Gabor Features for the Classification and Evaluation of Chromogenic In-Situ Hybridization Images

Stoyan Pavlov[1,3(✉)], Galina Momcheva[2,3], Pavlina Burlakova[2], Simeon Atanasov[2], Dimo Stoyanov[1,3], Martin Ivanov[1,3], and Anton Tonchev[1]

[1] Department of Anatomy and Cell Biology, Faculty of Medicine, Medical University "Prof. Dr Paraskev Stoyanov", Prof. Marin Drinov Str. 55, 9000 Varna, Bulgaria
[2] Department of Computer Science, Varna Free University "Chernorizets Hrabar", Yanko Slavchev str. 84, Chayka resort, 9007 Varna, Bulgaria
[3] Research Group for Advanced Computational Bioimaging, Research Institute of Medical University "Prof. Dr Paraskev Stoyanov", Prof. Marin Drinov Str. 55, 9000 Varna, Bulgaria

Abstract. High-throughput chromogenic in-situ hybridization (CISH) is a bright-field microscopic technique that reveals the spatial distribution of gene expression in animal cells and tissues by an easily detectable coloured precipitate. The "golden standard" for the grading of CISH-stained tissues involves qualitative scoring by a domain expert. This method is biased, suffers from low reproducibility, and lowers the efficiency of high-throughput experiments. A few quantitative image analysis approaches resolve these issues, but the proposed methods are sensitive to experimental conditions or require expert adjustment of multiple parameters. The idea of our research team is to extract textural information from CISH-images that will be used to generate a feature space for semantic segmentation and functional analysis of gene expression. In our current work, we explore the idea by unsupervised classification based on features generated via Gabor energy filters. The tissue was divided into overlapping 150 μm tiles and processed with a Gabor filter bank (5 wavelengths, 16 directions, bandwidth 1.4). The results for the 16 directions at each wavelength were combined by a maximum superposition into a single image, and the mean grey value, standard deviation and entropy were measured. After appropriate dimensionality reduction, the tiles were classified by a fuzzy C-means algorithm. Four experts without prior knowledge of the classification results evaluated the strength and pattern of gene expression of a set of randomly selected tiles, and independently each class in the original whole-slide images. A comparison between the class-scale and tile-scale evaluations was used to assess the usefulness of the selected features.

Keywords: Chromogenic in-situ hybridization · Texture analysis · Machine learning · Feature extraction

S. S. Sotirov et al. (Eds.): BioInfoMed 2020, LNNS 374, pp. 375–383, 2022.
https://doi.org/10.1007/978-3-030-96638-6_39

1 Introduction

The high-throughput chromogenic in-situ hybridization (CISH) is a method to detect known mRNA sequences in the cells of a specific tissue or organ by complementary hybridization with nucleotide probes tagged with enzyme-linked haptens [1]. The hybridization is revealed by an easily detectable coloured precipitate that can be documented using a standard bright-field microscopic imaging system. CISH is primarily used to localize the specific mRNA fragments of the sought gene product and the cell that produces it in fixed tissues [2, 3]. The positive signals correspond to cells that actively produce (aka express) the product of a gene under investigation.

This paper presents a proof of concept for the usability of spatial features extracted through Gabor filtering for the segmentation or the analysis of CISH whole slide images in high-throughput imaging experiments. The paper is organized as follows: In Sect. 2, we explain the rationale and motifs behind the proposed approach. Section 3 discusses the selected texture features. In Sect. 4, we outline the experiment and its results. Section 5 discusses the evaluation by domain experts. In Sect. 6, we conclude with some remarks on further developments.

2 Rationale

2.1 CISH Evaluation

Currently, the gold standard for gene expression grading in CISH stained tissue slides is the assessment by an expert. This approach involves visual inspection of the cell expression and manual labelling (positive or negative) or grading depending on the amount of the observed precipitate. The usual visual scoring is based on cellular gene expression strength ("negative", "low", "moderate", or "strong"), and patterns of expression ("ubiquitous", "regional" or "scattered") [4, 5]. This method is highly biased and suffers from low reproducibility as it strongly depends on the conditions (Fig. 1) and the expertise of the annotator. Furthermore, it is slow and is not particularly effective in high-throughput experiments, when large amounts of image data are generated at high speeds. There are several attempts to design automated workflows for an unbiased evaluation of gene expression. Celldetect is an open-source algorithm for automatic localization and grading of cellular gene expression in whole-slide CISH images [6]. The algorithm performs intensity-based thresholding, classification, and labelling at a reduced spatial resolution (approximately a single cell per pixel). While this approach significantly reduces bias and improves reproducibility, it still suffers from the high variability between images and staining batches inherent to the method. The good reproducibility of Celldetect relies heavily on the human operator to select the proper parameters to account for the intensity variability between images and batches. Another reliable quantitative workflow is developed by Allen Brain Institute and is implemented in the annotation of their Mouse Brain ISH Atlas [7, 8]. The workflow evaluates gene expression in high-resolution images with normalized brightness based on the grey value of the corresponding pixels in combination with advanced filtering to account for different patterns of expression [9, 10]. Both approaches are very efficient but are sensitive to brightness fluctuations, and noise

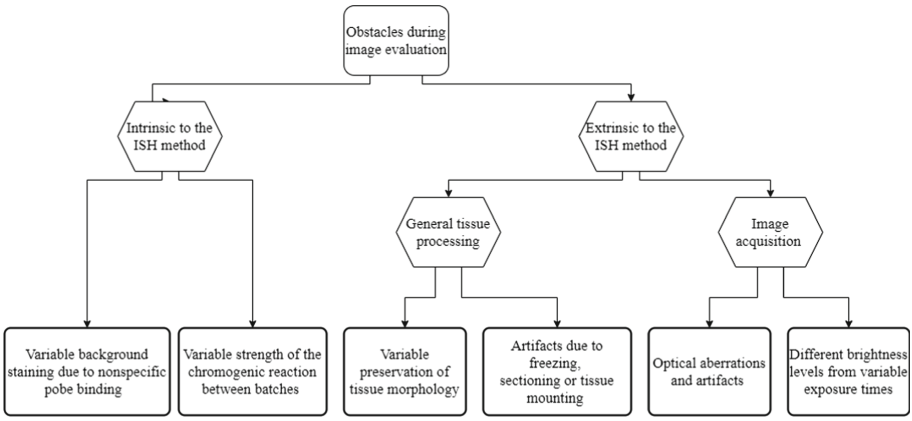

Fig. 1. Most common sources of variability in CISH-stained microscopic images.

introduced by the ISH procedure or the image acquisition and require stringent control of experimental conditions.

A robust and automated workflow can facilitate reproducibility between experiments and between labs and can ensure the acquisition of comparable data from imaging CISH experiments.

During the manual evaluation, experienced annotators can recognize similar levels of gene expression despite significant differences in the overall brightness between images. This process relies on an implicit evaluation of brightness and colour distribution within the locality of the observed tissue. The most prominent property of an image that is related to the local fluctuations of intensity and colour is the texture. Recent research showed that second-order textural features could be used successfully to classify and localize gene expression patterns to specific cerebellar cortical layers [11] or to identify mRNA-enriched sites in the hippocampal region of the brain [12].

2.2 Search for Spatial Features

The texture is a repeating pattern or a function of spatial variation of the brightness intensity of the pixels. Texture analysis plays an essential role in face recognition, surface defect detection, pattern recognition, and medical image analysis by the extraction of meaningful information from digital images. Texture feature extraction refers to the process of computing characteristics of an image, which numerically describes textural image properties and may include scalar numbers, discrete histograms, or empirical distributions [13, 14].

The spatial distribution of brightness variations in this type of images is highly variable and inhomogeneous. That is why a particular standard method for feature extraction can not satisfy all requirements of the analysis of CISH-stained images. The idea of our research team is to use different approaches in the analysis of CISH images that will be used for structural segmentation and functional analysis of gene expression.

As the level of expression is very difficult to quantify in CISH images due to its nonlinear relationship with the amount of precipitated dye, we decided to apply an

inverse approach to the problem. Instead of selecting specific quantitative measures and looking for a fit that can predict the evaluator's grades, we decided first to find features for an image segmentation that mimics the evaluator's decision. The most important features can be then used as an input to supervised classification and segmentation algorithms, and for the derivation of quantitative criteria for the unbiased evaluation and comparison of gene expression levels in CISH images.

3 Gabor Derived Features

The 2D Gabor filter is a sinusoidal wave modulated by a two-dimensional Gaussian envelope with a complex response. The convolution of an image with the spatial Gabor function extracts features with specific spatial frequency, orientation, and phase. The exact nature of the filter's response depends on its wavelength, orientation, phase, and the spread of the Gaussian envelope in the x- and y-direction [15].

The Gabor filter returns an image of the details with the specified spatial frequency and highest response in the direction of the filter. The scale of the filter (i.e., the size of the emphasized details) depends on its wavelength and bandwidth. The bandwidth, in turn, determines the ratio between the wavelength and the standard deviation of the Gaussian envelope. As the spatial selectivity of simple cells in the primate visual cortex has a median bandwidth of 1.4 (which corresponds to a ratio between wavelength and sigma approximately 0.42) [16, 17], we chose this value for the filter. We extracted the local features from Gabor Energy (root sum square of the real and imaginary output). The filter bank includes five wavelengths (2, 4, 8, 16, and 32 px) oriented in 16 directions each. The outputs from all directions at a given scale were combined in one omnidirectional image using maximum superposition. The final output contains all maximal responses for each direction of the filter at the given scale. The features for classification were measured in these multidirectional outputs of the Gabor energy filter. We used three measurements – Mean grey value, Standard Deviation, and Shannon entropy, which results in 15 features in total.

4 Implementation and Results

4.1 Image Filtering and Feature Extraction

The described features were measured in overlapping neighbourhoods with size 150 × 150 μm from the original whole slide images. First, the tissue portion of each image was masked, and only the pixels with coordinates within the tissue mask were included in the scanning process. The images were scanned at full scale (0.5 μm/px) with square neighbourhoods with size 150 μm and horizontal and vertical steps 100 μm. The overlap ensured that objects located at the border between regions would be captured adequately by the algorithm. Each tile was processed separately with the described Gabor filter bank, and fifteen features were extracted. The process was programmed in Python Programming Language (Python Software Foundation, https://www.python.org/) and libraries in the SciPy ecosystem [18–21]. For image navigation, we used the OpenSlide library for visualization of microscopic whole slide images [22].

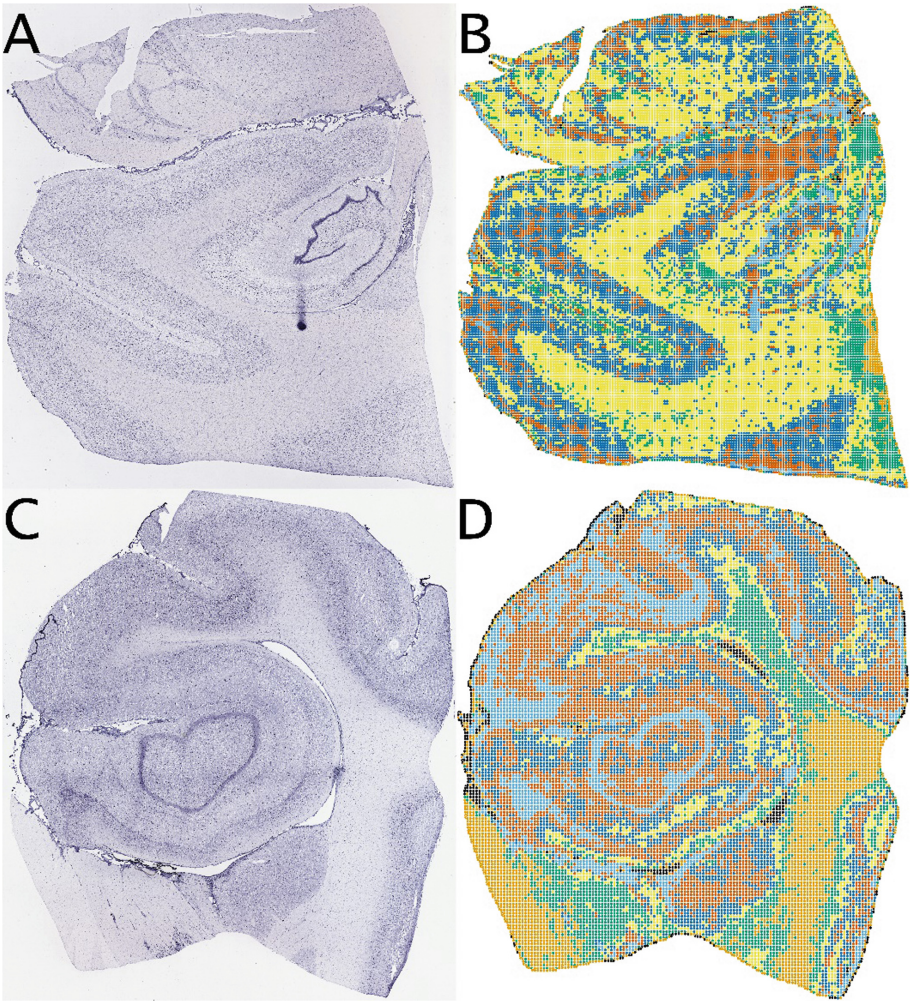

Fig. 2. Result of the Fuzzy C-Means clustering. Original whole slide images (A, C) stained with the CISH-technique for two different gene products and colour-coded spatial maps (B, E) of the clustering based on Gabor energy filter derived features.

4.2 Fuzzy C-means Clustering

To reduce the sources of measurement noise, improve accuracy in clustering, and reduce the computational cost, the dimensionality of the feature matrix was reduced by Principal Component Analysis (PCA) [23]. As the annotator is evaluating the CISH images mainly on two properties - strength and pattern (density) of expression, it seems appropriate to choose the first two principal components for further analyses. Additionally, a higher number of dimensions influences negatively the fuzzy C means algorithm (FCM), which we chose for the unsupervised clustering [24]. Our choice of the fuzzy counterpart of the more popular k-means algorithm was driven mainly by the notion that groups and

categories in living organisms are fuzzy by nature and an approach that captures this fuzziness will be more beneficial. For the selection of a cluster number, we looped the FCM between two and ten cluster centres on the two components and evaluated the results with the fuzzy partition coefficient [25]. While the FPC reaches a maximum value at two clusters, we chose seven clusters. This decision is once again dictated by the nature of the processed image data: the standard protocol for evaluation includes at least six separate groupings, and this number of classes allows us to capture gradual differences in gene expression while keeping the FPC at acceptable values. The fuzzy clustering of the reduced data produced classes that, on inspection, separated the images into regions with different expression levels (Fig. 2). Even for the unprepared observer, the contrast distribution in the unprocessed images corresponds to the cluster mapping.

5 Expert Evaluation

The original images and the colour-coded classifications were presented to four different experts for evaluation of the clusters as a whole. Each evaluator graded the expression strength (0–3) and the density (0–2) of expressing cells in each class (Table 1).

Since the evaluators had a different experience, the final grades were calculated as a weighted mean of the individual grades rounded to an integer score. The grades of evaluators one and two received weight 1.0, evaluator 3 grades had weight 2.0, and the most experienced grader received weight 3. Finally, we compared the tile-scale evaluations to the class-scale evaluations.

In the next stage, a set of 137 randomly selected tiles were evaluated by the same group of experts using the same criteria without knowledge about how the Fuzzy C Means algorithm classified the corresponding tile. In this way, the evaluators assigned strength and density values to selected tiles at the same scale (150 μm x 150 μm) as the algorithm performed its measurements and classification.

Table 1. Expert evaluation of the expression strength (**Str**) and density (**Den**) of the classes. Final scores are calculated as a weighted mean of the scores from each evaluator rounded to an integer (weights are given in parentheses after the evaluator's designation). For comparison with the tile-scale evaluations, we provide the median tile scores per class of 137 randomly selected tiles graded similarly.

Class	Evaluator 1 (weight = 1)		Evaluator 2 (weight = 1)		Evaluator 3 (weight = 2)		Evaluator 4 (weight = 3)		Weighted grade		Median tile score	
	Str	Den	Str	Den	Str	Den	Str	Den	Str	Den	Str	Den
Class 0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	**0**	**0**	**0**	**0**
Class 1	0.0	0.0	1.0	1.0	0.0	1.0	1.0	1.0	**1**	**1**	**1**	**1**

(*continued*)

Table 1. (*continued*)

Class	Evaluator 1 (weight = 1)		Evaluator 2 (weight = 1)		Evaluator 3 (weight = 2)		Evaluator 4 (weight = 3)		Weighted grade		Median tile score	
	Str	Den	Str	Den	Str	Den	Str	Den	Str	Den	Str	Den
Class 2	3.0	2.0	3.0	2.0	3.0	2.0	3.0	2.0	**3**	**2**	**3**	**2**
Class 3	1.0	1.0	3.0	1.0	1.0	1.0	1.0	1.0	**1**	**1**	**2**	**1**
Class 4	1.0	1.0	1.0	1.0	2.0	1.0	2.0	2.0	**2**	**1**	**2**	**1**
Class 5	1.0	1.0	3.0	2.0	2.0	2.0	2.0	2.0	**2**	**2**	**2**	**1**
Class 6	3.0	2.0	3.0	2.0	3.0	1.0	2.0	2.0	**3**	**2**	**3**	**2**

As we assumed by design that the clusters defined by the described algorithm must include tiles of similar expression level and density, we analyzed the agreement between the tile-scale scores and the corresponding class scores in a confusion matrix for the 137 randomly selected tiles (Table 2). While the median of the individual tile scores per class corresponds very well to the class grades (Table 1), the overall accuracy of the prediction of the tile's score from its class score is low. One explanation for this is that the evaluators looking at the individual tiles are assessing the expression strength and density with details at a smaller scale.

Table 2. Agreement between expert evaluations of the individual tiles and their corresponding classes assigned by the algorithm. The class grades were used as the conditional "ground truth".

Property	Levels	Precision	Recall	F1-score	N	Accuracy	Cohen's kappa
Expression strength	No expression	0.92	0.44	0.6	27	–	–
	Low	0.38	0.29	0.33	35	–	–
	Moderate	0.33	0.58	0.42	38	–	–
	High	0.65	0.54	0.59	37	–	–
	Overall	0.55	0.47	0.48	137	0.47	0.28
Density of expressing cells	No expression	1	0.45	0.62	29	–	–
	Sparse	0.58	0.55	0.56	64	–	–
	Dense	0.47	0.68	0.56	44	–	–
	Overall	0.68	0.56	0.58	137	0.57	0.31

Additionally, there are lots of "misclassifications" between neighbouring levels of the properties that reduce the calculated accuracy. Another reason is the presence of artefactual staining in some of the randomly selected tiles. Those tiles were accurately evaluated as nonexpressing by the human raters. The clustering algorithm, however, classifies the tiles based on the distribution of intensities in the image and does not discriminate between artefacts and valid staining.

6 Conclusion

We presented the use of a set of features derived from Gabor filtered images of CISH-stained tissues. The selected features allowed the clustering of the image regions into classes with generally good correspondence to the different expression patterns as described by expert evaluators in CISH images. The accuracy of prediction of the tile-scale evaluation by the class-scale grade was low, but it must be noted that this approach is not intended to (and does not) measure the ability of the algorithm to correctly discriminate between the different levels of expression strength and density of the expressing cells. The low accuracy stems mainly from the different scales at which the evaluators asses the randomly selected tiles without knowledge of their neighborhood and the classes in the whole image. As the median scores of the tiles in each class were in agreement with the overall class score assigned by the human raters, we may conclude that the proposed features have discriminating power similar to the human evaluation. Next step should be to analyze and select the most important of these features and combine them with other approaches into a feature space which can be used for efficient training of supervised segmentation algorithms and derivation of unbiased indicators for the quantitative comparison of gene expression patterns in CISH-stained tissues. An essential issue in the analysis of CISH-images is the distinction and removal of nonspecific staining. This problem can be addressed by the implementation of a similar carefully constructed feature space after differential analysis of positive and negative control images.

References

1. Corthell, J.T.: In situ hybridization. In: Basic Molecular Protocols in Neuroscience: Tips, Tricks, and Pitfalls, pp. 105–111. Elsevier (2014)
2. Jensen, E.: Technical review: *in situ* hybridization: AR insights. Anat. Rec. **297**, 1349–1353 (2014). https://doi.org/10.1002/ar.22944
3. McFadden, G.I.: In situ hybridization, chapter 12. In: Methods in Cell Biology, pp. 165–183. Elsevier (1995)
4. Eichele, G., Diez-Roux, G.: High-throughput analysis of gene expression on tissue sections by in situ hybridization. Methods **53**, 417–423 (2011). https://doi.org/10.1016/j.ymeth.2010.12.020
5. Visel, A., Thaller, C., Eichele, G.: GenePaint.org: an atlas of gene expression patterns in the mouse embryo. Nucleic Acids Res. **32**, D552–D556 (2004). https://doi.org/10.1093/nar/gkh029
6. Carson, J.P., Eichele, G., Chiu, W.: A method for automated detection of gene expression required for the establishment of a digital transcriptome-wide gene expression atlas. J. Microsc. **217**, 275–281 (2005). https://doi.org/10.1111/j.1365-2818.2005.01450.x

7. ISH Data: Allen Brain Atlas: Mouse Brain. https://mouse.brain-map.org/. Accessed 7 July 2020
8. Ng, L., Pathak, S.D., Kuan, C., et al.: Neuroinformatics for genome-wide 3D gene expression mapping in the mouse brain. IEEE/ACM Trans. Comput. Biol. Bioinform. **4**, 382–393 (2007). https://doi.org/10.1109/tcbb.2007.1035
9. Informatics Data Processing in: Documentation - Allen Mouse Brain Atlas. http://help.brain-map.org/download/attachments/2818169/InformaticsDataProcessing.pdf?version=1&modificationDate=1319667590884&api=v2. Accessed 7 July 2020
10. Lein, E.S., Hawrylycz, M.J., Ao, N., et al.: Genome-wide atlas of gene expression in the adult mouse brain. Nature **445**, 168–176 (2007). https://doi.org/10.1038/nature05453
11. Kirsch, L., Liscovitch, N., Chechik, G.: Localizing genes to cerebellar layers by classifying ISH images. PLoS Comput. Biol. **8**, e1002790 (2012). https://doi.org/10.1371/journal.pcbi.1002790
12. Ugolotti, R., Mesejo, P., Zongaro, S., et al.: Visual search of neuropil-enriched RNAs from brain in situ hybridization data through the image analysis pipeline hippo-ATESC. PLoS ONE **8**, e74481 (2013). https://doi.org/10.1371/journal.pone.0074481
13. Armi, L., Fekri-Ershad, S.: Texture image analysis and texture classification methods - a review. arXiv:190406554 (Cs) (2019)
14. Materka, A.: Texture analysis methodologies for magnetic resonance imaging. Dialogues Clin. Neurosci. **6**, 243–250 (2004)
15. Bianconi, F., Fernández, A.: Evaluation of the effects of Gabor filter parameters on texture classification. Pattern Recognit. **40**, 3325–3335 (2007). https://doi.org/10.1016/j.patcog.2007.04.023
16. De Valois, R.L., Albrecht, D.G., Thorell, L.G.: Spatial frequency selectivity of cells in macaque visual cortex. Vis. Res. **22**, 545–559 (1982). https://doi.org/10.1016/0042-6989(82)90113-4
17. Petkov, N., Kruizinga, P.: Computational models of visual neurons specialised in the detection of periodic and aperiodic oriented visual stimuli: bar and grating cells. Biol. Cybern. **76**, 83–96 (1997). https://doi.org/10.1007/s004220050323
18. Hunter, J.D.: Matplotlib: a 2D graphics environment. Comput. Sci. Eng. **9**, 90–95 (2007). https://doi.org/10.1109/MCSE.2007.55
19. Pedregosa, F., Varoquaux, G., Gramfort, A., et al.: Scikit-learn: machine learning in python. J. Mach. Learn. Res **12**, 2825–2830 (2011)
20. SciPy 1.0 Contributors, Virtanen, P., Gommers, R., et al.: SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat. Methods **17**, 261–272 (2020). https://doi.org/10.1038/s41592-019-0686-2
21. van der Walt, S., Colbert, S.C., Varoquaux, G.: The NumPy array: a structure for efficient numerical computation. Comput. Sci. Eng. **13**, 22–30 (2011). https://doi.org/10.1109/MCSE.2011.37
22. Goode, A., Gilbert, B., Harkes, J., et al.: OpenSlide: a vendor-neutral software foundation for digital pathology. J. Pathol. Inform. **4**, 27 (2013). https://doi.org/10.4103/2153-3539.119005
23. Han, J., Kamber, M.: Data Mining: Concepts and Techniques. Elsevier, Burlington (2012)
24. Winkler, R., Klawonn, F., Kruse, R.: Fuzzy c-means in high dimensional spaces. Int. J. Fuzzy Syst. Appl. **1**, 1–16 (2011). https://doi.org/10.4018/ijfsa.2011010101
25. Kim, D.-W., Lee, K.H., Lee, D.: Fuzzy cluster validation index based on inter-cluster proximity. Pattern Recogn. Lett. **24**, 2561–2574 (2003). https://doi.org/10.1016/S0167-8655(03)00101-6