Segmentation of fish chromosomes in microscopy images: A new dataset

Rodrigo Júnior Rodrigues*, Rubens Pasa†, Karine Frehner Kavalco†, João Fernando Mari*
Universidade Federal de Viçosa - Campus Rio Paranaíba,
*Instituto de Ciências Exatas e Tecnológicas,
†Laboratório de Bioinformática e Genômica,
Caixa Postal 22 – 38.810-000 – Rio Paranaíba – MG – Brazil
Email: {rodrigo.rodrigues, joaof.mari}@ufv.br

***Abstract*—The chromosome segmentation is the most important step in automatic karyotype assembling. In this work, we presented a brand new chromosome image dataset and proposed methods for segmenting the chromosomes. Chromosome images are usually low quality, especially fish chromosomes. In order to overcome this issue, we tested three filters to reduce noise and improve image quality. After filtering, we applied adaptive threshold segmentation combined with mathematical morphology and supervised classification methods. Support Vector Machine and k-nearest neighbors were applied to discriminate between chromosomes and image background. The proposed method was applied to segment chromosomes in a new dataset. To enable measure the performance of the methods all chromosomes were manually delineated. The results are evaluated considering the Hausdorff distance and normalized sum of distances between segmented and reference images.**

***Index Terms*—Fish karyotype, chromosome segmentation, computer vision, classification, new dataset.**

I. Introduction

The karyotype assembling is an important task in cytogenetics. It is useful in a number of practical and research activities, such as assist the diagnosis of genetic diseases and some types of cancer types [1]. The chromosomes are composed of supercoiled and associated DNA. Human chromosomes can suffer related anomalies to an atomic number of chromosomes or structural abnormality in one or more chromosomes [2]. The human cells contain 46 chromosomes including 22 pairs of chromosome and two sex chromosomes (XY: male and XX: female). Fishes have a variable number of chromosomes and they cannot be previously defined as in humans [3]. The process of chromosomal karyotyping is performed by pairing the chromosomes according to the similarity between them. The chromosomes are classified in one of the four classes according to the location of the centromere: metacentric, submetacentric, subtelocentric, and acrocentric [4] [5].

In the process of segmentation, the images are first converted to binary format. The binary images help to find details about the object shapes [6]. Chromosomes are cellular structures that contain genetic information. When chromosomes are imaged using a microscopy, information about the health of an individual. Since the 1980s, chromosome detection and classification systems have aroused great interest in research. The manual assembling of a karyotype is repetitive, exhausting, time consuming, and subject to error. It can be performed by visual analysis but requires specialized professionals.

The segmentation of chromosome is the most important step in automated analysis of chromosomes [7]. An automated system generally includes the following four steps: (1) image enhancement, (2) segmentation and alignment of the chromosomes, (3) chromosome feature selection, and (4) chromosome classification. The chromosome segmentation is the most important step because their results can affect the performance of the entire system [8].

In this work, we study and compare approaches for segmentation of fish chromosomes in digital images combining filtering operations, segmentation, and morphological operations. Mean filter, median filter, and Non Local Means filter are used to reduce the noise and improve the image quality. Segmentation is performed using adaptive threshold followed by morphological operations.

Supervised classifiers, such as Vector Support Machine (SVM) and k-nearest neighbor (KNN) are applied to discriminate the segmented objects in chromosomes and artifacts (objects that are not chromosomes). The methods were implemented and tested in an image dataset with ground-truth. Finally, we analyze the methods performance considering Hausdorff distance and NSD metrics and compare the implemented approaches.

This paper is organized as follows: This section introduces the subject. Section II shows some related work on chromosome segmentation. Section III describes the new image dataset we created, as well as the proposed methods to automatically segment the chromosomes and the validation methods. In Section IV we present and discuss the results and the conclusion and future works are in Section V.

II. Related Works

Aln W. and Jane Y. [9] developed an algorithm based on an adaptive local kernel (KAFCM) and a classifier of Probabilistic defuzzification to improve segmentation and classification of chromosomes. This is achieved on a window for each pixel and compensate for the intensity of the homogeneity caused during the process of generation of images and by the preparation of the physical chromosome itself. The algorithm was tested on a publicly available dataset and the results were compared with traditional fuzzy clustering algorithms. The classification results for the proposed method are for defuzzification of standard FCM were compared, and the proposed classification method demonstrated an improved overall.

Monika S. et al. [10] proposed a method to segment and classify chromosomes in healthy patients combining deep-learning and pre-processing methods and crowd-sourcing. The experiments are performed on 400 images taken from healthy patients. For the subset with better images quality, the classification rate is about 95%

Madian et. al. [11] studied the chromosome segmentation considering boundary information. Otsu threshold, morphological operations, and filling holes after binarization were applied. A curvature function was applied to find cut-off points in the object edges. The concavity points at the edges are used to detect chromosomes overlapping zones. The method has been tested on over 350 images with several degrees of overlap and obtained an overall accuracy of 96%.

Karvelis *et al.* [12] present a method for the segmentation of groups of chromosomes that touch each other and chromosomes superimposed on M-FISH images. Initially, the watershed transform is applied and the image is decomposed in regions. Gradient paths are calculated from points of high concavity and used to divide the groups of chromosomes. To validate the method they used a reference dataset composed of 183 M-FISH images. The algorithm resulted in a success rate of 90.6% for the chromosomes that are touched and of 80.4% for the groups of chromosomes that are superimposed.

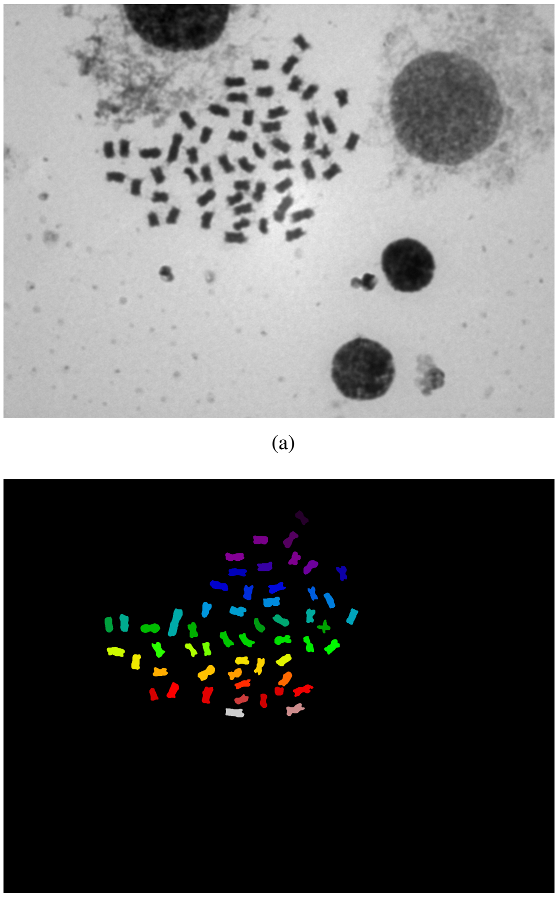
Rodrigues et al. [13] compared two approaches to segment overlapping chromosomes, one based on morphological skeleton and the other based on restricted Delaunay triangulation. Restricted Delaunay triangulation demonstrates to achieve better results then morphological skeleton.

Saiyod and Wayalum [14] developed an approach to compute the skeleton of the chromosomes. With the skeletons it is possible to search for points of intersection. The point of intersection is used to search candidate cut points. The cut-off points are found by calculating the Euclidean distance from the point of intersection to the points of curvature. The nearest four points are the points of interest

III. MATERIAL AND METHODS

A. Dataset

The dataset was constructed in the Bioinformatics and Genomics Laboratory, at the campus of UFV in Rio Paranaíba - Brazil. The images were captured using a microscope Olympus BX 41 (Olympus Inc., Japan) with a 3 MP and a magnification of 1000x using the software Qcapture Pro 6.0 (QImagine, Surrey, BC, Canada). The images were converted to grayscale and for each image, we created a reference image in which each chromosome were manually segmented using the image processing software Gimp. Each image has a size of 1250 x 1250 pixel and have been saved as hdf5 file forming (Figure 1).


(a)

(b)

Fig. 1: The Chromosome dataset. (a) Original image (b) Labeled image

B. Filtering

Generally, chromosome images have low quality, they are low contrast and it is possible to observe the presence of noise and artifacts. This problem is worse when we are dealing with fish chromosomes because they are smaller than human chromosomes. Thus, the preprocessing step is very important for the segmentation [14]. We tested three filters: (1) median filter with mask size of 3 x 3; (2) average filter with mask size of 5 x 5 [15]; and (3) Non-Local Means filter (NLM) [16] [17] with standard deviation of 0.08 and h of 0.6.

C. Image segmentation in the background and chromosomes

In order to segment the images in pixels belonging to chromosomes and background pixels, we used a local adaptive threshold with a block size of 45 pixels. The block size was chosen empirically. After some experiments, we noticed that very large block sizes tend to generate connected chromosomes.

Mathematical morphology algorithms were applied to improve the quality of the binary image. Fill holes algorithms, based on morphological reconstruction [15] were applied to prevent holes inside the objects that may interfere with the

[0] Available in https://www.gimp.org

chromosome classification. A morphological opening operation using a disk-shaped structuring element with radius 2 is applied to break some isthmus and smooth out the object contours. Finally, we removed small objects (less than 120 pixels) which consist of artifacts resulting from the threshold segmentation [15] and removed the objects in the image borders.

D. Classification

A set of five features were selected to classify the connected components in chromosome or artifacts: (1) the area, (2) solidity, (3) eccentricity, (4) equivalent diameter, and (4) mean intensity. We divided our dataset in 80% of the images for training and 20% for testing. The features of all objects in the training set were used for training a Support Vector Machine (SVM) and a K-Nearest Neighbor (KNN) classifier. The dataset was split in an image-wise fashion since the chromosomes on each test image should be presented to the classifier only in the testing step.

After the training, the models were evaluated classifying the objects in the testing set in chromosomes or artifacts (any other segmented object which does not correspond to a chromosome). The metrics used to evaluate the classification were those derived from the confusion matrices: Precision (Equation 1); Recall (Equation 2), and F1-score (Equation 3) [18] [19]. These metrics are used to evaluate the performance of classifiers in the proposed methods.

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$F1 - Score = \frac{2 \times (Precision \times Recall)}{Precision + Recall} \tag{3}$$

where TP, FP and FN are *True Positive*, *False Positive* and *False Negative*, respectively.

E. Validation

To verify the efficiency and compare the methods described in this work we used the Hausdorff distance and the normalized sum of distances (NSD) methods. The Hausdorff distance is the largest minimum distance between the object borders in the segmented image I and in the reference image R , according to the Equation 4:

$$Hausdorff(I, R) = \max D(i) : S_i \neq R_i \tag{4}$$

where $D(i)$ is the distance between the pixel i of the object and the border of the reference object. The value 0 indicates a perfect segmentation, however, the index does not have an upper limit [20].

The normalized sum of distances (NSD) between the segmented image I and the reference image R is defined by Equation 5.

$$NSD(I, R) = \frac{\sum_i I_i \neq R_i * D(i)}{\sum_i D(i)} \tag{5}$$

where $D(i)$ is the distance between the pixel i and the border of the reference object. The value 0 indicates a segmentation perfect while 1 indicates that there is no overlap between the segmented cell and to the reference cell.

IV. Results

All images in the dataset described in Section III-A, a total of 97 images, were filtered according to the procedures described in Section III-B. Then the images were segmented in chromosome pixels and background pixels as described in Section III-C. The objects in the segmented images were manually labeled in actual chromosomes and artifacts (all segmented objects that are not chromosomes). These images were split in training and test sets which a proportion of 80 % and 20 %, as described in Section III-D and used to train an SVM and KNN classifiers.

Table I shows the classification results when the images were submitted to the mean filter, Table II is for when the images were submitted to the median filter, and Table III is for the NLM. We can observe the KNN had better accuracy, recall, and f1-score for all filtering strategies. These values where computed over objects in the testing set.

TABLE I: Classification results between SVM and KNN when applying the **mean filter**.

	precision	**recall**	**f1-score**
SVM	0.77	0.82	0.78
KNN	0.79	0.83	0.80

TABLE II: Classification results between SVM and KNN when applying the **median filter**

	precision	**recall**	**f1-score**
SVM	0.78	0.82	0.79
KNN	0.78	0.82	0.79

TABLE III: Classification results between SVM and KNN when applying the **NLM filter**

	precision	**recall**	**f1-score**
SVM	0.78	0.82	0.78
KNN	0.79	0.83	0.80

Tables IV, V, and VI shows the final segmentation results in terms of Hausdorf Distance and NSD. Table IV is for when images where filtered with mean filter, Table V, is for median filter, and VI is for NLM. As we expected, based on results in Tables I to I, KNN overcomes SVM considering all filtering strategies. Regarding the filtering strategy, the results are very close to each other. Before the classification median and NLM filtering are slightly superior than mean filter in therms of Hausdorf distance. After the classification, the values continues very close to each other, but it is clear that the classification process is essential for a good segmentation. These results were obtained after applying the specific filter process (Section III-B), the segmentation strategy described in Section III-C and, finally, the object classification with the trained models whose results are described in Tables I to III.

TABLE IV: Hausdorff distances, and NSD between the images resulting from the SVM and KNN with **mean filter** and the reference images.

	Hausdorff	**NSD**
No class.	7.03	0.69
SVM	7.38	0.46
KNN	6.50	0.37

TABLE V: Hausdorff distances, and NSD between the images resulting from the SVM and KNN with **median filter** and the reference images.

	Hausdorff	**NSD**
No class.	6.98	0.69
SVM	7.30	0.46
KNN	6.53	0.36

TABLE VI: Hausdorff distances, and NSD between the images resulting from the SVM and KNN with **NLM filter** and the reference images.

	Hausdorff	**NSD**
No class.	6.97	0.69
SVM	7.27	0.46
KNN	6.57	0.36

Figure 2 shows some segmented images using NLM filter and after classification with KNN. The first row shows the original image in grayscale, the second row shows the segmentation before classification, the third row shows the final segmentation after the classification, and the fourth row shows the the image considered as ground-truth.

V. Conclusions

This paper presented a comparison of approaches to segment chromosomes in microscopy images combining filtering techniques, adaptive thresholding, and classification methods.

Experiments were carried out using a newly constructed dataset of fish chromosome images with ground-truth. The images were obtained from the Bioinformatics and Genomics Laboratory of the Federal University of Viçosa in Rio Paranaíba - Brazil. Each image had its chromosomes segmented manually and saved in an h5py dataset. This dataset allows the development of this study, which investigate chromosome segmentation methods in a pragmatic way but also will be useful to a number of future works.

A number of filtering methods are compared (mean filter, median filter, and NLM filter) and tested in conjunction with two supervised classifiers used to improve the segmentation results. It can be seen that the filtering strategies have small effect over the segmentation results, however the object classification has a high impact on the quality of the results. The KNN classifier showed to be better than SVM for this task. Even so, images of fish chromosomes in metaphase state have a very large amount of noise and filtering strategy is still very important.

As future work one can consider testing the using the individually transformed watershed algorithm on chromosomes that touch each other. Another approach using neural networks and deep learning to test the outcome of the segmentation. To apply a method of classification for the chromosomes in metacentric, submetacentric, subtelocentric, and acrocentric for mounting the fish karyotype. And finally, other different metrics to evaluate the quality of segmentation.

Acknowledgment

The authors are grateful to the Bioinformatics and Genomics Laboratory for providing the image dataset. Thanks also to Capes and FAPEMIG. We would like to thanks CNPq for financial support. We would like to thanks CAPES and FAPEMIG (Grant number CEX - APQ-02964-17) for financial support. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

References

[1] S. Minaee, M. Fotouhi, B. H. Khalaj, A Geometric Approach For Fully Automatic Chromosome Segmentation (2011) 1–8arXiv:1112.4164.

[2] N. Madian, K. B. Jayanthi, Overlapped chromosome segmentation and separation of touching chromosome for automated chromosome classification, Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS (2012) 5392–5395.

[3] O. M. Filho, L. A. C. Bertollo., Análise cromossômica de astyanax scabripinnis rivularis (characiformes, characidae) da região Três Marias MG., Cienc. Cult (1986) 35:855.

[4] A. Levan, K. Fredga, A. A. Sandberg, Nomenclature for centromeric position on chromosomes, Hereditas 52 (2) (1964) 201–220.

[5] R. Manohar, J. Gawande, Watershed and clustering based segmentation of chromosome images, in: Advance Computing Conference (IACC), 2017 IEEE 7th International, IEEE, 2017, pp. 697–700.

[6] D. Somasundaram, V. R. Vijay Kumar, Separation of overlapped chromosomes and pairing of similar chromosomes for karyotyping analysis, Measurement: Journal of the International Measurement Confederation 48 (1) (2014) 274–281. doi:10.1016/j.measurement.2013.11.024.

[7] M. V. Munot, M. Joshi, N. Sharma, G. Ahuja, Automated detection of cut-points for disentangling overlapping chromosomes, in: 2013 IEEE Point-of-Care Healthcare Technologies (PHT), 2013, pp. 120–123.

[8] W. Yan, D. Li, Segmentation algorithms of chromosome images, in: Proceedings of 2013 3rd International Conference on Computer Science and Network Technology, 2013, pp. 1026–1029.

[9] A. W. Dougherty, J. You, A Kernel-based adaptive Fuzzy C-Means algorithm for M-FISH image segmentation, 2017 International Joint Conference on Neural Networks (IJCNN) (2017) 198–205.

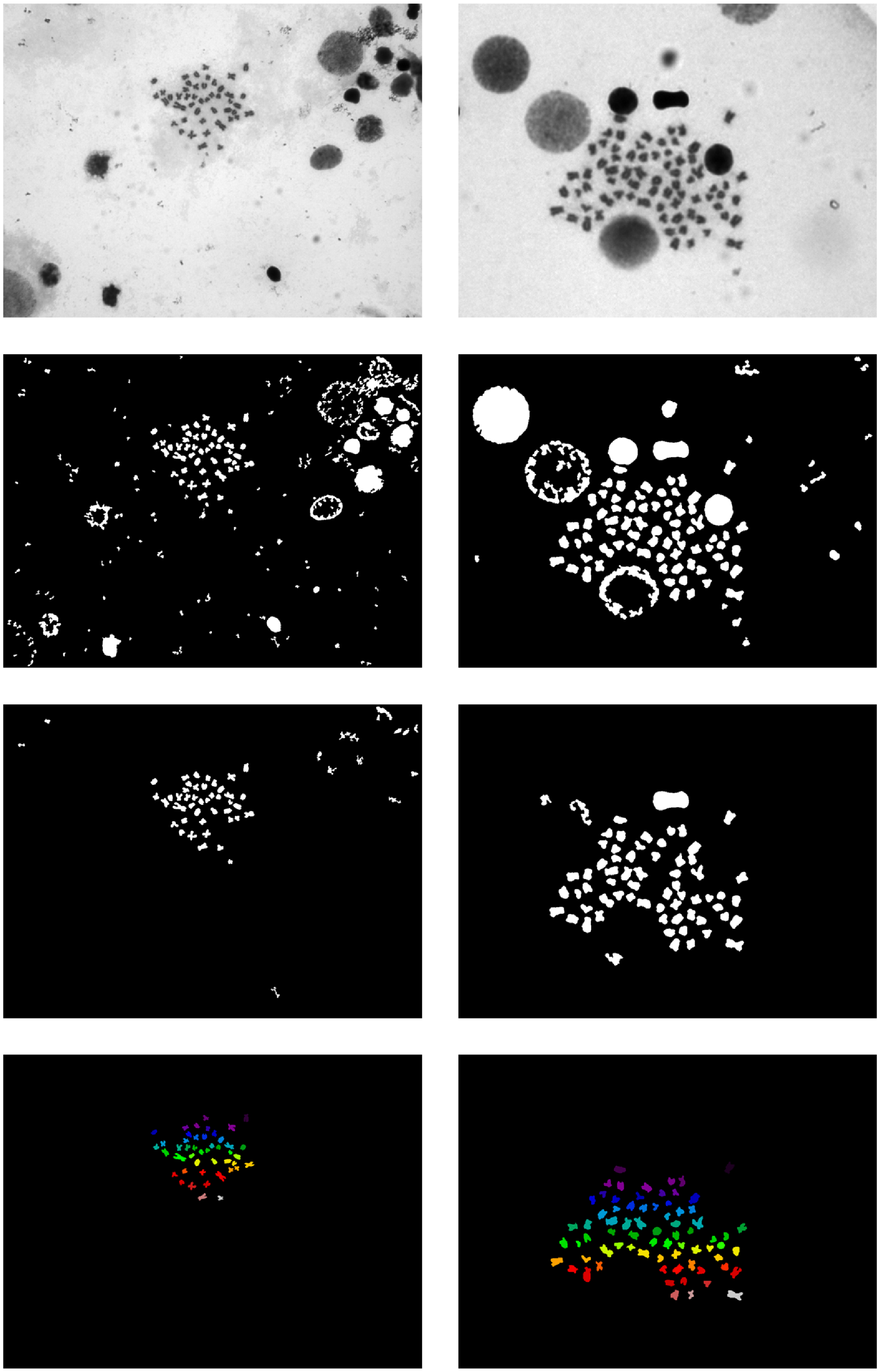


Fig. 2: Segmentation result using NLM filter. The first row shows the original image, the second column the segmented image without classification, the third row shows the final segmented images after classification, and the fourth row shows the reference image.

[10] M. Sharma, O. Saha, A. Sriraman, R. Hebbalaguppe, L. Vig, S. Karande, Crowdsourcing for Chromosome Segmentation and Deep Classification, IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops 2017-July (2017) 786–793. `doi:10.1109/CVPRW.2017.109`.

[11] N. Madian, K. B. Jayanthi, S. Suresh, Contour based segmentation of chromosomes in g-band metaphase images, in: 2015 IEEE Global Conference on Signal and Information Processing (GlobalSIP), 2015, pp. 943–947.

[12] P. Karvelis, A. Likas, D. I. Fotiadis, Identifying touching and overlapping chromosomes using the watershed transform and gradient paths, Pattern Recognition Letters 31 (16) (2010) 2474–2488.

[13] R. J. Rodrigues, W. F. Gonçalves, J. F. Mari, A comparison between two approaches to segment overlapped chromosomes in microscopy images, in: Anais do XIII Workshop de Visão Computacional, 2017, pp. 118–123.

[14] S. Saiyod, P. Wayalun, A hybrid technique for overlapped chromosome segmentation of g-band mataspread images automatic, in: 2014 Fourth International Conference on Digital Information and Communication Technology and its Applications (DICTAP), 2014, pp. 400–404.

[15] R. Gonzalez, R. Woods, Digital Image Processing, Pearson/Prentice Hall, 2008.

[16] A. Buades, B. Coll, J.-M. Morel, A non-local algorithm for image denoising, in: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), Vol. 2, IEEE, 2005, pp. 60–65.

[17] J. V. Manjón, J. Carbonell-Caballero, J. J. Lull, G. García-Martí, L. Martí-Bonmatí, M. Robles, Mri denoising using non-local means, Medical Image Analysis 12 (4) (2008) 514 – 523.

[18] J. Fritsch, T. Kuehnl, A. Geiger, A new performance measure and evaluation benchmark for road detection algorithms, in: 16th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013), IEEE, 2013, pp. 1693–1700.

[19] F. Ge, S. Wang, T. Liu, Image-segmentation evaluation from the perspective of salient object extraction, in: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), Vol. 1, IEEE, 2006, pp. 1146–1153.

[20] L. P. Coelho, A. Shariff, R. F. Murphy, Nuclear segmentation in microscope cell images: A hand-segmented dataset and comparison of algorithms, Proceedings - 2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro, ISBI 2009 (2009) 518–521.