Journal of Microscopy

Journal of Microscopy, Vol. 0, Issue 00 2014, pp. 1–21 doi: 10.1111/jmi.12186
Received 28 March 2014; accepted 25 August 2014

Performance evaluation of image segmentation algorithms on microscopic image data

MIROSLAV BENEŠ*,† & BARBARA ZITOVÁ*
*Institute of Information Theory and Automation, Academy of Sciences of the Czech Republic, Prague, Czech Republic
†Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic

Key words. Image analysis, image segmentation, microscopic images, performance evaluation.

Summary

In our paper, we present a performance evaluation of image segmentation algorithms on microscopic image data. In spite of the existence of many algorithms for image data partitioning, there is no universal and 'the best' method yet. Moreover, images of microscopic samples can be of various character and quality which can negatively influence the performance of image segmentation algorithms. Thus, the issue of selecting suitable method for a given set of image data is of big interest. We carried out a large number of experiments with a variety of segmentation methods to evaluate the behaviour of individual approaches on the testing set of microscopic images (cross-section images taken in three different modalities from the field of art restoration). The segmentation results were assessed by several indices used for measuring the output quality of image segmentation algorithms. In the end, the benefit of segmentation combination approach is studied and applicability of achieved results on another representatives of microscopic data category – biological samples – is shown.

Lay description

The image segmentation is one of several parts of image analysis process. Its role is to partition an image to meaningful nonoverlapping regions – segments – which serve as an input to following stages of the analysis. There is plenty of different approaches addressing this issue and although many of them deliver high-quality results, there is no universal segmentation method which would outperform the others on any kind of data. Thus, there is always a dilemma which method to choose for segmentation of given data set. In our paper we compare performance of many segmentation methods on data set of microscopic images. We give suggestions on which method to use under which circumstances and we show that combination of several methods outperforms even the best one from the studied set. The findings are supported by large number of experiments and statistical testing.

Correspondence to: Miroslav Beneš, Institute of Information Theory and Automation, Academy of Sciences of the Czech Republic, Pod Vodárenskou věží 4, Prague, Czech Republic. Tel: (+420)266052864; fax: (+420)286890378; e-mail: benem9am@utia.cas.cz

Introduction

The fundamental objective of image segmentation is to partition the input image into meaningful nonoverlapping regions – segments – for further analysis or visualization. There is a variety of approaches addressing this task, exploiting various image properties to achieve the given goal. They span from low-level techniques using intensity thresholds, edge tracing or region growing (RG), over graph-based and statistical approaches, to model-based algorithms and other higher level methods (see e.g. Pal & Pal, 1993 or Dey *et al.*, 2010 for surveys, the latter from optical remote sensing perspective). Survey (Freixenet *et al.*, 2002) presents also quantitative comparison next to the review of segmentation techniques which integrate boundary and region information. Recently, the combination-based solution has been introduced, where the final partition is formed using a combination of results of several segmentation methods and thus inhibiting their shortcomings.

Despite the longtime effort to develop high-quality segmentation algorithms, there has not been any universal segmentation method proposed. Under these circumstances, there is a dilemma which method to choose for given particular data set and whether the combination of segmentation results would be beneficial. Our article tries to answer these questions for defined category of image processing data set of images of microscopic samples (see Fig. 1), moreover taken in different modalities (visible spectrum (VIS), ultraviolet spectrum (UV) and scanning electron microscope (SEM)). From the image processing point of view, the origin of the samples often does not play an important role. The factual meaning of particular intensity levels can be irrelevant for the segmentation algorithm.

We limit our study to the microscopic image data that contain the sample located in the inner part of the image, mostly

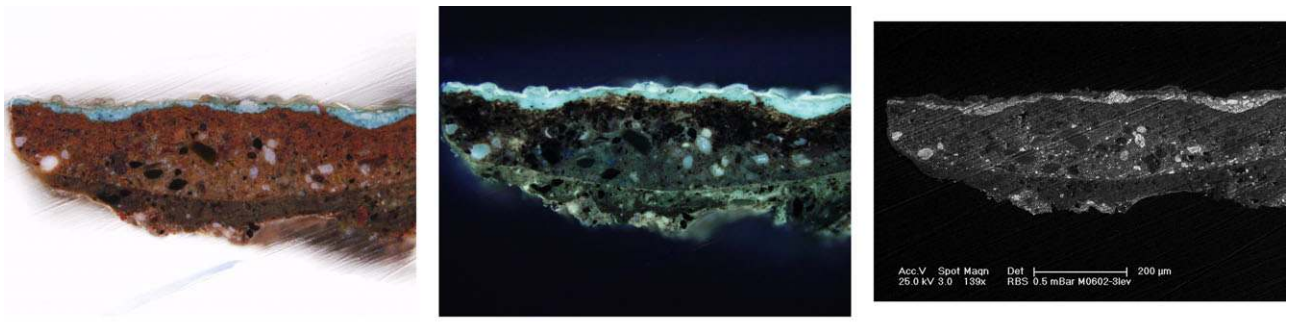

Fig. 1. The images of the cross-section samples are acquired in three modalities – visible spectrum (VIS), ultraviolet spectrum (UV) and scanning electron microscope (SEM). Image courtesy of ALMA, Prague.

not reaching to the top and bottom image borders. The data may come from an analysis of painting materials used in art restoration (Fig. 1), which is the case of the data set used in our evaluation. They can be samples of various biological materials, such as tissues, cells, or other biological structures. The task at hand can be seen as the two-target problem where an image has to be labelled with either foreground or background label and where the foreground is usually the inner part of the image and the background is separated and/or removed. The problem can be viewed as image binarization, too.

At first glance it might seem to be a simple task solvable by means of basic thresholding, however the situation is often more complex. Due to the setting of data collection process, acquired images are often unfit to the chosen segmentation method and following complications are usually inevitable – surroundings of analysed samples can be semitransparent, with nonuniform cutting-plane and various debris, to name a few examples. High number of samples can negatively influence precision of sample scanning in terms of noise level and blurring.

The objective of the paper is to evaluate the noninteractive segmentation methods in terms of their accuracy, assessed by several indices used for measuring the output quality of image segmentation algorithms. Furthermore, efficiency of combination of segmentation results is addressed, too. Finally, the applicability of the achieved conclusions is demonstrated on different data set – the biological samples. Sections segmentation algorithms and quality indices introduces the participating methods and indices. The full explanation of the analysed methods is out of the scope of our paper. If necessary, please consult given references. Section algorithms evaluation forms the key part of this paper with evaluation and comparison of the image segmentation algorithms. Insight into their performance and guidelines for their use are given there. Also application of the obtained results to different data set is shown. Section combination of image segmentation methods presents exploitation of the results for achieving even better segmentation output via combination approach. The paper is concluded in conclusion Section.

Segmentation algorithms and quality indices

First, a survey of the image segmentation algorithms analysed in this paper (i.e. studied set) is presented. The second part focuses on indices used for measuring the output quality of the image segmentation algorithms. The abbreviations are assigned to each method and index for future references and their list is presented in Table 1.

Segmentation algorithms

There is a variety of segmentation methods available to be used to solve the image segmentation problem which differ in many ways (see e.g. Pal & Pal, 1993, Dey *et al.*, 2010 or Freixenet *et al.*, 2002 for surveys). The algorithms in our study are selected with respect to the following criteria. Methods with different fundamentals are considered to provide a diversity. The performance and computational (time) efficiency are taken into account with preference for short execution time. Finally, the public availability of the implementation and thus related popularity of the segmentation method are considered too. Last criterion is also important because it can be expected that potential users of image segmentation algorithms would choose exactly such popular methods. There exists a lot more segmentation algorithms (e.g. Grady, 2006; Malcolm *et al.*, 2007; Arbelaez *et al.*, 2011) but inclusion of each of them is beyond the scope of this paper.

The selected algorithms can be divided into groups according to their fundamental approach to solve the image segmentation problem. The following paragraphs briefly describe the groups and particular algorithms.

Thresholding. Thresholding is probably the most popular method for image segmentation. The aim is to find an optimum threshold which separates the input image to two distinct groups of pixels by their intensity. Plenty of different methods for threshold detection exist and many of them are selected to participate in the evaluation.

Table 1. List of image segmentation methods in studied set and of quality indices used for their comparison. The abbreviations widely used in text are in the first column.

Segmentation methods	
IMJ_*	Various thresholding methods from ImageJ (Huang & Wang, 1995; Prewitt & Mendelsohn, 1966; Ridler & Calvard, 1978; Li & Tam, 1998; Kapur *et al.*, 1985; Kittler & Illingworth, 1986; Tsai, 1985; Otsu, 1975; Doyle, 1962; Shanbhag, 1994; Zack *et al.*, 1977; Yen *et al.*, 1995)
HT_*	Various thresholding methods from HistThresh (Rosenfeld & De La Torre, 1983; Glasbey, 1993; Otsu, 1975; Ridler & Calvard, 1978; Prewitt & Mendelsohn, 1966; Dempster *et al.*, 1977; Kittler & Illingworth, 1986; Tsai, 1985)
TNC	Tao's thresholding method (Tao *et al.*, 2008)
RG	Region growing (Pratt, 2007)
KM	K-means clustering (MacQueen, 1967)
MS	Mean Shift algorithm (Comaniciu & Meer, 2002)
GC_FH	Felzenszwalb's method (Felzenszwalb & Huttenlocher, 2004)
GC_R	GrabCut (Rother *et al.*, 2004)
GC_CV	Daněk's optimization of Chan-Vese (Daněk, 2012; Chan & Vese, 2001)
GC_RD	Daněk's optimization of Rousson–Deriche (Daněk, 2012; Rousson & Deriche, 2002)
MNC	Multiscale normalized cut (Cour *et al.*, 2005)
Quality indices	
HD	Hamming distance (Hamming, 1950)
BHD	Boundary Hamming distance (Kohli *et al.*, 2009)
RI	Rand index (Rand, 1971)
ARI	Adjusted Rand index (Hubert & Arabie, 1985)
DC	Dice coefficient (Dice, 1945)
FMI	Fowlkes–Mallows index (Fowlkes & Mallows, 1983)
NMI	Normalized mutual information (Strehl & Ghosh, 2003)
VI	Variation of information (Meilă, 2007)
HAUSD	Hausdorff distance (Sluimer *et al.*, 2005)
MASD	Mean absolute surface distance (Sluimer *et al.*, 2005)

The methods of the Auto Threshold plugin[1] for ImageJ software package[2] are included. Namely Huang method (IMJ_HUANG) (Huang & Wang, 1995) which minimizes the measures of background/foreground fuzziness, Intermodes (IMJ_IM) (Prewitt & Mendelsohn, 1966) with iterative histogram smoothing, Isodata (IMJ_ISO) (Ridler & Calvard, 1978) and its variation (IMJ_DEF) which iteratively update the threshold according to background and foreground intensity means, Li's method (IMJ_LI) (Li & Tam, 1998) for cross entropy minimization, Kapur–Sahoo–Wong maximum entropy method (IMJ_MAXENT) (Kapur *et al.*, 1985), mean of the grey levels as threshold (IMJ_MEAN), iterative version of minimum error thresholding (IMJ_IME) (Kittler & Illingworth, 1986), minimum method (IMJ_MIN) (Prewitt & Mendelsohn, 1966), moment-preserving method (IMJ_MOM) (Tsai, 1985), Otsu's method (IMJ_OTSU) (Otsu, 1975) for minimizing the intraclass variance, percentile method (IMJ_PER) (Doyle, 1962), method using Renyi's entropy (IMJ_RENYI) (Kapur *et al.*, 1985), Shanbhag's extension (IMJ_SB) (Shanbhag, 1994) to Kapur's maximum entropy method, geometric Triangle algorithm (IMJ_TRIANGLE) (Zack *et al.*, 1977) and Yen's method (IMJ_YEN) (Yen *et al.*, 1995) based on a maximum correlation criterion.

In addition to the plugin several other thresholding methods from MATLAB HistThresh toolbox[3] are studied[4] – concavity method by Rosenfeld (HT_CONCAV) (Rosenfeld & De La Torre, 1983), Glasbey's entropy method (HT_ENT) (Glasbey, 1993), maximum likelihood via EM algorithm (HT_MAXLIK) (Dempster *et al.*, 1977), Intermeans (HT_INTER) as equivalent to Otsu's method and its iterative version (HT_INTERI) which is equivalent to IsoData method mentioned above. Then there is median method (HT_MEDIAN) (Glasbey, 1993) which assumes that half of the pixels belong to the background and other half to the foreground, and noniterative minimum error thresholding (HT_ME) (Kittler & Illingworth, 1986).

[1] http://fiji.sc/Auto_Threshold
[2] http://rsbweb.nih.gov/ij/

[3] http://www.cs.tut.fi/~ant/histthresh/
[4] There are more thresholding methods in the toolbox. Most of them are the same as in ImageJ plugin. However, we found out that their implementation often slightly differed and so did the results of the segmentation. For this reason all methods are included in the studied set with corresponding suffices in their abbreviations (so there are, for example, both IMJ_MEAN and HT_MEAN in the studied set).

Finally, a Tao's method for image thresholding (TNC) (Tao *et al.*, 2008), which uses a normalized graph-cut to detect an optimum threshold, is included in the evaluation below.

Region growing. The RG (Pratt, 2007) is another common segmentation approach included in our selection. The algorithm partitions the input image to segmented regions by growing from the seed points (picked automatically or by the user) to the neighbouring pixels depending on a membership criterion such as intensity or texture similarity.

Clustering methods. The goal of clustering methods is to group the input objects by their similarity or dissimilarity with respect to a given criterion such as colour, spatial coordinates etc. K-means clustering and Mean Shift (MS) algorithm are selected representatives of this approach.

K-means clustering (MacQueen, 1967) assigns the input objects to the clusters with the nearest means which are iteratively updated. The method strongly depends on the initialization and favours final clusters/segments of similar spatial extent. The MS algorithm represents more complex approach. Comaniciu and Meer (Comaniciu & Meer, 2002) exploited the nonparametric MS procedure for detecting multiple modes in a feature space in order to delineate the final clusters in such space.

Graph-based algorithms. Graph-based image segmentation algorithms generally model the image as a graph in which the nodes represent the pixels and the edges of the graph correspond to some relation between pixels (usually their similarity or dissimilarity). A graph partitioning method is then used to obtain final partition and by doing so also the final segmentation of the input image.

In their paper Felzenszwalb & Huttenlocher (2004, GC_FH) introduced the efficient greedy algorithm for partitioning an image graph to obtain a final segmentation that is not too coarse or too fine given a dissimilarity predicate. GrabCut algorithm by Rother *et al.* (2004, GC_R) uses graph cut optimization technique (min-cut/max-flow algorithm) to minimize energy function derived from an input image using intensity values.[5] The OpenCV[6] implementation of this algorithm is examined. The graph cut minimization (Daněk, 2012) of both Chan–Vese active contour model for image segmentation (GC_CV) (Chan & Vese, 2001) and Rousson–Deriche Bayesian model (GC_RD) (Rousson & Deriche, 2002) is included. A multiscale version of normalized cut graph partitioning framework (MNC) (Cour *et al.*, 2005) is considered too. The multiscale adjustment added to the original algorithm by Shi & Malik (2000) allows to segment large images thanks to its computational efficiency.

[5] Although GrabCut is user interactive algorithm, its initialization can be done automatically with no effort (see Section of the input data set and evaluation setup). Interactivity is thus no handicap.

[6] http://www.opencv.org

Quality indices

Quality indices form the second important part of the evaluation. In order to objectively evaluate the performance of the image segmentation methods and quality of their results, the quality indices (or measures) are necessary to adopt. The pursuit of objectivity is motivated by an effort to suppress the subjective (and still often empirical) evaluation of the segmentation algorithms in the original papers.

There exist two main approaches to design an objective measure – *unsupervised evaluation* and *supervised evaluation*. The unsupervised quality indices do not require comparison with any additional reference standard and their evaluation is solely based on a given segmented image. These indices usually exploit such criteria as intraregion homogeneity, interregion difference etc. For a survey of unsupervised evaluation methods, see Zhang *et al.* (2008). Conversely the supervised performance evaluation approach requires the ground truth reference image (GT) which the actual segmented image is compared to. The GT image is often obtained manually by experts and reflects the optimum of the resulting segmentation. In our case the supervised evaluation is more appropriate because of the better ability to distinguish the slight disparities between the results of various segmentation algorithms thanks to the comparison with this ideal GT.

The following sections present quality indices used in this paper. They are selected mainly to keep the diversity of the final set. On top of that they are widely used in relevant papers. Each index usually favours certain properties of the segmentation results and penalizes others (they are biased in this sense). Therefore, it is important to incorporate larger set of indices and handle their possibly different evaluation of given segmented image in order to keep the evaluation objective as much as possible. Only one or two indices would be insufficient and would probably distort the results.

It is worth mentioning that there exist more quality indices than are described in this paper. Nevertheless a lot of them are equivalent to the ones selected, like F-measure (Rijsbergen, 1979), Jaccard index (Jaccard, 1912) or Classification accuracy used, for example, in Kuncheva *et al.* (2006). Some are inappropriate for the task, for example, LCE and GCE (Martin *et al.*, 2001), which try to deal with refinements in context of multilabel segmentation. We assume that the indices are correct, that is, their values are meaningful and not random. The theoretical range of values is specified for each index.[7] In formulas I denotes segmented image for which the quality index is computed, GT is the corresponding ground truth, F and B subscripts denote foreground and background, respectively.

Hamming distance. Hamming distance (HD) is well-known metric from the information theory (Hamming, 1950). Originally it counts differences between two strings. In image processing it can be used to count the number of misclassified or

[7] Extremities of the range do not necessarily have to be reached in practice.

missegmented pixels. The distance is normalized with the total number of pixels and therefore the range is in the interval of 0 to 1, where 0 is for absolute mismatch and 1 for equality to the GT.

$$\mathrm{HD} = 1 - \frac{|I_B \cap GT_F| + |I_F \cap GT_B|}{|I|}.$$

Huang & Dom (1995) introduced a variation called normalized HD, which can deal with multilabel and not only with binary segmentation. However, in binary case Huang's normalized version is equivalent to plain HD.[8]

Boundary Hamming distance. Boundary Hamming distance (BHD) introduced in Kohli *et al.* (2009) is the variation of HD that stresses the accuracy of the segmentation result on an object's boundary. Kohli *et al.* argue that the ordinary HD is not appropriate if the user is interested more in accurate object boundary (and so in the accurate segmentation as well), because a large qualitative improvement on the object border results in only a negligible increase of the performance measure. The quality in boundary version is then evaluated by counting the number of missegmented pixels in the region surrounding the object boundary with the specified width. As with the previous case, the distance is normalized and range is between 0 and 1.

$$\mathrm{BHD} = 1 - \frac{|I_B \cap \mathrm{GT_F}|_{\mathrm{BOUNDARY}} + |I_F \cap GT_B|_{\mathrm{BOUNDARY}}}{|\mathrm{BOUNDARY}|}.$$

In our case it makes sense to include both the HD and its boundary version, because even though we are interested in fine object boundary in the resulting image the complete missegmentation might happen and such case is better reflected (and penalized) by common HD.

Rand index and adjusted Rand index. Rand index (RI) (Rand, 1971) and adjusted Rand index (ARI) (Hubert & Arabie, 1985) are quality indices originally developed for comparing the clusterings. They are based on counting pairs of objects which two clusterings agree or disagree on (which leads to what is often called contingency table or confusion matrix). In the same manner they can compare segmentation results to the GT.

$$m_{ij} = |I_i \cap GT_j|,\ i,\ j \in \{F,\ B\},$$

$$m = \sum_{i,\,j \in \{F,\,B\}} m_{ij} \qquad m_{i+} = \sum_{j \in \{F,\,B\}} m_{ij} \qquad m_{+j} = \sum_{i \in \{F,\,B\}} m_{ij},$$

$$T = \frac{1}{2}\left[\sum_{i,\,j \in \{F,\,B\}} m_{ij}^2 - m\right],$$

$$P = \sum_{i \in \{F,\,B\}} \binom{m_{i+}}{2}, \qquad Q = \sum_{j \in \{F,\,B\}} \binom{m_{+j}}{2}, \qquad N = \binom{m}{2},$$

$$\mathrm{RI} = \frac{N + 2T - P - Q}{N}.$$

The adjusted Rand index corrects the original RI for chance agreement between two clusterings by normalizing RI with its expected value. The range of RI (values between 0 and 1, where 0 is for absolute noncompliance with GT) is thus corrected to the interval of -1 and 1. It is questionable if this correction stays practical in the area of image segmentation where assumptions do not have to hold, but experimental results (Vinh *et al.*, 2009) show that it is worth considering.

$$\mathrm{ARI} = \frac{2(NT - PQ)}{N(P + Q) - 2PQ}.$$

The RI and ARI are also in some sense equivalent to other well-known criteria like Cohen's Kappa statistic (Cohen, 1960; Warrens, 2008) or Mirkin's metric (Mirkin, 1996), which is another adjusted form of RI (Meilă, 2007).

Dice coefficient. Dice coefficient (DC) (Dice, 1945) is popular quality index for evaluating the results of image segmentation, especially in the medical imaging domain. Its range is again from 0 to 1 (1 for perfect match with GT).

$$\mathrm{DC} = \frac{2|I_F \cap GT_F|}{|I_F| + |GT_F|}.$$

Other indices are equivalent to Dice coefficient, for example, Jaccard index (Jaccard, 1912) and in binary case the popular F-measure (Rijsbergen, 1979).

Fowlkes–Mallows index. Fowlkes–Mallows index (FMI) (Fowlkes & Mallows, 1983) is another index based on the contingency table. It has different properties than both RI and ARI mentioned earlier. It handles the independent clusterings in a better way and behaves stably in the presence of noise (see the original paper). As with the RI the range of this index is between 0 and 1. The smaller the degree of missegmentation is, the closer the index is to 1.

$$W_1 = \frac{T}{\sum_{i \in \{F,\,B\}} |I_i|(|I_i| - 1)/2},$$

$$W_2 = \frac{T}{\sum_{j \in \{F,\,B\}} |GT_j|(|GT_j| - 1)/2},$$

$$\mathrm{FMI} = \sqrt{W_1\, W_2}.$$

Normalized mutual information. Mutual information is information theoretic index which measures the amount of mutually shared information between two random variables (i.e. partitions or segmented images in our case). The more the segmented result resembles the GT, the more information is shared. Since the mutual information has no argument-independent upper bound, Strehl & Ghosh (2003) normalized it using the geometric mean of the entropies. The normalized

[8] Except for the matching problem between segmented regions. See the paper Huang & Dom (1995) for details.

version (NMI) thus ranges from 0 to 1 with 1 for equality to the GT.

$$\mathrm{NMI} = \frac{MI(I, GT)}{\sqrt{H(I)H(GT)}},$$

where *MI(I,GT)* denotes the mutual information between I and GT, and $H(I)$ denotes the entropy of I.

Variation of information. The variation of information (VI) (Meilă, 2007) is distance metric derived from the mutual information. Contrary to the mutual information it measures the amount of information (or entropy) which is not shared between two random variables. It would seem that VI is only a complement of NMI and their results would be equivalent. Comparison of the results however shows that they may differ, so both indices are used in evaluation. The nonnormalized version of VI is used with values 0 for absolute match to the GT and positive values for the opposite.

$$\mathrm{VI} = H(I) + H(GT) - 2MI(I, GT).$$

Hausdorff distance and mean absolute surface distance. Two last indices take the boundary of the segmented foreground into account. Hausdorff distance (HAUSD) measures the largest minimal distance between two boundaries. Mean absolute surface distance (MASD) measures the average minimal distance between two boundaries (e.g. Sluimer *et al.*, 2005). Both indices are symmetric and their values approach 0 with increasing resemblance between the segmented image and the GT. Both are directly connected to the distance distribution signature (Huang & Dom, 1995).

$$d_{\min}(\mathbf{x}, B_j) = \min\left\{d_E(\mathbf{x}, \mathbf{y}) | \mathbf{y} \in B_j\right\},$$

where $d_E(\mathbf{x}, \mathbf{y})$ denotes the Euclidean distance between points $\mathbf{x}$ and $\mathbf{y}$, B_j denotes set of boundary points of either I or GT. So $d_{\min}(\mathbf{x}, B_j)$ is the minimum distance of a point $\mathbf{x}$ (for example on boundary B_i) to boundary B_j.

$$h(B_I, B_{GT}) = \max\left\{d_{\min}(\mathbf{x}, B_{GT}) | \mathbf{x} \in B_I\right\},$$

$$\mathrm{HAUSD} = \max\left\{h(B_I, B_{GT}), h(B_{GT}, B_I)\right\},$$

$$\mathrm{MASD} = \frac{1}{2}\left[\bar{d}_{\min}(B_I, B_{GT}) + \bar{d}_{\min}(B_{GT}, B_I)\right],$$

where $\bar{d}_{\min}(B_I, B_{GT})$ denotes average (minimum) distance from all points $\mathbf{x}$ from B_I to B_{GT}.

Algorithms evaluation

The study of image segmentation algorithms performance is presented in this section. First, few remarks connected to the input data set and experimental setup are made. They are necessary to correctly interpret the results. Then the evaluation is carried out which mainly consists of answering two important questions – whether there is such segmentation method that would outperform the others in the studied set, and (if not) whether it is possible to choose method that is sufficiently good in the majority of cases. In the following part (Section discussion of the achieved results) the results are analysed in more detail and the generally applicable recommendations concerning the performance of the algorithms are proposed. Finally, the applicability is shown on different but related data set, that is, microscopic biological images.

The input data set and evaluation setup

The algorithms for image segmentation in this paper are evaluated on a data set of the cross-section images of the artworks. They originate from the painting restoration process in which the minute samples are taken away from the artwork, embedded in polyester resin, grounded at a right angle to a surface plane and ground to expose the painting layers. Afterwards the samples are captured in three modalities – visible (VIS) and ultraviolet (UV), complemented with a study under scanning electron microscope (SEM). The microscope Olympus BX-60 (Olympus, Tokyo, Japan) equipped with digital camera Olympus DP70 is used for acquisition of VIS and UV images. In case of UV the radiation of 330–380 nm is produced by Hg discharge tube. SEM images are acquired by Philips XL30 CP scanning electron microscope (Philips, Amsterdam, Netherlands) at working voltage 25 kV with the use of Robinson detector of back-scattered electrons. The typical magnification is between 100× and 320×. The images come from the Academic Materials Research Laboratory of Painted Artworks (ALMA),[9] where they help the art restorers to choose the proper materials and appropriate technique for the very restoration. The images do not always form the triplet (SEM modality is often missing). There are 148 VIS images, 148 UV images and 89 SEM images. The SEM images are greyscale, the other two modalities are in RGB colourspace. This also permits to evaluate the performance of the image segmentation algorithms in different colourspaces (or their subspaces) like LUV or LAB (Pratt, 2007)[10].

Some of the artefacts, which are present in the cross-section images and make their segmentation difficult, can be diminished. The polyester resin, which the minute sample is embedded in, has to be ground by fine sandpaper to expose the painting layers. The grinding produces the artefacts in the captured image in the form of omnipresent parallel lines which have undesired impact on outcome of specific image segmentation methods. The method for removal of such artefacts is based on the Fourier transform and makes use of the distinct properties of the artefacts. For details see (Beneš *et al.*, 2011). The removal of the artefacts may improve the performance of the image segmentation methods evaluated in this paper (see Fig. 2). A study

[9] http://www.alma-lab.cz

[10] Naturally this applies only to UV and VIS images. SEM images are processed as greyscale. Also not every colourspace or its subspace is used for every segmentation method. Only those with meaningful results are included in the studied set.

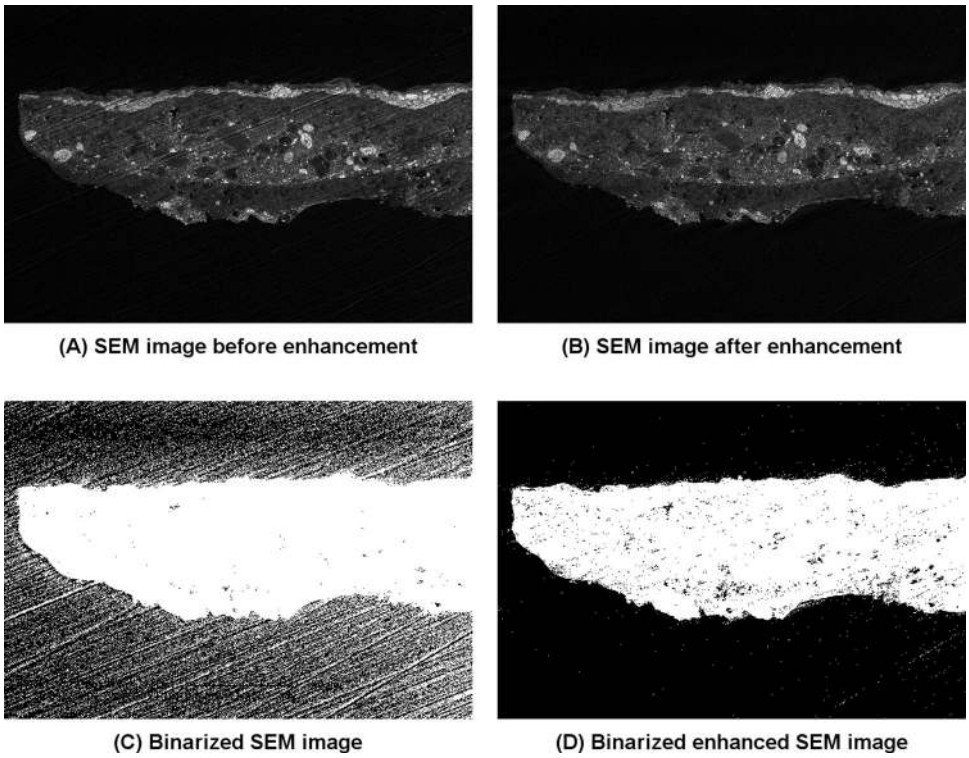

Fig. 2. The background artefacts might influence the outcome of the segmentation algorithm. In (A) there is an image with the artefacts, in (B) the image is after enhancement (artefacts are removed). (C) and (D) illustrate the influence of artefacts (non)presence on segmentation method. Image in (A) courtesy of ALMA, Prague.

was conducted to find out which segmentation methods from the set are liable to grinding artefacts. In case of such methods the preprocessed images with suppressed artefacts are used as input for segmentation. This ensures that the segmentation results are not influenced by the presence of artefacts.

Next remark regarding the input data set concerns GT images as the reference standard for the evaluation of the image segmentation algorithms performance. They were obtained manually for each image in the input data set. The delineation of the sample boundary (i.e., the foreground) is a troublesome process even for the art restorer because of the difficulties mentioned earlier. The object boundary is not always clear. Sometimes the top or the bottom material layer is not even visible because the lack of contrast to the background. However, the final binary masks produced in cooperation with ALMA represent suitable reference standard.

The second group of remarks is dedicated to the algorithms' parameters setting and their initialization. The behaviour and so the output of the selected image segmentation algorithms can be considerably influenced by various setting of their input parameters. The parameters of some methods are plainly interpretable and as such they can be adjusted appropriately to obtain the best results. For the rest the experiments with different sets of parameters were performed and the parameter set with the best output was selected. The same goes for the parameter of BHD quality index, which is the only quality index with parameter.

The second issue is the initialization of some segmentation methods. For example the RG demands the indication of the initial seed points. Considering the properties of the images the pixels with the most typical intensity on the border of the image (i.e. the mode) can be taken as the seed points. The algorithm then groups the pixels similar to the seeds by intensity with given tolerance (given as a parameter and added to the abbreviation, e.g. RG_25; there are 7 different parameters used in the studied set). The Grabcut algorithm requires user initialization in the form a rectangle with a potential foreground inside. This task is done automatically in our case and the rectangle is set to cover the most of the image except for the narrow band of pixels around the image border.

Finally, the aim is to obtain the final masks without small noisy regions in the background and with the smooth border of the foreground. Hence, the resulting binary masks after the segmentation are slightly postprocessed using mathematical morphology.

Single best segmentation method

The goal of this subsection is to find out whether there is such image segmentation method in studied group of methods

that solely outperforms the others in processing the input images in terms of quality. That means if there is method which gives better segmentation result for significant majority of images (or for each image in extreme case) in the data set than every other method in the group. If so, use of such method would be of general preference to solve background removal problem of similar data.

To study prevalence of any method first we need to denote the best segmentation algorithm for every image in input data set separately (see algorithm 1 for pseudocode). Ten quality indices (described in section quality indices) have to be computed for every such image and every segmentation algorithm. Then the algorithm with the best result may be picked by each index for each image. It is the algorithm with the best correspondence to the respective GT, so the algorithm with maximum (or minimum) index value is picked. After this, there are 10 possibly different segmentation methods selected by each quality index for every image. To obtain single decision for every image some combination rule has to be applied. Since the quality indices can be interpreted as 10 different voters, voting rules can be successfully used in this situation. In our case the relative majority rule is considered. It means that for every image the segmentation method which is the most frequently selected as the best one by individual indices is the best segmentation method for the particular image overall. This gives us the best segmentation method for every image in input data set.

Algorithm 1. Denotes the best segmentation algorithm for an image

Require: image I
  for all Q from the set of quality indices **do**
    $result \leftarrow$ empty vector
    for all M from the set of segmentation methods **do**
      compute Q on the result of M on I to obtain value val_Q
      $result(M) \leftarrow val_Q$
    end for
    $M_Q \leftarrow \arg\max_M \{result(M)\}$ {or min depending on the index}
  end for
  apply majority vote on all M_Q to obtain M_{BEST}
  return M_{BEST}

It would be useful to verify that the best segmentation method selected by quality indices according to the described procedure is also visually the best segmentation method from the set available for each image. Therefore visual comparison of all the segmentation results for every image was performed with extra focus on cases where the result of the selected best method was not too close to the GT (we need to verify that there is no better result available). The analysis leads to conclusion that the quality indices behave correctly in a vast majority of cases. The selected result is either one of the many proper ones or it is the only viable output. If there is no satisfactory result of any segmentation method, then the one visually most plausible is often selected. However, there are some cases where the indices (or majority vote) do not decide entirely correctly. The selected result is not visually the best available though it is very similar to it. In such cases the decision of the indices is usually far from being unanimous. Each index may favours a different method and final decision using majority vote would be supported by small number of indices.

In any case, we have the best segmentation method denoted for every image in input data set. The key conclusion of this section is based on a distribution of segmentation methods among the best methods selected by quality indices and voting for each image. In this section we focus only on the most frequent segmentation methods which have potential to be the best. Deeper analysis with additional material is given in the Appendix. The results are presented separately for each modality. They naturally differ due to distinct character of those modalities and their input images. This gives us opportunity to study performance of the algorithms in different conditions.

The two most frequent segmentation methods in SEM modality are Felzenszwalb's method (GC_FH) and RG (with parameter equal to 5 – RG_5) with 12 occurrences out of 89 possible (number of SEM images in total) each among the best methods. The situation in UV modality is rather different. MS is clearly the most successful method. It is better than any other method in 34 cases out of 148 (the total number of UV or VIS images). In VIS modality, MS stays the most frequent among the best methods for each image with 40 occurrences out of 148 possible. Nevertheless, the most frequent segmentation methods outperform the others only in fraction of cases (13–27% depending on modality).

Based on these facts we can say that there is no segmentation method which significantly outperforms the other segmentation algorithms in the set. The use of the most frequent method mentioned in previous paragraph (e.g. MS for UV modality) for background removal in images similar to those in our data set is not sufficient for achieving perfect results (see Fig. 3 for example of an image where the best method does not perform that well). It is important to keep in mind that potential user usually does not have the GT images, so he cannot select the individual best method for every sole image. Additional conclusions can be made from the results. MS, GC_FH, GC_R and MNC often perform well. But also more straightforward approaches such as RG or thresholding can be used to achieve good results (see the Appendix for reasoning).

Best average segmentation methods

The evaluation in the previous section is not entirely fair. The focus was on finding a segmentation method which was the best for significant majority of images. There was no such method in the studied set. However, what if there is a method which is good enough (and not necessarily the best) for vast

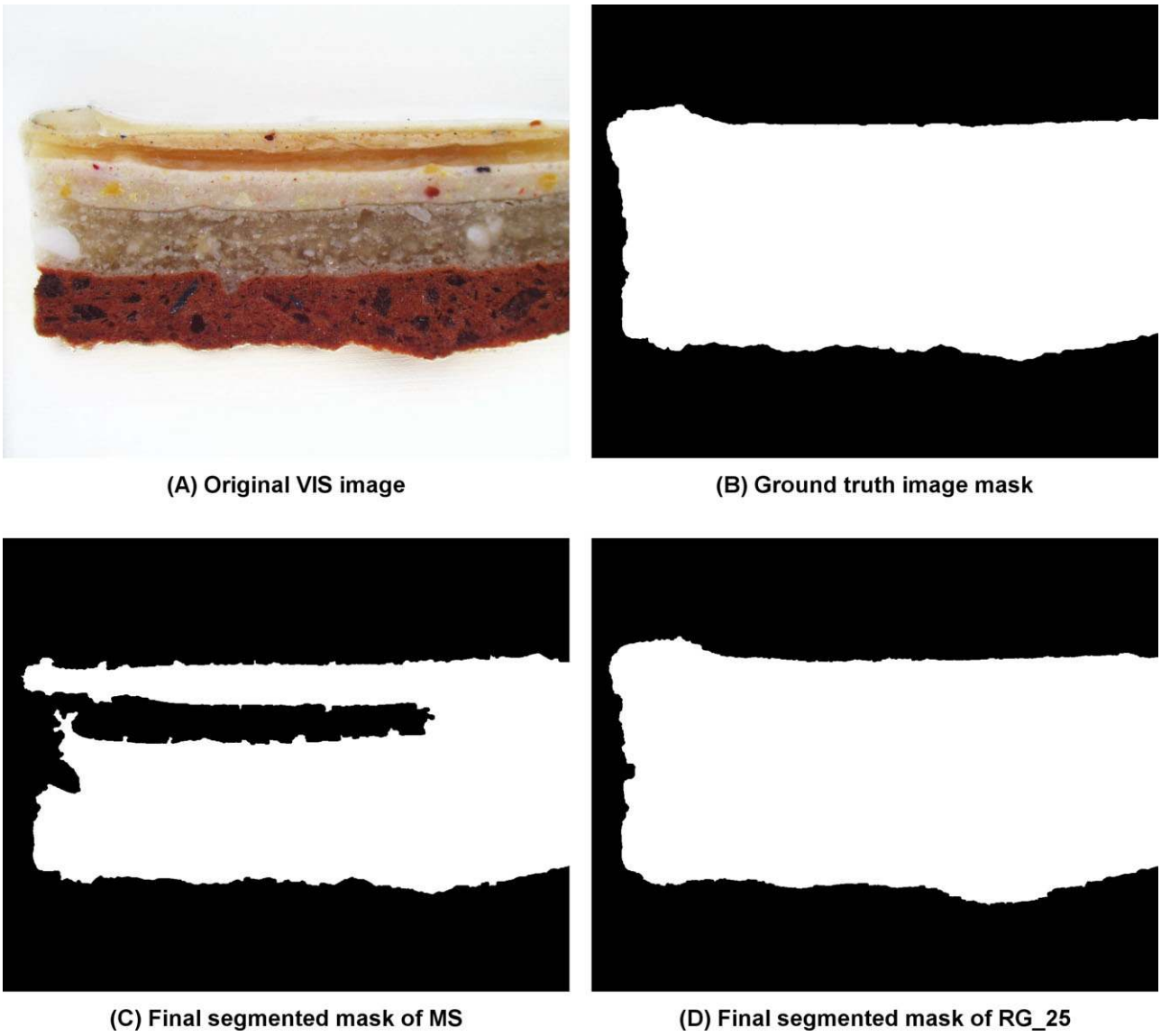

Fig. 3. Demonstration that the selected best method is not perfect for all images. The image in (A) is better segmented by RG with parameter 25 (RG_25, in D) than MS (in C) which is the best method in VIS modality. RG does not perform nearly that well overall. In (B) there is a GT image for reference. Image in (A) courtesy of ALMA, Prague.

majority of the images. We look for method which is comparable to the best method in case of easy to segment images (majority methods can segment this image with satisfactory results) and does not completely fail in case of worse images (where most of the methods fail), that is, the best average segmentation method. Such method (if found) could be used as number one choice to solve the image segmentation problem.

The starting point for the evaluation is the same as in the previous section. The values of 10 quality indices are computed for each image and segmentation method. However, following steps differ from the previous procedure (see Algorithm 2). There are so many values as there are images for every pair of quality index and image segmentation method. Median of these values is the average performance of segmentation method according to the respective index. The best average method is thus the method with the highest median (or the lowest depending on the index). Finally, the majority rule denotes the best average segmentation method as a consensus of all quality indices. The median is preferred over the mean because vectors of numbers often contain several outliers which would distort the results inappropriately.[11]

[11] Outlier means that segmentation method segments some image exceptionally well or poorly. Outlier is the value of the quality index for such image. We are interested in average performance which has to be stable despite the outliers. That is why the median is more suitable for the task.

Table 2. Table with median values and interquartile ranges in brackets (both rounded to two decimal places) of all 10 quality indices for several selected segmentation methods in SEM modality. Median value is the average performance of a segmentation method on a set of images according to a quality index. There are the six most successful methods, several representative methods in the middle and the two worst methods according to evaluation in Section best average segmentation methods (in this order). SEM modality is chosen for demonstration due to bigger variance in indices values for different methods in different places of the ranked list than it is in other two modalities.

	Quality indices				
Segmentation methods	BHD [0,1]	HD [0,1]	RI [0,1]	ARI [−1, 1]	VI [0, . . .)
GC_RD	0.84 (0.12)	0.98 (0.03)	0.96 (0.06)	0.90 (0.15)	0.29 (0.31)
GC_FH	0.82 (0.13)	0.98 (0.03)	0.96 (0.06)	0.90 (0.20)	0.28 (0.24)
MS	0.82 (0.14)	0.97 (0.04)	0.95 (0.08)	0.88 (0.23)	0.33 (0.34)
GC_CV	0.84 (0.14)	0.97 (0.06)	0.95 (0.11)	0.88 (0.31)	0.32 (0.37)
IMJ_IME	0.81 (0.14)	0.97 (0.04)	0.94 (0.07)	0.89 (0.20)	0.32 (0.31)
RG_10	0.82 (0.15)	0.97 (0.05)	0.94 (0.10)	0.87 (0.23)	0.33 (0.38)
IMJ_TRIANGLE	0.77 (0.18)	0.97 (0.09)	0.93 (0.15)	0.86 (0.39)	0.39 (0.47)
GC_R	0.73 (0.22)	0.96 (0.10)	0.93 (0.17)	0.82 (0.41)	0.37 (0.45)
KM	0.63 (0.17)	0.87 (0.20)	0.78 (0.27)	0.40 (0.55)	0.77 (0.57)
IMJ_OTSU	0.61 (0.17)	0.84 (0.18)	0.74 (0.24)	0.38 (0.53)	0.82 (0.48)
TNC	0.49 (0.29)	0.81 (0.28)	0.70 (0.34)	0.01 (0.84)	0.81 (0.55)
RG_70	0.49 (0.09)	0.75 (0.19)	0.64 (0.17)	0.02 (0.19)	0.93 (0.28)
MNC	0.50 (0.05)	0.57 (0.17)	0.51 (0.05)	0.01 (0.09)	1.66 (0.35)
IMJ_SB	0.46 (0.04)	0.70 (0.19)	0.58 (0.12)	0.00 (0.00)	0.88 (0.24)
	FMI [0, 1]	DC [0, 1]	NMI [0, 1]	HAUSD [0, . . .)	MASD [0, . . .)
GC_RD	0.96 (0.05)	0.97 (0.07)	0.82 (0.22)	40.31 (65.19)	4.57 (8.03)
GC_FH	0.96 (0.04)	0.96 (0.10)	0.83 (0.26)	32.60 (54.43)	4.43 (7.28)
MS	0.96 (0.07)	0.95 (0.11)	0.81 (0.23)	45.50 (68.99)	5.71 (10.63)
GC_CV	0.96 (0.08)	0.94 (0.16)	0.79 (0.33)	53.48 (71.93)	5.79 (13.93)
IMJ_IME	0.96 (0.06)	0.95 (0.09)	0.80 (0.25)	48.71 (94.05)	5.82 (11.98)
RG_10	0.95 (0.08)	0.95 (0.10)	0.78 (0.26)	57.24 (97.56)	6.40 (12.15)
IMJ_TRIANGLE	0.95 (0.10)	0.93 (0.26)	0.77 (0.38)	54.58 (127.65)	7.00 (22.40)
GC_R	0.94 (0.11)	0.92 (0.16)	0.74 (0.37)	56.04 (71.01)	7.98 (21.90)
KM	0.85 (0.17)	0.62 (0.46)	0.39 (0.43)	118.17 (177.56)	29.42 (47.81)
IMJ_OTSU	0.82 (0.16)	0.60 (0.48)	0.36 (0.40)	124.39 (192.89)	34.48 (58.20)
TNC	0.81 (0.19)	0.12 (0.89)	0.05 (0.71)	361.82 (449.15)	111.70 (129.26)
RG_70	0.78 (0.11)	0.11 (0.37)	0.08 (0.20)	370.18 (333.21)	98.91 (82.07)
MNC	0.59 (0.09)	0.39 (0.31)	0.02 (0.11)	197.50 (75.12)	63.94 (23.54)
IMJ_SB	0.76 (0.10)	0.00 (0.01)	0.01 (0.02)	523.64 (181.98)	144.72 (41.20)

Table 2 shows median values for each quality index and several selected segmentation methods in SEM modality.

Algorithm 2. Denotes the best average segmentation algorithm overall

for all Q from the set of quality indices **do**
  $result, medians \leftarrow$ empty vectors
  for all M from the set of segmentation methods **do**
    for all I from the set of input images **do**
      compute Q on the result of M on I to obtain value val_Q
      $result(M, I) \leftarrow val_Q$
    end for
    $medians(M) \leftarrow \underset{I}{median}\{result(M, I)\}$
  end for
  $M_Q \leftarrow \underset{M}{arg\,max}\{medians(M)\}$ {or min depending to the index}
end for
apply majority vote on all M_Q to obtain M_{BESTAVG}
return M_{BESTAVG}

Felzenszwalb's method (GC_FH) and Rousson–Deriche approach (GC_RD) are the two best average methods for SEM modality (they were selected equally by the indices). If we look on the problem of finding the best average segmentation method even in more detail and consider first five methods for each quality index (assuming that the lists for each index are sorted by median values, thus by performance), we can see that GC_FH and GC_RD occupy the first two positions of

Table 3. Lists with first five segmentation methods (rows) according to every quality index (columns) in SEM modality. Lists are sorted by median values, thus by average performance of segmentation methods.

Quality indices									
BHD	HD	RI	ARI	VI	FMI	DC	NMI	HAUSD	MASD
GC_RD	GC_RD	GC_RD	GC_RD	GC_FH	GC_FH	GC_RD	GC_FH	GC_FH	GC_FH
GC_CV	GC_FH	GC_FH	GC_FH	GC_RD	GC_RD	GC_FH	GC_RD	GC_RD	GC_RD
RG_15	MS	MS	IMJ_IME	GC_CV	GC_CV	IMJ_IME	MS	MS	MS
RG_10	GC_CV	GC_CV	GC_CV	IMJ_IME	MS	MS	IMJ_IME	IMJ_IME	GC_CV
MS	IMJ_IME	IMJ_IME	MS	RG_10	IMJ_IME	RG_10	GC_CV	GC_CV	IMJ_IME

almost every list (there is only one exception) in SEM modality (see Table 3). Considering the median values there is a noticeable gap between these two and next methods in the list. This second cluster is formed by Chan–Vese approach (GC_CV), MS and minimum error thresholding (IMJ_IME). Apart from them there are several occurrences of RG with parameters 10 and 15 on lower positions. MS holds its superiority in UV modality even as the best average method. It is first for 9 out of 10 quality indices (only HAUSD votes for GC_FH) with substantial performance gap from the second position which is occupied almost only by GC_FH (except for HAUSD naturally). Two colourspace versions of multiscale normalized cut (MNC, RGB and greyscale) fill the third and the fourth position. The last one with other noticeable loss in performance is mainly RG with parameter 25 (RG_25). There are sporadic occurrences of other methods from studied set on lower positions, but nothing of importance. The result in VIS modality is not so clear. Majority vote denotes MS to be the best average method, since five quality indices vote for it. Nonetheless four indices are for MNC (in RGB) and one for GC_FH. The rest of the first five positions is shared by plenty of different methods including thresholding, RG, K-means etc. The conclusion is that there exist four very good methods which can be used as number one choice depending on the modality. It is GC_FH and GC_RD for SEM, MS for both UV and VIS modality, in the latter case supported by MNC (in RGB).

The evaluation of previous paragraph can be done more rigorously with the removal of the following shortcoming in addition. The choice of the best average method (and four runners up) was based on the position within ten sorted list coming from ten quality indices. Unfortunately the situation when one method was chosen as the best one by several indices and given a lower rank by others was not taken into account because only first five positions were considered. Therefore, the results could be little bit inaccurate. This drawback can be amend by exploiting the information about performance of all the methods from all the indices, that is, by processing complete sorted lists of indices' values. The goal is to combine all ranked lists to the single ordering which would express input preferences in the best way. This is called a rank aggregation problem and is extensively studied in different fields (elections, web search etc.). See, for example, Dwork *et al.* (2001) in context of web searching. We use RankAggreg package (Pihur *et al.*, 2009) for R statistical software[12] for our evaluation. It implements optimization techniques necessary to produce final ranked list.[13] The rank aggregation algorithm from the RankAggreg package minimizes the objective function to obtain final ranked list δ^*

$$\delta^* = \underset{\delta}{arg\,min} \sum_{i=1}^{m} d(\delta, L_i),$$

where L_i is ith input list and d is a distance function. Spearman distance is used as a distance function d because it suits our problem better than Kendall's tau distance (see Pihur *et al.*, 2009, for more details on problems). Spearman distance is equal to the summation of the absolute differences between the ranks (positions) of all unique segmentation methods from two ordered lists.

$$d(L_i, L_j) = \sum_{t \in L_i \cup L_j} \left| r^{L_i}(t) - r^{L_j}(t) \right|,$$

where $r^{L_i}(t)$ is the position of method t in a list L_i. Finally, the Cross-Entropy Monte Carlo algorithm is selected for minimization (see the mentioned paper for details). As a result there is one list of image segmentation methods sorted by their performance (according to quality indices) for each modality. This list represents consensus of ten input lists as individual voters with preferences.

It is impossible in this limited space to deeply analyse positions of every segmentation method in the final lists. Hence we focus only on several prominent methods, interesting results and general position of different approaches (comprehensive analysis is given below in Section discussion of the achieved results). The complete lists are appended in Table 4. Rousson–Deriche approach (GC_RD) and Felzenszwalb's method (GC_FH) stay the best average methods in SEM modality with

[12] http://www.r-project.org.

[13] Optimization is unavoidable because due to amount of data (10 relatively long lists) the exact solution cannot be computed in feasible time. However, exact solution can be computed for short input lists and they more or less match the corresponding part of presented optimization results. Unfortunately implemented optimization algorithms do not necessarily find a global optimum and can get stuck in a local one. The scripts were therefore executed many times to obtain as best solution as possible.

Table 4. Final lists of segmentation methods sorted according to their average performance (the best in the first place) in all three modalities.

SEM	GC_RD, GC_FH, MS, GC_CV, IMJ_IME, RG_10, RG_15, HT_ME,IMJ_TRIANGLE, IMJ_MEAN, HT_MEAN, HT_IME, GC_R, IMJ_HUANG, RG_20, IMJ_LI, RG_5, RG_25, KM, HT_INTER, HT_INTERI, IMJ_DEF, HT_CONCAV, IMJ_ISO, IMJ_OTSU, RG_50, HT_MOM, IMJ_MOM, IMJ_PER, HT_IM, TNC, IMJ_IM, HT_MEDIAN, IMJ_RENYI, RG_70, IMJ_YEN, HT_MIN, HT_MAXLIK, HT_ENT, IMJ_MAXENT, IMJ_MIN, MNC, IMJ_SB
UV	MS, GC_FH, MNC_GRAY, MNC_RGB, RG_20, GC_R_LAB(AB), RG_25,GC_CV, RG_15, IMJ_TRIANGLE, KM_LAB(AB), HT_MEAN, IMJ_HUANG, RG_50, TNC, GC_R_LAB, IMJ_LI, IMJ_MEAN, RG_10, KM_GRAY, KM_LAB, RG_70, HT_INTER, HT_ME, HT_INTERI, IMJ_DEF, KM_RGB, MNC_LAB(AB), IMJ_OTSU, GC_R_LAB(L), HT_CONCAV, IMJ_ISO, MNC_LUV(L), HT_MOM, GC_RD, IMJ_MOM, HT_IM, HT_MAXLIK, IMJ_IM, GC_R_RGB, HT_MIN, IMJ_YEN, IMJ_MIN, IMJ_RENYI,HT_ENT, IMJ_MAXENT, IMJ_IME, RG_5, HT_IME, IMJ_PER, HT_MEDIAN, IMJ_SB
VIS	MS, MNC_RGB, KM_RGB, IMJ_OTSU, IMJ_ISO, IMJ_DEF, IMJ_HUANG,HT_INTERI, TNC, MNC_LUV(L), GC_CV, HT_INTER, KM_LAB, KM_GRAY, IMJ_MEAN,IMJ_MOM, IMJ_IME, HT_MEAN, HT_MOM, IMJ_IM, RG_70, RG_50, IMJ_LI,MNC_GRAY, HT_IM, IMJ_RENYI, GC_FH, IMJ_MIN, HT_MAXLIK, HT_MIN,GC_R_LUV(UV), IMJ_YEN, KM_LAB(AB), RG_25, MNC_LAB(AB), RG_20, HT_ENT,GC_R_LUV(L), RG_15, IMJ_TRIANGLE, GC_RD, HT_CONCAV, HT_ME, IMJ_PER,IMJ_MAXENT, HT_MEDIAN, RG_10, HT_IME, RG_5, IMJ_SB

that GC_RD is the best one. This result is little bit surprising, because GC_RD was not so successful as the best method overall (in previous Section single best segmentation method) and nothing indicated that it would outperform the others on average. MS algorithm MS and Chan–Vese approach (GC_CV) follow the two. Iterated and normal version of minimum error thresholding is very successful (both ImageJ and HistThresh, i.e. IMJ_IME, HT_IME and HT_ME), as well as Triangle and Mean approaches (IMJ_TRIANGLE and IMJ_MEAN). RG with parameters 10 and 15 occupies position 6 and 7 in the list, other parameters are scattered in the middle. From already mentioned methods K-means (KM) and GrabCut (GC_R) rather disappoint with its results and multiscale normalized cut (MNC) completely fails with the last but one position.

MS is the best average algorithm in UV modality, which only confirms its dominance. It is followed by GC_FH and greyscale and RGB versions of MNC, which is very opposite from SEM modality, where greyscale version fails. Parameters 15, 20 and 25 of RG are suitable for UV modality as they are placed in top 10 also with GC_CV method. IMJ_TRIANGLE, IMJ_MEAN, IMJ_HUANG and IMJ_LI are the most useful thresholding methods. Several colourspace alternatives of KM are ranked in the top half. Contrary to SEM modality GC_RD method is not very good as it is ranked in bottom half of the list. The least successful method is Shanbhag (IMJ_SB) approach to thresholding. It is interesting that this method was voted as the best one overall for one image (previous Section single best segmentation method) despite its uselessness on average.

MS is the best average algorithm also in VIS modality, but otherwise the situation differs a lot compared to previous two modalities. In the second and third place there are RGB version of MNC and RGB version of KM algorithm. Apart from them top 10 consists further from thresholding methods, IMJ_OTSU, IMJ_ISO, IMJ_HUANG and Tao's thresholding method (TNC) to name several. GC_CV algorithm produces satisfactory results. GC_FH, GC_R or GC_RD do not perform very well. Concerning RG approach its results are generally worse than in the previous two modalities. However, higher values of parameter like 50 or 70 are definitely better than smaller ones. IMJ_SB thresholding is again the worst segmentation method on average.

The evaluation in this section delivers very interesting results. The most important is the construction of lists of segmentation methods sorted by algorithms' performance according to 10 selected quality indices. The ordering allows the future user to pick the suitable segmentation method for his problem and character of data (which are represented by different modalities in this paper). The lists also provide an insight to performance of different segmentation methods and their comparison. The conclusions about the performance depend on the specific modality, but generally some resume can be made. MS algorithm performs very well in all three modalities and can be declared the best average method overall. Felzenszwalb's method, Rousson–Deriche and Chan–Vese approaches, and multiscale normalized cut may deliver excellent results as well. RG is not a bad choice either, but its performance depend on the chosen parameter. Thresholding can be good alternative too, but the choice of specific algorithm has to respect the properties of data. Segmentation methods which take place at the end of the lists perform badly on average, however that does not necessarily mean that they perform badly on every image (for example see Fig. 3, where RG

outperforms the best method on average – MS. RG_25 is ranked in the bottom half). Furthermore, they may provide important diversity for segmentation fusion/combination or other processing (Section combination of image segmentation methods). More discussion and conclusions are presented in Section discussion of the achieved results.

One remark concerning correctness of the above evaluation has to be made before closing this section. The comparison does not take into account the absolute values of quality indices. So it is possible that the best average segmentation method is certainly better than the rest of the methods in the studied set, but absolutely its performance is poor with useless results. However, it is not the case. The segmentation methods at the top of the lists obtained relatively high values from the quality indices (and vice versa for the methods at the bottom). See Table 2 for reference in case of SEM modality. The further evaluation was performed to support this conclusion more precisely. The output of segmentation method on one image was marked good if its index value was above specified threshold (and bad if it was below another). Afterwards all the methods were ranked according to the number of their occurrences in a set of good outputs and a set of bad outputs. The results of this evaluation did not differ much with the results of this section described above.

Discussion of the achieved results

In this section, deeper analysis of the evaluations and their results is presented. We will use it to make recommendations for the application of studied image segmentation methods in different situations, that is, for different (but still related – microscopic) data. First, the distinct features of each modality (SEM, UV and VIS as shown in Fig. 1) are examined in more detail. Then the performance of each segmentation approach and its connection to input images (or modality) is evaluated to make clear in which situations which image segmentation methods perform the best.

SEM modality images are products of scanning electron microscope. This technique enables to study chemical contrast of different materials. In the image it is expressed by varying texture of the cross-section in contrast to relatively homogeneous background. Thus, the boundary edges between the cross-section as foreground object and the background are usually sharp and clear. The cross-section has generally different intensity values than the background. All this could make the segmentation quite easy. However, in case of our data set the task is sometimes complicated with the artefacts induced by scanning microscope, and certain materials used in the paintings do not have sufficient contrast response so the boundary edge is not sharp enough.

UV modality is similar to SEM in that the background is homogeneous. UV light reveals a possible fluorescent property of certain materials. Such materials have bright response (typically green, turquoise or blue) in the image. Nonfluorescent materials are on the other hand often dark and they blend with the background which is dark by definition due to absence of fluorescent property of polyester resin. Another problem is that the nonsurface parts of the cross-section can shine through transparent resin and form blurred shadows on the borders of the cross-section. Satisfactory background removal can therefore be quite challenging.

VIS modality captures optical properties in visible spectrum. The sharpness of cross-section boundary varies from high contrast edge to fluent transition to background depending on the material colour. The transparency of polyester also remains a problem in VIS modality. The difficulty of background removal is thus similar to UV modality in this aspect. In addition, the background is not uniform. The lighting can be reflected unevenly and there can be lot of different artefacts like air bubbles which are not visible in other modalities. Also grinding artefacts may be a problem as was mentioned before. Figure 4 gives examples of distinct properties of VIS images.

To summarize key properties of the modalities SEM modality generally represents microscopic images with sharp and contrast boundary edges, relatively homogeneous background and often clear separation of object and background intensity values. UV modality images have uniform background, but unclear boundary edges between background and certain (nonfluorescent in our case) parts of the foreground object, also transparency of the resin is the problem. VIS images are similar to UV in problems with unclear boundary edges and transparency of the resin. Difference is in more problematic background which is not uniform and contains artefacts.

Discussion about the usability of studied segmentation methods starts with simpler approaches, that is, RG, thresholding and K-means.[14] *RG* generally delivers satisfactory results when there is relatively homogeneous background and boundary between desired segmented object and background is apparent. In our case it is demonstrated on SEM and UV modalities where the background surrounding the cross-section is more or less uniform. Tolerance to nonuniformity is given by parameter. The smaller values of parameter are sufficient for images in SEM, whereas slightly higher values are required for UV to compensate the transparency mentioned above. RG is then placed in top 10 of the best average methods. VIS modality is different. The background there is more variable in such way that it almost prohibits compensation with high parameter values (RG would easily cross the border between background and foreground object in that case). This being said high values of parameter are more suitable in VIS. Overall RG approach can provide satisfactory results comparable to more complicated methods if the assumptions of relatively uniform background and clear border are met.

Thresholding methods (not only those in the studied set) differ in the way they find the threshold to divide pixels

[14] Concerning different colourspaces RG and thresholding exploited only the greyscale information in all three modalities. K-means was evaluated in more colourspaces.

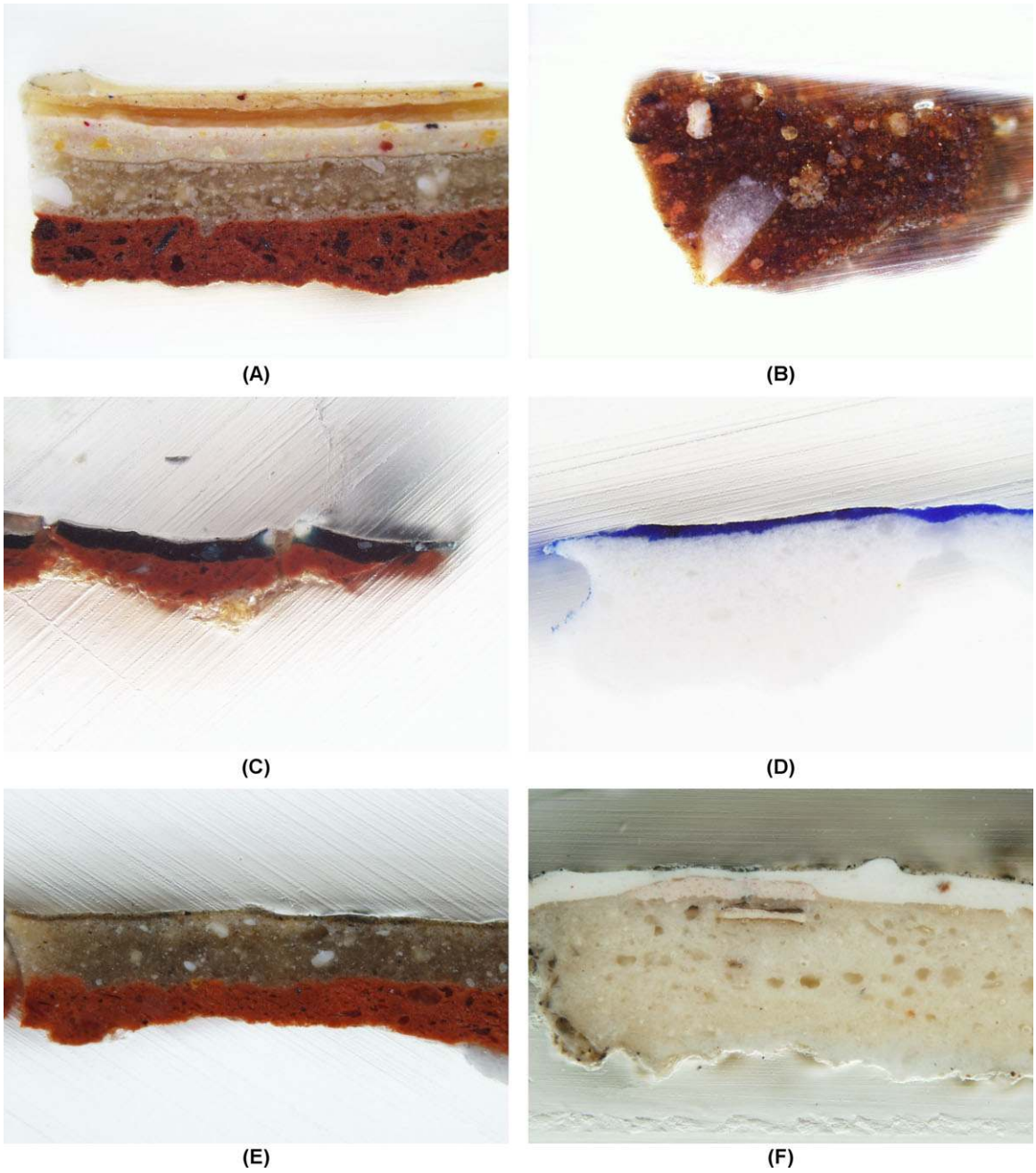

Fig. 4. Set of six VIS images demonstrating different properties which cause problems for image segmentation. In (A) there is neat and relatively easy to segment image for comparison. Other images demonstrates nonuniform illumination of the background (E and F), problematic transparency of the polyester resin (B and C), grinding artefacts (C–F), air bubbles and defects in the background (C and F) and finally unclear boundary edge between cross-section and background (D). Image courtesy of ALMA, Prague.

into two groups. Strictly bimodal histogram would be an optimum situation, however such case is not very common in our input data set (and in real images neither). Therefore, some methods are more successful in handling nonoptimum case than others. In SEM modality where the background pixels in histogram are easier to separate Triangle (IMJ_TRIANGLE), Mean (IMJ_MEAN), and minimum error method (IMJ_IME) are the most successful. On the other side of spectrum there are entropy-based methods (IMJ_MAXENT, IMJ_RENYI, IMJ_SB, HT_ENT) and several others (HT_MAXLIK, IMJ_YEN, IMJ_MIN). In UV modality the intensity values of the foreground often blend with those of the background, which is difficult condition for thresholding. Triangle, Huang (IMJ_HUANG), Mean and Li (IMJ_LI)

methods handle it well on average. The spectrum of failing methods stays the same as in SEM modality. IMJ_IME produces disappointing results too. Though the image properties of VIS modality are similar to those of UV mostly different thresholding methods are satisfactory in VIS. Otsu (IMJ_OTSU and HT_INTER), IsoData (IMJ_ISO, IMJ_DEF and HT_INTERI) and Huang are among the most successful methods. Concerning Tao's thresholding approach (TNC) it succeeds in UV and VIS modalities, whereas it fails in SEM. Thus, it deals better with visually hard cases with smooth transitions between background and foreground than in cases where the intensity values of the foreground object are clearly separated from those of the background.

The results of *K-means* (KM) approach are highly dependent on colourspace (or subspace) which the input data are in and on overall colour profile of the images in different modalities. Greyscale (the only one for SEM), LAB (plus AB subspace) and RGB variants are analysed. KM in greyscale produces merely mediocre results on average in all three modalities. Same thing can be said on account of full LAB space variant (in case of UV and VIS) with slightly better results in VIS. However, interesting results appear concerning KM in AB subspace of LAB and RGB. Both can perform well depending on colour profile of the image. In UV modality where the images are mainly darker with dominant responses in blue or green, the AB variant is placed in top positions of the ranked list. RGB variant performs much worse. The situation is opposite in VIS modality. RGB variant is the third best average method whereas AB variant takes place in two thirds of the ranked list. It is clear that successful use of K-means depends on the overall colour dominance of input images. Generally, its results can be quite satisfactory.

After more straightforward approaches were analysed we will now focus on more complex segmentation methods in the studied set.[15] *Felzenszwalb's method* (GC_FH) performs very well being the second most successful average segmentation method in SEM and UV modalities. However, it does not perform that well in the remaining VIS modality. The algorithm has apparent problems with converging to stable result when the border of the object is unclear and background is not homogeneous (and in that sense resembles the foreground object). In such cases the segmented result is often blank image. Apart from that GC_FH can be excellent method for segmentation which copes with other mentioned problematic image properties appropriately. *Daněk's optimization of Chan–Vese and Rousson–Deriche functionals* is very successful for the easy to segment images with clear and sharp border between object and surrounding background (GC_RD is the best average method in SEM, GC_CV being the fourth). Otherwise they struggle with unclear transitions and transparency. GC_RD fails in UV and VIS modality, GC_CV still manages to take position in top third of the average ranked list, but its results are often dissatisfactory. The results of *multiscale normalized cut* (MNC) differ with various colourspace configurations. MNC produces very good results when the original RGB colourspace is conserved (second place in VIS modality and fourth place in UV modality average ranked list). Also the exploitation of only the intensity channel (greyscale or lightness from LUV) can be profitable in case of UV and VIS. In all other cases MNC rather fails, especially in SEM modality. *GrabCut algorithm* (GC_R) provides perhaps the worst results from group of more advanced segmentation methods and cannot be recommended for unsupervised segmentation in similar setting. Originally, it is based on user interaction and its power lies in additional adjustment of initial segmentation. *MS* is the last algorithm to discuss. According to the results of evaluation it is the best average segmentation method in the studied set. It can handle problematic image properties well and its outputs often outperforms the rest (see Section single best segmentation method).

With regard to the analyses above MS algorithm should be number one choice for image segmentation of related data. However, several other methods could perform well while respecting above conditions, that is, MNC, GC_CV, GC_RD or GC_FH. Should the execution time be an issue GC_FH especially would be an excellent choice. In that situation even plenty of thresholding methods or RG could provide good results with some limitations. Concerning three modalities it is confirmed that SEM images are easier to segment thanks to clear boundaries between foreground object and relatively uniform background. Segmentation methods perform there generally much better than in UV and VIS where the segmentation is complicated by image properties. Table 5 offers recommendations on the use of segmentation methods depending on the input image properties in the context of microscopic images.

One more evaluation was performed in addition to already described procedures. The idea was to find out what were the various segmentation methods sensitive to in the input images. For each method the images could be clustered to three groups – where the output is good, bad and the rest. If some common features for the images in such groups could be found, it would provide a lead on which segmentation method should be used when such features happen to be present in an input image. Unfortunately no common features in addition to described properties could be found in defined groups.

Finally one remark to close the evaluation. It is important to keep in mind that behaviour of some algorithms can be influenced with parameter setting. In our evaluation parameters are tuned to specific input data and we assume that same thing has to be done for different data set.

[15] From those Felzenszwalb's method is applied to the images in original colourspaces. That means greyscale in case of SEM modality and RGB colourspace in case of UV and VIS. Processing in different colourspaces delivers comparable results. MS segmentation followed the original paper and LUV space is used. Daněk's version of Chan–Vese and Rousson–Deriche use the greyscale information. So only the performances of multiscale normalized cut and GrabCut algorithm are analysed in different colourspaces.

Table 5. Table contains findings of the evaluation generalized to use in the context of microscopic images. Image in the left column of the table stands for microscopic image with essentially similar properties to the images in studied data set, preferably in one of the three studied modalities (as are described in the introduction and at the beginning of Section discussion of the achieved results). The conclusion is that MS algorithm should be number one choice segmentation method. Use of other methods depends on the input image specific properties. Details and further results can be found in the text.

Images in general	– Mean Shift algorithm would be number one choice
Image with relatively homogeneous background and apparent boundary edge between object and background	– Region growing with appropriate parameters – Felzenszwalb's method (even in the case of not so clear boundary edge and partial blending of the object and the background) – Chan-Vese and Rousson-Deriche approaches optimized by Daněk
Image with possibly unclear boundary edges between object and background, presence of shadows or halos around boundaries	– Multiscale normalized cut in RGB or applied to intensity/luminance channel
Image with easier to separate histogram	– Thresholding methods Triangle, Mean or minimum error thresholding
Image with more blended histogram	– Thresholding methods Triangle, Huang, Otsu or IsoData – Tao's thresholding approach
Image with colour composition and properties similar to UV modality	– K-means in AB subspace of LAB colourspace could deliver interesting results
Image with colour composition and properties similar to VIS modality	– K-means applied to whole RGB image could be good choice

Fig. 5. Mouse retina coloured with hematoxylin–eosin. Boundary of segmented result by Mean Shift algorithm is depicted by red line. Courtesy of Jan Cendelín, Faculty of Medicine in Pilsen.

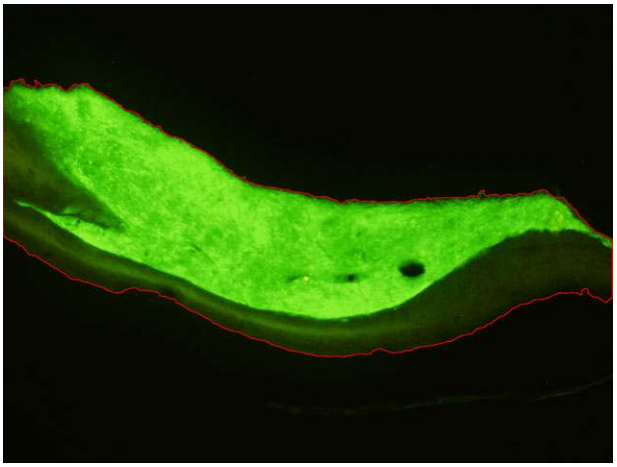

Fig. 6. Transplant mouse cerebellum. Boundary of segmented result by Mean Shift algorithm is depicted by red line. Courtesy of Jan Cendelín, Faculty of Medicine in Pilsen.

Demonstration of evaluation results applicability on different data

In this section the applicability of evaluation results to different data set – biological images – is shown. In Figures 5, 6 and 7 there are segmentation results of biological images. The first figure shows the mouse retina. Specimen is coloured with hematoxylin–eosin and captured with optical microscope in visible spectrum. It closely resembles VIS modality of cross-section images, because boundary edges are not clear enough and the background contains plenty of debris. The second figure shows transplant mouse cerebellum. Cells of the transplant generate enhanced green fluorescent protein (EGFP) so they are easily distinguishable from recipient tissue under a fluorescent microscope. The aim is to segment whole tissue (both original and transplant) from the background. The third figure shows 2D projection of 3D rendering of an early stage mouse heart, acquired by optical projection tomography. The image shows fluorescence excitation and emission. Last two figures resemble UV modality of cross-section images. The background is homogeneous and boundary edges are not so clear. The debris and other unwanted structures are also present in the background. Although it is not as visible as in the case of Figure 5, it makes segmentation problematic. The best average segmentation method for UV and VIS modality is applied, that is, MS algorithm. The results are depicted by red boundary line in respective figures. Also combination of the best three methods was generated following findings of the next section. However, in case of these three images combination results were very similar to those of MS with negligible differences, so they are not shown in the figures.

Fig. 7. 2D projection of 3D rendering of an early stage mouse heart. Boundary of segmented result by Mean Shift algorithm is depicted by red line. Courtesy of Martin Čapek, Institute of Physiology AS CR, Prague.

Combination of image segmentation methods

In Section best average segmentation methods we found (for each data modality) the image segmentation method which performed the best on average on input data set. The average means that this segmentation method often offers satisfactory results but sometimes it can fail (but not in such scale as other methods in the studied set). Next methods in ranked list (second, third, ...) can behave differently (and due to their different fundamentals they often do) with failing on other images than the best method. Therefore, it would be useful to somehow combine the results of several segmentation methods to remove unfavourable results and by doing so improve the overall performance of the segmentation process. The idea of combination comes from the classifiers domain. Kittler *et al.* (1998) in their paper provided theoretical framework for combining classifiers. Key idea is to exploit advantages of different classifiers and eliminate their misclassification (sets of misclassified patterns do not necessarily overlap). Similar concept exists in clustering domain, that is, cluster ensemble. Different clusterings of the same data set are combined to obtain final clustering of improved quality (see Vega-Pons & Ruiz-Shulcloper, 2011, for an extensive survey of various combination methods and techniques). The idea of combination can be straightforwardly extended from classification and clustering also to the problem of image segmentation, because the segmentation method can be considered as a special kind of classifier or clustering method. See, for example, Franek *et al.* (2011) and Vega-Pons *et al.* (2011) for application of cluster ensembles to image segmentation.

In our case we have to decide which segmentation methods to combine and what method of combination to use. Generally it holds that the input set of methods (results, clustering or classifiers) has to be sufficiently diverse to achieve the best possible result of combination but at the same time if there are frequently failing methods included the final combination is spoiled (see e.g. Sharkey, 1996, in context of neural networks classification). In terms of image segmentation we need to combine such segmentation methods which perform very well generally, do not fail too often and their results differ in important details (boundaries). We use evaluation results from previous section to achieve this. The best three average methods form the input set to combination in each modality. They perform the best from the studied set of methods, do not fail to often and their results are sufficiently diverse thanks to different fundamentals of each segmentation method. The combination of more than three methods was found dissatisfactory because the input results were more frequently bad which negatively influenced the output of combination. Concerning combination method the majority vote is used. Therefore the pixel of an input image is labelled as foreground if at least two of the three methods label it as foreground. Otherwise it is background. We show that even such uncomplicated combination method can achieve considerable improvement of the image segmentation.

Results of segmentation combination are thus generated for every image in each modality using the three best average methods. It is Rousson–Deriche approach, Felzenszwalb's method and MS for SEM modality, MS, Felzenszwalb's method and multiscale normalized cut in greyscale for UV modality, and finally MS, multiscale normalized cut in RGB and K-means in RGB for VIS modality (see Table 4). The aim now is to compare the results of the combination to the best average method. Again quality indices are necessary to ensure objective evaluation. We compute 10 indices already used in previous evaluations for every image and compare them to those of the best average segmentation methods (Rousson–Deriche approach for SEM and MS for UV and VIS modalities). We use statistical evaluation with hypothesis testing to determine which of the two is better. The Wilcoxon signed-rank test (Wilcoxon, 1945) is used as good trade-off between plain sign test (which does not consider the magnitude of differences at all) and *t*-test (which considers the magnitude in much stronger way and also the stronger assumptions have to be met). Level of significance is set to 0.05.

Combination is statistically significantly better than the best average method in SEM and UV modality. In VIS modality the situation is little bit more complicated. Only 4 out of 10 indices claim that the combination is significantly better. Conversely two indices claim that the best average method is significantly better. The rest stays rather undecided. Thus, it cannot be decided which of the two approaches is better in VIS modality. If we compare combination to the second best average method (which is multiscale normalized cut in RGB) situation

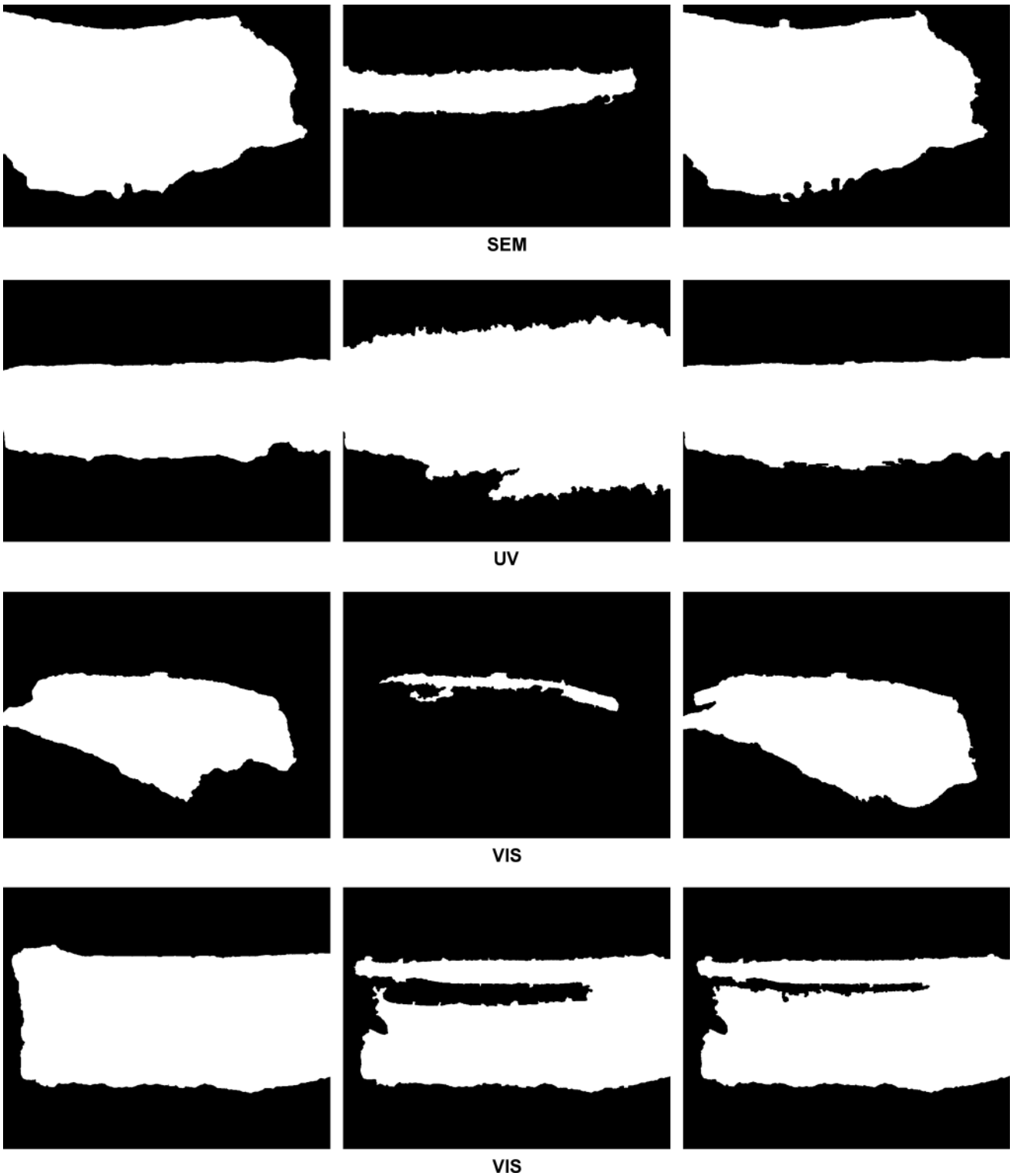

Fig. 8. Demonstration of improvement using combination of segmentation methods compared to the best average method. In each triplet in rows there is GT mask (left column), result of the best average method (middle column, GC_RD in SEM and MS in UV and VIS) and result of combination (right column). Last triplet corresponds to the images in Figure 3. Combination there is certainly better than MS's result. However even better result can be achieved with pure RG in this case as is shown in Figure 3.

gets much clearer. Combination is significantly better in this case. For these reasons the choice of combination approach is appropriate even for VIS modality thanks to its robustness.

Visual evaluation was done as well to support the findings from statistical testing. Combination pays off also from this point of view. It is usually better than the best average methods in SEM and UV modality. In UV the difference is even more prominent and it is easy to see how combination of several segmentation methods amend inaccuracies of MS algorithm as the best average method (see Fig. 8 for examples). Perhaps surprisingly the same holds for VIS modality. The results of combination are often more plausible. In those cases where MS is better than combination, the difference is often minute. In the opposite cases difference between combination and MS is much larger and combination resembles GT more accurately.

Conclusion is that combination of several segmentation methods can significantly outperform use of single (even the best average) segmentation method. This clearly holds for SEM and UV modality but also in case of VIS it is safe to use combination approach. Combination there is almost identical or only slightly worse than the best average method in vast majority of cases and occasionally it gives much better results. See Figure 8 for examples of the results of segmentation combination.

Conclusion

In this paper the performance of several segmentation methods on images of microscopic samples in three different modalities was analysed. The set of 10 quality indices was used to achieve evaluation as objective as possible. We showed that there was no single segmentation method which significantly outperformed the others in the studied set. The average performance of the methods was then evaluated with conclusion that MS algorithm performed the best and can be considered the best segmentation method on average. Concerning other methods in the studied set, the recommendations on their usability in different situations were proposed. Finally, it was demonstrated that performance of even the best average method could be further improved by using combination of several segmentation methods. This was confirmed with statistical tests. Moreover, the applicability of the evaluation results on different but related biological data was shown.

Acknowledgements

We would like to thank Janka Hradilová and David Hradil from ALMA laboratory (joint workplace of the Academy of Fine Arts in Prague and the Institute of Inorganic Chemistry of the Academy of Sciences) for providing the cross-section images and invaluable insights to the field of material research and art restoration. We thank Jan Flusser for his comments and ideas and Jiří Dvořák for invaluable help with statistical evaluation and for his advice. Thanks also go to Jan Cendelín from Faculty of Medicine in Pilsen and to Martin Čapek from Institute of Physiology AS CR in Prague for provision of biological image. The work has been supported by the Czech Science Foundation under project GAP103/12/2211. The work of Jan Cendelín was supported by Charles University Research Fund under project P36.

References

Arbelaez, P., Maire, M., Fowlkes, C. & Malik, J. (2011) Contour detection and hierarchical image segmentation. *IEEE T. Pattern Anal.* **33**, 898–916.

Beneš, M., Zitová, B., Blažek, J., Hradilová, J. & Hradil, D. (2011) Removing the artifacts from artwork cross-section images. In *Proceedings of the IEEE Image*. pp. 3537–3540. IEEE, Brussels, Belgium.

Chan, T.F. & Vese, L. A. (2001) Active contours without edges. *IEEE T. Image Process.* **10**, 266–277.

Cohen, J. (1960) A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* **20**, 37–46.

Comaniciu, D. & Meer, P. (2002) Mean shift: a robust approach toward feature space analysis. *IEEE T. Pattern Anal.* **24**, 603–619.

Cour, T., Benezit, F. & Shi, J. (2005) Spectral segmentation with multiscale graph decomposition. In *Proceedings of the CVPR IEEE*. Vol. **2**, pp. 1124–1131. IEEE, San Diego, CA, USA.

Daněk, O. (2012) *Graph cut based image segmentation in fluorescence microscopy*. Ph.D. thesis, Masarykova univerzita, Brno.

Dempster, A.P., Laird, N. M. & Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Stat. Soc. B Met.*, **39**, 1–38.

Dey, V., Zhang, Y. & Zhong, M. (2010) A review on image segmentation techniques with remote sensing perspective. In *Proceedings of the ISPRS TC VII Symposium-100 Years ISPRS*, 5–7 July 2010, Vol. XXXVIII, Part 7A, pp. 31–42 (eds. by W. Wagner & B. Székely), IAPRS, Vienna, Austria.

Dice, L.R. (1945) Measures of the amount of ecologic association between species. *Ecology* **26**, 297–302.

Doyle, W. (1962) Operations useful for similarity-invariant pattern recognition. *J. ACM* **9**, 259–267.

Dwork, C., Kumar, R., Naor, M. & Sivakumar, D. (2001) Rank aggregation methods for the web. In *Proceedings of the WWW*. pp. 613–622. ACM, Hong Kong.

Felzenszwalb, P.F. & Huttenlocher, D. P. (2004) Efficient graph-based image segmentation. *Int. J. Comput Vision* **59**, 167–181.

Fowlkes, E.B. & Mallows, C.L. (1983) A method for comparing two hierarchical clusterings. *J. Am. Stat. Assoc.* **78**, 553–569.

Franek, L., Abdala, D.D., Vega-Pons, S. & Jiang, X. (2011) Image segmentation fusion using general ensemble clustering methods. In *Lect. Notes Comput. Sc.* **6495**, 373–384. Springer.

Freixenet, J., Mu noz, X., Raba, D., Martí, J. & Cufí, X. (2002) Yet another survey on image segmentation: region and boundary information integration. In *Proceedings of the Lecture Notes in Computer Science*. Vol. **2352**, pp. 408–422. Springer, Berlin-Heidelberg, Germany.

Glasbey, C.A. (1993) An analysis of histogram-based thresholding algorithms. *CVGIP-Graph. Model. Im.* **55**, 532–537.

Grady, L. (2006) Random walks for image segmentation. *IEEE T. Pattern Anal.* **28**, 1768–1783.

Hamming, R.W. (1950) Error detecting and error correcting codes. *Bell Syst. Tech. J.* **29**, 147–160.

Huang, L.-K. & Wang, M.-J.J. (1995) Image thresholding by minimizing the measures of fuzziness. *Pattern recogn.* **28**, 41–51.

Huang, Q. & Dom, B. (1995) Quantitative methods of evaluating image segmentation. In *Proceedings of the IEEE Image*. Vol. **3**, pp. 53–56. IEEE, Washington, DC, USA.

Hubert, L. & Arabie, P. (1985) Comparing partitions. *J. Classif.* **2**, 193–218.

Jaccard, P. (1912) The distribution of the flora in the alpine zone. *New phytol.* **11**, 37–50.

Kapur, J., Sahoo, P.K. & Wong, A. (1985) A new method for gray-level picture thresholding using the entropy of the histogram. *Comput. Vision Graph.* **29**, 273–285.

Kittler, J., Hatef, M., Duin, R.P. & Matas, J. (1998) On combining classifiers. *IEEE T. Pattern Anal.* **20**, 226–239.

Kittler, J. & Illingworth, J. (1986) Minimum error thresholding. *Pattern recogn.* **19**, 41–47.

Kohli, P., Ladický, L. & Torr, P.H. (2009) Robust higher order potentials for enforcing label consistency. *Int. J. Comput Vision* **82**, 302–324.

Kuncheva, L., Hadjitodorov, S. & Todorova, L. (2006) Experimental comparison of cluster ensemble methods. In *Proceedings of the FUSION*. pp. 1–7. IEEE, Florence, Italy.

Li, C. & Tam, P. K.-S. (1998) An iterative algorithm for minimum cross entropy thresholding. *Pattern recogn. Lett.* **19**, 771–776.

MacQueen, J. (1967) Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symp. Math. Stat.* Vol. **1**, pp. 281–297. Berkeley, CA, USA.

Malcolm, J., Rathi, Y. & Tannenbaum, A. (2007) A graph cut approach to image segmentation in tensor space. In *Proceedings of the CVPR IEEE*. pp. 1–8. IEEE, Minneapolis, MN, USA.

Martin, D., Fowlkes, C., Tal, D. & Malik, J. (2001) A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *IEEE I. Conf. Comp. Vis.* Vol. **2**, pp. 416–423. IEEE, Vancouver, BC.

Meilă, M. (2007) Comparing clusterings: an information based distance. *J. Multivariate Anal.* **98**, 873–895.

Mirkin, B. (1996) *Mathematical Classification and Clustering*. Kluwer Academic Publishers, Dordrecht, Netherlands.

Otsu, N. (1975) A threshold selection method from gray-level histograms. *Automatica* **11**, 23–27.

Pal, N.R. & Pal, S.K. (1993) A review on image segmentation techniques. *Pattern recogn.* **26**, 1277–1294.

Pihur, V., Datta, S. & Datta, S. (2009) Rankaggreg, an R package for weighted rank aggregation. *BMC Bioinformatics* **10**, 62.

Pratt, W.K. (2007) *Digital Image Processing: PIKS Scientific Inside*. 4th edn. Wiley-Interscience, Hoboken, NJ, USA.

Prewitt, J. & Mendelsohn, M.L. (1966) The analysis of cell images. *Ann. N.Y. Acad. Sci.* **128**, 1035–1053.

Rand, W.M. (1971) Objective criteria for the evaluation of clustering methods. *J. Am. Stat. Assoc.* **66**, 846–850.

Ridler, T. & Calvard, S. (1978) Picture thresholding using an iterative selection method. *IEEE T. Syst., Man Cyb.* **8**, 630–632.

Rijsbergen, C.J.V. (1979) *Information Retrieval*. 2nd edn. Butterworth-Heinemann, Newton, MA, USA.

Rosenfeld, A. & De La Torre, P. (1983) Histogram concavity analysis as an aid in threshold selection. *IEEE T. Syst., Man Cyb.* **SMC-13**, 231–235.

Rother, C., Kolmogorov, V. & Blake, A. (2004) Grabcut: interactive foreground extraction using iterated graph cuts. *ACM T. Graphic* **23**, 309–314.

Rousson, M. & Deriche, R. (2002) A variational framework for active and adaptative segmentation of vector valued images. In *Proceedings of the MOTION*. pp. 56–61. IEEE, Orlando, FL, USA.

Shanbhag, A.G. (1994) Utilization of information measure as a means of image thresholding. *CVGIP-Graph. Model. Im.* **56**, 414–419.

Sharkey, A.J.C. (1996) On combining artificial neural nets. *Connect. Sci.* **8**, 299–314.

Shi, J. & Malik, J. (2000) Normalized cuts and image segmentation. *IEEE T. Pattern Anal.* **22**, 888–905.

Sluimer, I., Prokop, M. & van Ginneken, B. (2005) Toward automated segmentation of the pathological lung in CT. *IEEE T. Med. Imaging* **24**, 1025–1038.

Strehl, A. & Ghosh, J. (2003) Cluster ensembles – a knowledge reuse framework for combining multiple partitions. *J. Mach. Learn. Res.* **3**, 583–617.

Tao, W., Jin, H., Zhang, Y., Liu, L. & Wang, D. (2008) Image thresholding using graph cuts. *IEEE T. Syst., Man Cyb.* **38**, 1181–1195.

Tsai, W.-H. (1985) Moment-preserving thresolding: a new approach. *Comput. Vision Graph.* **29**, 377–393.

Vega-Pons, S., Jiang, X. & Ruiz-Shulcloper, J. (2011) Segmentation ensemble via kernels. In *Proceedings of the ACPR*. pp. 686–690. IEEE, Beijing, China.

Vega-Pons, S. & Ruiz-Shulcloper, J. (2011) A survey of clustering ensemble algorithms. *Int. J. Pattern Recogn.* **25**, 337–372.

Vinh, N.X., Epps, J. & Bailey, J. (2009) Information theoretic measures for clusterings comparison: is a correction for chance necessary? In *Proceedings of the ICML*. pp. 1073–1080. ACM, Montreal, Canada.

Warrens, M.J. (2008) On the equivalence of Cohen's kappa and the Hubert-Arabie adjusted Rand index. *J. Classif.*, **25** 177–183.

Wilcoxon, F. (1945) Individual comparisons by ranking methods. *Biometrics Bull.* **1**, 80–83.

Yen, J.-C., Chang, F.-J. & Chang, S. (1995) A new criterion for automatic multilevel thresholding. *IEEE T. Image Process.* **4**, 370–378.

Zack, G., Rogers, W. & Latt, S. (1977) Automatic measurement of sister chromatid exchange frequency. *J. Histochem. Cytochem.* **25**, 741–753.

Zhang, H., Fritts, J.E. & Goldman, S.A. (2008) Image segmentation evaluation: A survey of unsupervised methods. *Comput. Vis. Image Und.* **110**, 260–280.

Appendix: Additional material to Section single best segmentation method

This appendix contains additional material to Section single best segmentation method. Deeper analysis of distribution image segmentation methods among the best methods selected by quality indices is presented here.

The two most frequent segmentation methods in SEM modality are Felzenszwalb's method (GC_FH) and RG (with parameter equal to 5 – RG_5) with 12 occurrences out of 89 possible each among the best methods. They are followed by Mean Shift algorithm (MS) and Rousson–Deriche approach (GC_RD). The rest is featured in Figure A1. Nineteen methods out of 43 have zero number of occurrences. Several important conclusions can be made based on this histogram. First and the most important, there is no segmentation method which clearly outperforms the others (12 occurrences for MS

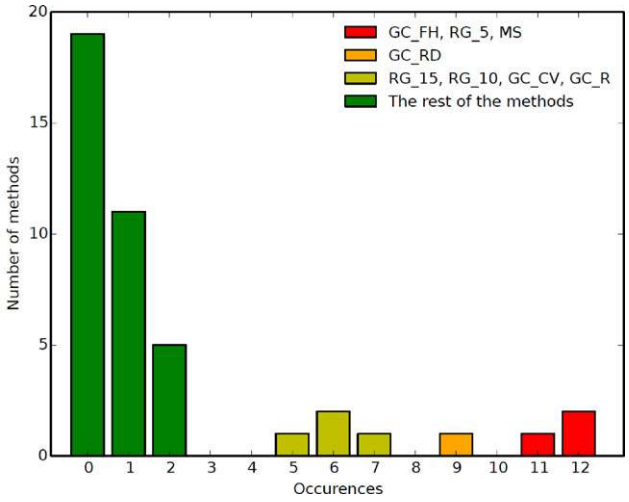

Fig. A1. Graph of number of occurrences among the best segmentation methods for each method in SEM modality. Felzenszwalb's method, region growing (with parameter 5) and Mean Shift algorithm are the most successful methods. The majority of methods has however two occurrences at most.

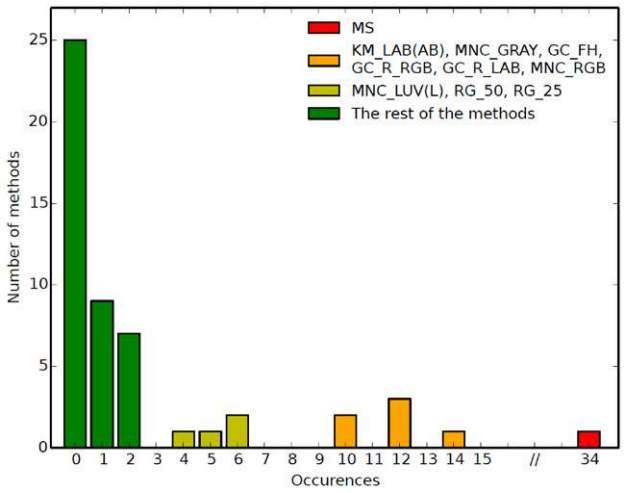

Fig. A2. Graph of number of occurrences among the best segmentation methods for each method in UV modality. Mean Shift is by far the most successful method with colourspace versions of multiscale normalized cut, K-means, GrabCut and Felzenszwalb's method behind.

out of 89 are not sufficient enough). Second, RG methods are quite successful, especially with smaller values of the parameter. Finally, thresholding algorithms do not perform well individually (though there are 16 occurrences in total for thresholding).

The situation in UV modality is rather different. MS is clearly the most successful method. It is better than any other method

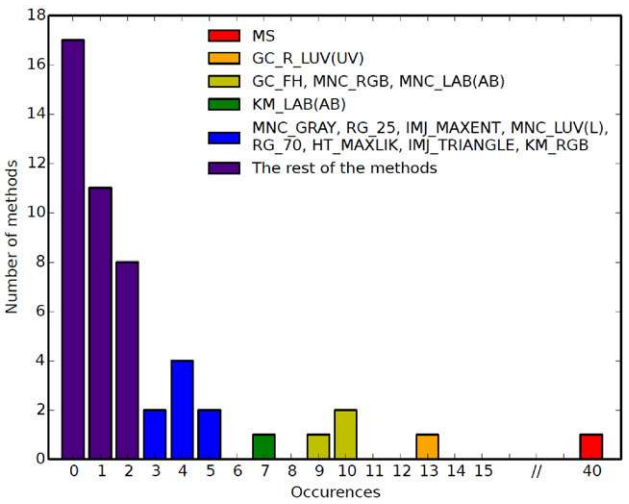

Fig. A3. Graph of number of occurrences among the best segmentation methods for each method in VIS modality. Mean Shift is a method with the most occurrences. GrabCut follows with large gap and Felzenszwalb's method and colourspace variations of multiscale normalized cut are behind. Lots of methods have two occurrences at most.

in 34 cases out of 148 (the total number of UV images). K-means (KM, in AB subspace of LAB colourspace), GC_FH, GrabCut (GC_R, in RGB) and multiscale normalized cut (MNC, in greyscale) follow with 12–14 occurrences. Half of the methods (25 out of 52 precisely) are not among the best methods in at least one case. The rest is displayed in Figure A2. As in SEM modality there is no clear winner which could be mechanically used for segmentation of UV images. MS is indeed very successful, but it outperforms the others only in quarter of cases which is not sufficient. Surprisingly, GC_RD and Chan–Vese approach (GC_CV) fail completely with one and zero occurrences respectively. RG does not perform that well as in SEM modality. Thresholding methods represent only a complement to more successful methods.

Finally, the results for VIS modality are presented. MS stays the most frequent among the best methods for each image with 40 occurrences out of 148 possible. Versions of GC_R and MNC in various colourspaces and GC_FH follow with roughly 10 occurrences. The rest can be seen in histogram in Figure A3. Seventeen methods out of 50 are not selected as the best method at least once. The conclusions for UV modality hold also here. MS outperforms the other methods in lots of cases, nevertheless not in the significant majority. GC_RD and GC_CV approaches fail again. RG is not very successful and where it is, the bigger parameter values are used. In contrast to UV, thresholding methods represent alternative to more sophisticated methods. They are selected as the best ones for 31 images in total.