Elektrotehniški vestnik 74(1-2): 13–18, 2007
Electrotechnical Review, Ljubljana, Slovenija

Objective assessment of image segmentation algorithms

Dušan Heric, Božidar Potočnik

University of Maribor, Faculty of Electrical Engineering and Computer Science, Smetanova 17, 2000 Maribor, Slovenia
E-mail: dusan.heric@uni-mb.si

Abstract. An image segmentation algorithm quality assessment is usually visual, subjective and based on non-standard measures. Non-transparencies among assessment results present a problem in objective evaluation of segmentation algorithms accuracy. This paper introduces an image processing assessment tool (IPA-tool) for objective assessment with standard accuracy measures and reference annotations. These annotations are considered as a ground truth. It is difficult to get the ground truth of real-world images in practice. The IPA-tool, therefore, supports also a mean observer mechanism which creates the ground truth from several annotations. This paper proposes an assessment chain, mean observer mechanism and supported assessment measures inside the IPA-tool.

Key words: Image processing, Objective assessment, Mean observer

Objektivno ocenjevanje slikovnih segmentacijskih algoritmov

Povzetek. Ocenjevanje segmentacijski postopkov je pogosto izvedeno vizualno, subjektivno, z nestandardnimi in težko primerljivimi merami. Takšna neskladnost med ocenami pa je pereč problem, ko želim lasten postopek dobro oceniti in ga primerjati z že obstoječimi segmentacijskimi postopki.

V delu je predstavljeno orodje IPA, ki je namenjeno za objektivno ocenjevanje segmentacijskih postopkov in je podprto s standardnimi merami za ocenjevanje natančnosti. Ocenitev temelji na primerjavi računalniških označb z referenčnimi. Slednje natančno opisujejo segmentacijska območja, kar pa je v praksi težko zagotoviti. Zato orodje IPA podpira mehanizem za izračun povprečnega ocenjevalca iz množice referenčnih označb. V delu so predstavljeni ocenjevalna veriga, mehanizem za ocenjevanje povprečnega ocenjevalca in podprte mere za ocenjevanje natančnosti segmentacijskih postopkov z orodjem IPA.

Klučne besede: obdelava slik, nepristrano ocenjevanje, povprečni ocenjevalec

1 Introduction

Segmentation algorithms split an image into regions. These are either a) analyzer additional information or b) inputs into simulation environments. In both cases high accuracy is required. Unfortunately, authors of segmentation algorithms usually use custom-made measures or visual estimation. Visual assessment does not consider intra- and inter-expert variability, while custom-made measures are non-transparent. An additional problem is deficiency of objective assessment methods [10]. Therefore, authors frequently expose the need for measures and tools for evaluation/comparison of segmentation algorithms [12]. The segmentation accuracy assessment depends on the ground truth availability [2]. If the ground truth is known, only assessment measures are calculated. However, situations of missing the ground truth are permanent in real systems (e.g. medical images) where a) the exact regions positions are non-detectable due to noise/artefacts and b) optimum-quality acquisition is impossible [3]. In such situations the ground truth is usually constructed from several reference annotations. The major drawback are high variations in annotations or inter-experts variability which can be up to 15% [8]. Unfortunately, such assessment approach is inevitable in a) where fiducial markers can not be placed in known areas or b) if it is difficult to generate synthetic images that capture the complexity and deformability of original images. The proposed IPA-tool implements the ground truth mechanism based on several reference annotations.

This paper introduces a robust and reliable accuracy assessment of segmentation algorithms, novel

Received 11 April 2006
Accepted 6 December 2006

mean observer mechanism and supported measures implemented in the IPA-tool. In comparison with available assessment tools, which mostly do not support mean observer mechanisms, our tool evaluates segmentation algorithm with respect to several reference annotations.

The paper is organized as follows: Section 2 brings a short overview of the related works. Section 3 introduces the proposed assessment protocol. Sections 4 and 5 present a mean observer mechanism and assessment measures, respectively. Section 6 evaluates the segmentation algorithm, while Section 7 draws conclusions.

2 Related work

Objective segmentation accuracy assessment methodologies came in sight at the beginning of 1980s [16]. Since then, several quality assessments measures have been proposed which are classified either as or empirical methods [17]. The focus of the formers is on the segmentation algorithms working (e.g. expressions like good detection, good localization in [19]). The latter use disparity between the segmentation results and the ground truth. Several volume spatial overlap measures have been introduced in [11, 14]. The hausdorff distance and its derivatives [20] represent measures of a spatial distance between two sets of points. These measures are denoted as surface or contour-based measures. Recently, a hybrid measure of the boundary measurements and spatial overlap has been introduced in [13]. Generally, all these measures are spatial disparity measures between the segmented and the reference objects. Regions/bodies are here denoted as objects while voxels/pixels as points. It should be noted that some measures do not contain relevant information and cannot give a satisfactory assessment of segmentation quality [15]. It is also important to know the expected/desired quality of the segmentation algorithm.

The measures proposed in [4] are useful if ground truth segmentation results are available, while [5] evaluates segmentation methods quantitatively, if ground truth segmentation results are unavailable. Representative assessment of the image segmentation accuracy and expert quality are in [1], where reference annotations are calculated from a group of expert segmentations.

Various alternative methods, especially in medicine, have been sought to allow for a statistical assessment [21]. A useful method is to construct phantoms, either physically or digitally [22]. Nevertheless, sophisticated phantoms and image synthetic models may frequently not yield images with a full range of characteristics such as intensity inhomogeneous, noise, partial volume artefacts, and pathologic anatomic variability. Therefore, it is reasonable to evaluate the segmentation accuracy comparing computer segmented annotations (detected) and experts annotations.

3 Assessment Protocol

The segmentation accuracy assessment is a degree for a measurement correctness [18]. The object accuracy assessment is defined as a difference between the computed values and the ground truth. The proposed protocol is based on the ground truth and disparity measures calculation. The protocol is based on a language of set theory and supports a) plane (surface) assessment when volume information is unavailable and b) volume assessment when volume information is available. The protocol chain consists of two processes: a) mean observer creation and b) disparity measures calculation.

3.1 Mean observer

The mean observer is an approximation for the ground truth in our assessment protocol, calculated from a set of annotations.

Let I be an image, p a point (x, y) in image I, S_p a set of points p ($S_p = \{x, y\}\,; x, y \in \Re$), $A = (S_p, I)$ an annotation, and S_a a set of annotations in image I. The ground truth function signature is prescribed as g_t. The function implementation is not considered from the assessment protocol point of view. The result of g_t is the ground truth which is treated as a plane mean expert E_p:

$$E_p = g_t(S_a); \quad E_p \in S_p. \tag{1}$$

Plane expert prescription can be used only for 2D images. If volume information is available, it is better to make a volume assessment. A volume expert is structured as a body or as a partial body. The partial body is a cross-section through an object composed into the 3D-world with the width equal to the slice thickness.

Let v be a voxel (x, y, z), P a patch defined by three non-collinear voxels and bounded by straight lines between them, and P_v a set of patches or surfaces. The volume annotation is defined as $A_v = (P_v, I)$. Body volume expert E_v structure is a set of volume annotations. Additional information is necessary only for the partial body definition. Let S_a be a set of experts annotations and D_v volume data. The volume data includes the volume position for each annotation. Partial body volume mean expert signature g_{vi} is defined as:

$$E_v = g_{vi}(E_p, D_v). \quad (2)$$

The output of ground truth calculation is the volume mean expert. This expert can be considered as a ground truth for the volume and for plane segmentation results.

3.2 Measure calculation

The volume or plane disparity measures are calculated between the ground truth and the segmented annotations. Specialty is a partial body assessment where disparity measures are calculated for each cross-section individually. The result is thus presented as a list of plane disparity measures.

4 IPA-tool mean-observer mechanism

Reliable and quality evaluation of the segmentation algorithm accuracy is a challenging task. To assure an objective assessment, it is necessary to known the ground truth.

The IPA-tool implements the ground truth calculation method named BigSmall regions (BS-method). The method follows the idea from [1] and is based on the set of expert annotations. It is suited for both the convex and the concave regions (objects).

4.1 BigSmall regions (BS-method)

The problem of ground-truth calculation is formulated as searching of the mean or average among all the experts objects (annotations). The computational framework for the two-dimensional regions is outlined in the sequel, while extension to the three-dimensional volumes is relatively straightforward. Firstly, the mean observer between two manual annotations is described and then followed by an extension to an arbitrary number of annotations.

Let A and B be two intersecting regions. The BS-method uses the hypothesis that the mean point between regions lies in the middle of the straight line connecting two nearest regions' points, where the first point belongs to the contour of region A and the second to the contour of region B.

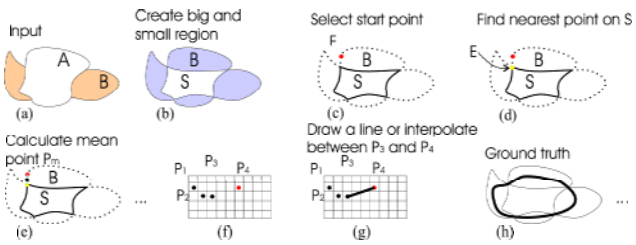

Figure 1. BigSmall region creation.

The algorithm of ground truth calculation between two experts annotations consists of six steps (see Figure 1):

1. Let B_R and S_R denote a big and small region, respectively. The big region is a union of regions A and B, while the small region is an intersection between both regions (Figure 1 (a) and (b)):

$$B_R = A \cup B, S_R = A \cap B$$

2. Select an arbitrary starting point F on the contour of the big region B_R (Figure 1 (c)).
3. Find the nearest point E on the small region contour. Use the Euclidian distance (Figure 1 (d)).
4. Calculate a mean point P_{mean} between points F and E (Figure 1 (e)).
5. If P_{mean} is not a neighbouring point of the previous mean point $P_{mean-old}$, then connect points P_{mean} and $P_{mean-old}$ with a straight line (see Figure 1 (f) and (g)). The contour continuity is preserved in this way.
6. Select the next point on the big region boundary.
7. Repeat steps from (3) to (6) until all points on the big region boundary are processed (see Figure 1 (h)).

The following lines formalize the BS-method described above:

1. $B_c = contour(B_R), S_c = contour(S_R), E_p = \{\}$
2. $\forall s \in B_c \quad \min_t \|s - t\|_2 \quad \wedge \quad t \in S_c$
3. $E_p = E_p \cup \{\frac{s+t}{2}\}$,

where E_p is the ground truth between two experts. The function *contour* returns the boundary of the input region. Extension to the arbitrary number of experts is relatively straightforward. The new mean-observer mechanism with this extension is defined as:

1. Let E_{pi} denotes E_p calculated between two arbitrary experts, $i = 0$.
2. Let B_R and S_R denote big and small regions between E_{pi} and unapplied expert annotation A.
3. $B_c = contour(B_R), S_c = contour(S_R)$
   $E_c = contour(E_{pi}), A_c = contour(A), E_{pi+1} = \{\}$
4. $\forall s \in B_c \quad \min_t \|s - t\|_2 \quad \wedge \quad t \in S_c$
5. $s_i = \{s\} \quad \wedge \quad t_i = \{t\}$

6. $E_{pi+1} = E_{pi} \cup \{\frac{i \cdot s_i + t_i}{i+1}\}, \quad i = i + 1$
7. Steps (2) to (6) are repeated until all experts are taken in the ground truth calculation.

Major BS-method benefits are: a) low time complexity, b) low global sensitivity to the outliers because the neighbour pixels compensate outliers positions, and c) designed for both the convex and the concave regions. On the other hand, local sensitivity is relatively high. Thus, an averaging filter could be used for reduction of such variations. The question is whether local sensitivity controlling has any sense. Nonetheless, there is no need for annotations to differentiate evidently between experts.

The presented mean-observer mechanism is fast, reliable and treats each expert annotations with equal probability.

5 IPA tool-supported measures

The proposed IPA tool supports empirical disparity measures which can be classified into four groups: a) region-based measures - the expert and calculated objects regions are compared [6], b) contour-based measures - the experts and calculated objects boundaries are compared [7], c) body-based measures - the expert and calculated volumes are compared and d) surface-based measures - the expert and calculated objects surfaces are compared to each other.

In the sequel, the IPA tool-supported disparity measures for measuring the difference (accuracy) between the calculated and mean observer annotations are described.

5.1 Ratio R1 and R2

Ratios R1 and R2 are measures considering spatial objects properties by a pair-wise comparison of two binary images. Images are analyzed in a point-by-point manner. Ratios are based on the intersection between the segmented object and the reference object. The ratio R1 is thus defined as the ratio between the intersection and reference object, while the ratio R2 is defined as the ratio between the intersection and segmented region. Both ratios give 1 for perfect agreement and 0 for complete disagreement between calculated and reference objects. For example, if ratios R1 and R2 are both 1, then this signifies the perfect alignment of the calculated object with the reference object, and vice verse. In another example with ratio R1 at 0.6 and ratio R2 at 0.3 this alignment is poor. The ratio R1 explains that 60% of the reference object is fitted by the calculated object and only 30% of the calculated object fits the reference object. However, the ratios R1 and R2 depend on the size and object complexity. Their main drawback is the penalty function. In this function small objects get much higher penalty than the large for the same boundary or surface errors.

5.2 Mean absolute distance (MAD)

The mean absolute distance determines an average between the two contours/surfaces. It is defined according to the following equation:

$$MAD = \frac{1}{2n}\sum_{i=1}^{n} d\left(a_i, B\right) + \frac{1}{2m}\sum_{i=1}^{m} d\left(b_i, A\right), \quad (3)$$

where A and B are curve/surface, and a_i and b_i stand for the curve/surface points. The distance d is a minimal Euclidean distance between the point on the object boundary/surface A, and the object boundary/surface B. The main advantages of MAD measure are: a) low sensitivity to the outliers and b) independence of the object size.

5.3 Hausdorff distance (HD)

The Hausdorff distance measures the maximum distance between two objects boundaries/surfaces. The distance from the boundary/surface of calculated object to the nearest point on the boundary/surface of reference object is measured, and vice versa. It is determined in three steps: 1) the Euclidian distances between the boundary/surface of the reference object to the boundary/surface of the calculated object is calculated, 2) the shortest distance from the boundary/surface of the reference object to the calculated object is kept for each point, and 3) the largest distance between all distances is taken:

$$HD = \max\left(\mathrm{h(A,B)}, \mathrm{h(B,A)}\right)$$

$$h(A,B) = \max_{\mathrm{a}\in\mathrm{A}}\min_{\mathrm{b}\in\mathrm{B}}\mathrm{d(a,b)} = \max_{\mathrm{a}\in\mathrm{A}}\mathrm{d(a,B)}$$

where h is a directed Hausforff distance. This measure is sensitive to outliers and does not reflect the actual situation along the whole boundary/surface. If segmentation accuracy remained within the certain limits (are prescribed), this measure would be the metrics of choice. The directed measure is non-symmetric.

5.4 Spherical distance (SD)

In the IPA-tool, our novel body/volume based measure–spherical distance, interpreted as the distance between two surfaces, is introduced and implemented. The measure defines the distance between

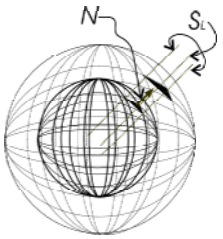

Figure 2. Two spheres - the smaller and the larger one. The arrow denotes the normal to patch on the smaller sphere, while the dotted straight lines with the same direction as normal run through the patch edge points.

two objects surfaces with respect to the normal direction of their patches and their distribution. This measure is our original contribution.

Figure 2 depicts two spheres - the small and the large one, where the small sphere is inside the larger one. The black triangles represent two selected patches. The arrow is the normal on the patch of the small sphere and the dotted straight lines through the patch edge points have the same direction as the normal.

The spherical distance is calculated in three steps:

1. Calculate normals for all patches on evaluated surface (e.g. calculated/segmented).

2. Calculate the Euclidean distance between the patch edge point and the nearest point (on reference surface) determined by the straight line piercing through a reference surface. the straight line, which pierces the patch edge point, has the same direction as the normal of the patch (see Figure 2). A distance e_d between the patch edge point from A and surface B is defined as

$$e_d(T_1, B) = ||T_1 - T_2||_2; T_1 \in B \wedge T_2 \in S_L \cap A$$

where T_2 is the nearest point where the straight line S_L pierces the reference surface B. Distance E_d between surfaces A and B is calculated as follows

$$E_d(A, B) = \sum_{T_1 \in A} e_d(T_1, B).$$

3. The spherical distance SD is calculated as the minimum between the averages of the Euclidian distances $E_d(A, B)$ and $E_d(B, A)$:

$$SD = \min\{\frac{1}{n} E_d(A, B), \frac{1}{m} E_d(B, A)\},$$

where n and m are the number of patches on the surfaces A and B, respectively.

The main measure advantages are: a) low sensitivity to the outliers, b) independence of the object size, and c) measure symmetry.

6 Example of segmentation algorithm assessment

The proposed IPA tool was used in the SimBio project [9] for the assessment of image segmentation/registration routines. This tool was tested on a set of high-quality static MR human knee joint images with dimensions of 512x512 pixels, acquired with T1 weighted sequence. Slice thickness was 2 mm and an effective pixel size was 0.4 mm. Figure 3 depicts the example of annotated MR testing image.

Figure 3. Example of annotated knee joint: a) expert readings overlaid on the original MR image; b) computer-detected knee structures.

The patient MR knee image sequences were manually annotated by an orthopaedic surgeon (i.e. ground truth references) and, afterwards, processed by the segmentation algorithm–Vreglocal3d [9]. The evaluation was done on three bones (i.e. femur, tibia and patella) and their corresponding cartilages. The accuracy of bones and cartilages detection was measured by ratios R1 and R2, Hausdorff distance (HD), Mean absolute distance (MAD), and Spherical distance (SD). The disparity measures calculated for all patients' image sequences are presented in the sequel. The average calculated ratio R1 is 0.82, the standard deviation is 0.15, the minimum average ratio is 0.57 and the maximum average ratio is 0.94. The average ratio R2 is 0.73, the standard deviation is 0.11, the minimum is 0.46 and the maximum is 0.94. The average MAD distance is 0.52 mm, the standard deviation is 0.34 mm, the minimum is 0.28 mm and the maximum is 0.75 mm. The average HD distance is 2.78 mm, the standard deviation is 1.89 mm, the minimum is 1.18 mm and the maximum is 4.71 mm. The average SD distance is 1.69 mm, the standard deviation is 3.21 mm, the minimum is 1.18 mm and the maximum is 4.91 mm. These results point out the rough assessment of the segmentation algorithm. Evident segmentation inaccuracies are indicated.

7 Conclusion

The IPA tool, assessment protocol scheme, novel mean-observer method and disparity measures with

our novel spherical distance were presented in this paper.

The assessment protocol is a prescription whose major preferences are a) transparent plane and volume assessment, and b) simple integration with other metrics. The supported disparity measures were classified into four classes with respect to plane and volume information. The IPA tool overcomes the assessment problems with one biased expert using mean observer BS method. This mechanism ensures reliable results, because each expert has equal weight in ground truth calculation. Specialities of the proposed IPA tool are: a) original mean-observer mechanism (BS-method) and b) original volume spherical distance. The assessment results are meaningless for the non-experts, because it is important to known and correctly interpret the most appropriate measures for the segmentation algorithm quality assessment! Our future research will be towards development novel statistical measures and extension of the mean-observer mechanism, with the outliers having low influence on the final mean observer.

8 References

[1] K.S. Warfield, H.K. Zou, R. M. Kaus, M.W. Wells, Simultaneous validation of image segmentation and assessment of expert quality, *International Symposium on Biomedical Imaging IEEE*, 2002, pp. 1494-1498.

[2] L. P. Correia, F. Pereira, Objective evaluation of video segmentation quality, IEEE Transaction on Image Processing, Vol. 12, No. 2, 1993, pp. 850-863.

[3] C. J. Russ, The Image Processing Handbook, second ed., CRC Press Boca Raton, 1995.

[4] E. C. Erdem, B. Sankur, Performance evaluation metrics for object-based video segmentation, *Im Proc. X European Signal Processing Conference*, Vol. 2, 2000, pp. 917-920.

[5] E. C. Erdem, M. A. Tekalp, B. Sankur, Metrics for performance evaluation of video object segmentation and tracking without ground-truth. *Image Processing Proceedings*, 2001, pp. 69-72.

[6] B. Potočnik, D. Zazula, Automated analysis of a sequence of ovarian ultrasound images, Part I: segmentation of single 2D images, *Image and Vision Computing*, No. 20, Vol. 3, 2002, pp. 217-225.

[7] P. D. Huttenlocher, A. G. Klanderman, J. W. Rucklidge, Comparing images using the Hausdorff distance, *IEEE Transaction on Pattern Analysis and Machine Intelligence*, No. 15 Vol. 9, 1993, pp. 850-863.

[8] M. W. Wells, L. E. W. Grimson, R. Kikinis, A. F. Jolesz, Statistical intensity correction and segmentation of MRI data, *In Proceedings of the SPIE: Visualization in Biomedical Computing*, 1994.

[9] SimBio Consortium, SimBio - A Generic Environment for Bio-numerical Simulation, http : //www.simbio.de/, 2003.

[10] T. Kapur, L. E. W. Grimson, M. W. Wells, R. Kikinis, Segmentation of Brain Tissue from Magnetic Resonance Images, *Medical Image Analysis*, Vol. 1, 1996, pp. 109-127.

[11] V. C. Chalana, Y. Kim, A methodology for evaluation of boundary detection algorithms on medical images, *IEEE Transactions on Medical Imaging*, No. 16, Vol. 5, 1997, pp. 642-652.

[12] K. W. Pratt, Digital Image Processing, John Wiley & Sons, 2001.

[13] L. B. Vicente, A. M. Jose, E. Eoldan, R. Lopera, J. Francisco, Measure of quality for evaluating methods of segmentation and edge detection, *Pattern Recognition*, No. 34, Vol. 6, 2001, pp. 1127-1146.

[14] W. K. Bowyer, P. Phillips, Empirical evaluation techniques in computer vision. *IEEE Computer Society*, 1998.

[15] J. W. Niessen, J. C. Bouma, L. K. Vincken, A. M. Viergever, Error metrics for quantitative evaluation of medical image segmentation, *Theoretical Foundations of Computer Vision, Kluwer Academic Publishers*, 1998, pp. 275-284.

[16] J. W. Zhang, A survey on Evaluation Methods for Image Segmentation, *Pattern Recognition*, No. 29, 1996, pp. 1335-1346.

[17] H. K. Zoh, Statistical Validation of Image Segmentation Quality Based on a Spatial Overlap Index. *Academic Radiology*, No. 11, 2004, pp. 178-189.

[18] S. C. Goodman, Introduction to Health care Technology Assessment. Nat. Library od Medicine/NICHSR, http://www.nlm.nih.gov/nichsr/ta101/ta101.pdf, 1998.

[19] J. Canny, A computational approach to edge detection, *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 1986, 679-698.

[20] P. D. Huttenlocher, A. G. Klanderman, J. W. Rucklidge, Comparing imges using the hausdorff distance, *IEEE Transaction on Pattern Analysis and Machine Intelligence*, No. 15, 1993, pp. 850-863.

[21] S. K. R. Kwan, C. A. Evans, B. G. Pike, MRI Simulation-based Evaluation of Image-processin and classification methods, *IEEE Transaction on Medical Imaging*, No. 18, 1999, pp. 1085-1097.

[22] L. D. Collins, Design and Construction of a Realistic Digital Brain Phantom, *IEEE Transaction on Medical Imaging*, No. 17, 1998, pp. 463-468.

Dušan Heric received his B.Sc. degree from the Faculty of Electrical Engineering and Computer Science of University of Maribor, Slovenia, in 2002. He is currently working as a young researcher at the same university. His main research interests are image and signal processing.

Božidar Potočnik received his Ph.D. degree from the University of Maribor, Slovenia, in 2000. He is currently working as an Assistant Professor at the same faculty. His main research interests include, digital image segmentation, pattern recognition and machine learning.