICOM standards. *J. Pathol. Inform.* **2**, 23 (2011)
4. Bay, H., Tuytelaars, T., Gool, V.G.: Speeded up robust features. *Comput. Vis. Image Underst.* **110**(3), 346–359 (2008). <https://doi.org/10.1016/j.cviu.2007.09.014>
5. Mehta, N., Alomari, R.S., Chaudhary, V.: Content based sub-image retrieval system for high resolution pathology images using salient interest points. In: Annual International Conference of IEEE Engineering in Medicine and Biology Society, EMBC 2009 (2009)
6. Shyu, C. R., Brodley, C. E., Kosaka, A.C., Aisen, A., Broderick, L.: Local versus Global Features for Content-Based Image Retrieval. School of Electrical and Computer Engineering. Purdue University. Indiana University Medical Center
7. Velmurugan, K., Santhosh Baboo, S.: Content-based image retrieval using SURF and . Global J. Comput. Sci. Technol. **11**(10), Version 1.0 (2011)
8. Govindaraju, S., Ramesh Kumar, G.P.: A novel content based medical image retrieval using SURF features. *Int. J. Comput. Sci. Inf. Technol.* **4**(2), 242–245 (2016). <https://doi.org/10.17485/ijst/2016/v9i20/89786>
9. Goode, A., Gilbert, B., Harkes, J., Jukic, D., Satnarayanan, M.: Openslide: a vendor-neutral software foundation for digital pathology. *J. Pathol. Inform.* **4**, 27 (2013)
10. Goudas, T., Maglogiannis, I.: An advanced image analysis tool for the quantification and characterization of breast cancer in microscopy images. *J. Med. Syst.* **39**(3), 13 (2015). <https://doi.org/10.1007/s10916-015-0225-3>