31st Annual International Conference of the IEEE EMBS
Minneapolis, Minnesota, USA, September 2-6, 2009

QUANTITATIVE COMPARISON OF SEGMENTATION METHODS FOR IN-BODY IMAGES

Farhan Riaz[1], Mario Dinis Ribeiro[2], Miguel Tavares Coimbra[1], Member, IEEE

[1]Instituto de Telecomunicações, Faculdade de Ciências da Universidade do Porto
[2]CINTESIS / Faculdade de Medicina do Porto and Instituto Português de Oncologia

ABSTRACT

In this paper, we present a numerical comparison of how well segmentation algorithms approximate the manual segmentation of gastroenterologists for a set of endoscopic images. Different areas in these images demand different levels of analysis by a clinician and some provide critical information about the patient. Our objective is thus to segment endoscopic images so that the results mimic as closely as possible the areas that were considered relevant by doctors. We focus on a detailed quantitative comparison of two popular segmentation algorithms, mean shift and normalized cuts, when applied to in-body images, most specifically for vital-stained magnification endoscopy. Segmentation results are compared with the manual annotations of the same images performed by two specialist clinicians. Results show that if we simply consider the most relevant segmented patch, normalized cuts performs better. However, if we allow the annotated area to be represented by multiple patches, mean shift is clearly a better choice, although automatic ways to determine its kernel's bandwidth are highly desirable.

Index Terms— image segmentation, medical imaging, gastroenterology, endoscopy.

1. INTRODUCTION

Recent advances in medical science have contributed to a steady decline in mortality due to gastrointestinal cancer. However, this is still considered one of the deadliest forms of cancer and one of the main reasons behind this fact is late diagnosis [1]. Currently it is possible to directly observe all parts of gastrointestinal tract through endoscopy but these examinations are either time consuming, invasive or in need of standardization given their low classification reliability. Given this, it is imperative that the medical and signal processing communities join efforts in creating assisted diagnostic systems that can reduce the financial and temporal effort required for various Endoscopy modalities present in Hospitals.

A recent landmark that was achieved by medical scientists is the Dinis-Ribeiro classification proposal for subdividing the patients into different groups based on the color and texture patterns of the gastric mucosa [2]. This proposal helps doctors classify patients in three different groups based on the severity of the disease and our research group has shown that, given a manual segmentation, it is possible to replicate this classification using computers with reasonable accuracy [3, 4]. In this paper, we will expand this work by answering a specific question: How well can current state-of-the-art segmentation algorithms approximate the manual annotation performed by a clinician for in-body images?

The structure of the paper is as follows: In Section 2 we discuss the two segmentation algorithms used, followed by the description of the reference dataset and the measures used for evaluation in Section 3. In Section 4, we present the obtained results followed by a discussion in Section 5.

2. SEGMENTATION

For the purposes of the study described in this paper, we have selected two of the currently most popular segmentation algorithms: mean shift and normalized cuts. Our choice is motivated by the high degree of success that these methods have achieved in the recent past for biomedical images [5]. Due to space limitations, we will only present a short description of each and suggest references for additional details.

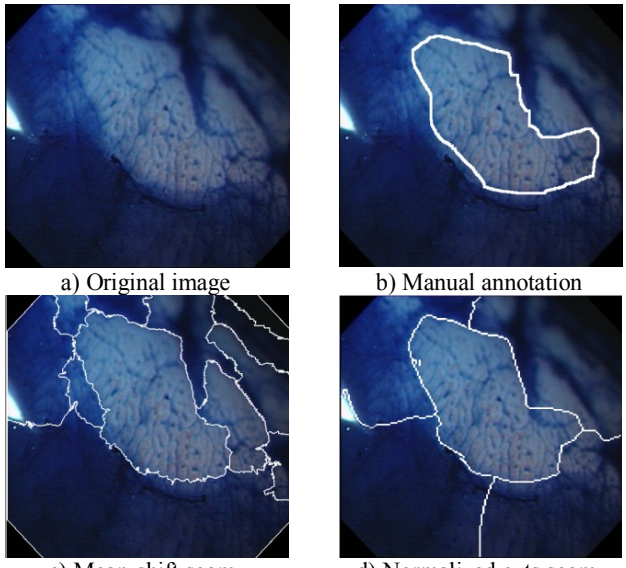
2.1. Mean Shift

The mean shift method was proposed in 1975 by Fukunaga and Hostetler [6] and it was mainly sidelined until Cheng's work [7] in 1998 in which he used it for mode seeking and clustering in a distribution. The mean shift method is motivated by the iterative calculation of the gradient of the kernel density estimation to find the densest region in a distribution.

2.2. Normalized cuts

The *normalized cuts* method is a graph theoretic approach for solving the perceptual grouping problem in vision. In normalized cuts, all the sets of points lying in the feature space are represented as a weighted, undirected graph. The weight of each arc is assigned using a set of pre-defined criteria. These can be based on the spatial distance among the pixels, their brightness values, etc. Usually the easiest

way to perform segmentation in graph theoretic algorithms is to disconnect the edges having small weights usually known as the *minimum cut* [8]. The problem with minimum cuts is that it typically results in over segmentation since the method basically finds local minima. Shi and Malik [9] proposed in 2000 a new approach that aims at extracting the global impression of an image instead of focusing on its local features. In this approach known as *normalized cuts*, the cut between two graphs is normalized by the volumes of the resulting graphs.



a) Original image b) Manual annotation

c) Mean-shift segm. d) Normalized cuts segm.

Fig. 1. Examples of the various segmentation methods used.

3. MATERIALS AND METHODS

All images used in this study were obtained using an Olympus CV-180 endoscope at the Portuguese Institute of Oncology (IPO) Hospital in Porto, Portugal during routine clinical work. Optical characteristics of this endoscope include 140º field of view and four way angulation (180º vertical and 160º horizontal), and allows depths of field between 2 and 100 mm. The endoscopic videos used were recorded on tapes using a Digital Video (DV) recorder while performing real endoscopic examinations. A total of 4 hours of video was analyzed and 144 images were selected given their clinical relevance. This was first determined by pre-selecting images that were annotated during the procedure by the clinician performing the exam, and later each image was individually selected for this study by an expert clinician. Images were saved as graphics files of type PNG (Portable Network Graphics) with a resolution of 518x481.

3.1. Comparison Methodology

Given our proposed objective, we need a set of measures that can quantify how well our segmentation approximates the manual annotation performed by clinicians. Some factors were considered important: number of resulting segments, area similarity and shape similarity. For notation purposes, let us consider the following definitions:

- An *annotated region* is an image section resulting from manual annotation, with an area equal to A_a.
- A *patch* is an image section resulting from automatic segmentation, with an area equal to S_a.
- The *area* of an image section is equal to its total number of pixels.
- *Patches* with *area* lower than 1% of the image size were ignored for this comparison.

3.1.1. Number of patches

Total number of image patches produced by the segmentation algorithm. We want this to be as low as possible for computational cost purposes, assuming that each will need to be individually analyzed by statistical pattern recognition classifiers.

3.1.2. Patch Index

We are interested in measuring how much a patch overlaps with the annotated region defined by the doctor. Since we need a relative measure for patch comparison, we can compare this with the total patch area or with the total annotated region area. We will call *Patch Index (PI)* to the former. If we call O_a to the number of common pixels between both regions, our PI is:

$$PI = \frac{O_a}{S_a} \tag{1}$$

A high PI means that the patch is almost entirely inside the annotated region, hopefully mimicking its visual characteristics faithfully, meaning it can be successfully processed by statistical classifiers.

3.1.3. Annotated Area Covered

A complementary measure to PI is the percentage of the annotated region covered by the patch:

$$AAC = \frac{O_a}{A_a} \tag{2}$$

A high *Annotated Area Covered* (*AAC*) means that the patch covers most of the annotated region and it is not just representing the visual characteristics of a very small area.

3.1.4. Dice Similarity Coefficient

The Quantitative accuracy two different overlapping contours can also be computed using *Dice Similarity Coefficient (DSC)* [12].

$$DSC = \frac{2O_a}{2\,O_a + (S_a - O_a) + (A_a - O_a)} \tag{9}$$

Its values range between *0* and *1* for zero overlap and identical contours respectively.

3.1.5. Euclidean distance using point correspondence

Besides area similarity between patches and annotated regions, we are also interested in how similar the shape of both regions is. A relative comparison can be done by *Euclidean distance point correspondence* (*EPC*). In this method we calculate the point correspondence between the annotated and segmented contours and use the average Euclidean distance between them as a way to express the quality of the segment. The most comprehensive work on shape correspondence is the work by Gold [10] and Chui and Rangarajan [11]. Given our space limitations, we refer to these papers for additional details.

4. RESULTS

4.1. Single Patch Analysis

In single patch analysis, we consider the patch with the highest *patch index (PI)* and compare it with the annotated region using all the measures which have been described above. The underlying question of this choice is thus: Is there a single patch that provides us with a good approximation to a clinician's manual annotation?

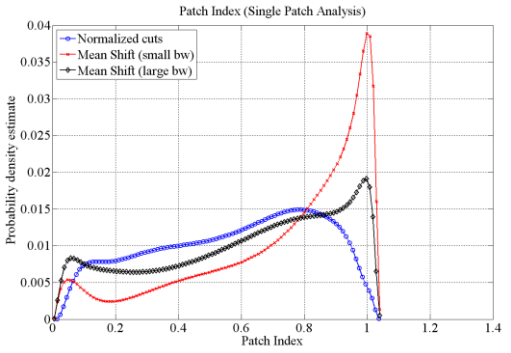
Table I compares the average values of all measures for normalized cuts and mean shift. For the latter, we consider kernels having a range of bandwidths. We can observe that normalized cuts yields a lesser number of patches and amongst them, the best ones have inferior average 'quality' (*PI*) as compared to mean shift. They cover, however, a much higher percentage of the annotated region. We obtain a smaller number of larger patches when using higher kernel bandwidths for mean shift, increasing *AAC* at the cost of *PI*. These effects can be observed in the density plots of *PI* in Fig. 2a. The strong peak for smaller kernel bandwidths is greatly suppressed by increasing kernel bandwidths. For mean shift, the EPC density is much higher in high value ranges compared with normalized cuts (Fig. 2b). Based on these results, the higher stability of normalized cuts, along with its independence from external parameters such as kernel bandwidth, makes it a good choice if our objective is a simple solution for single patch analysis.

Measures	MS (low BW)	MS (med BW)	MS (high BW)	NC
Nr. Seg.	8.17	5.46	4.07	2.66
PI	0.744	0.684	0.6215	0.574
AAC	0.570	0.634	0.715	0.786
EPC	123	127	128	109
DSC	0.4656	0.4786	0.4930	0.591

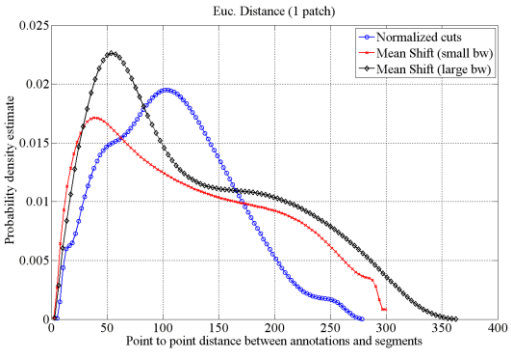
Table I. Average values of comparison measures using single patch analysis (MS – mean shift; NC – normalized cuts; BW – kernel bandwidth).

4.2. Multiple Patch Analysis

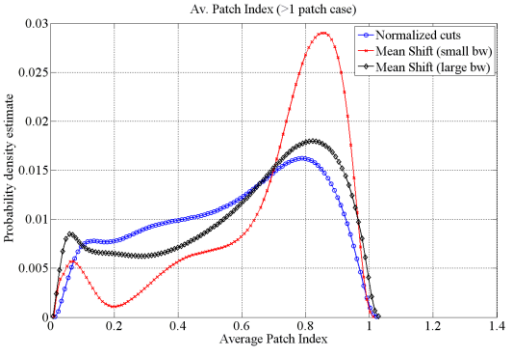
Let us now assume that we are dealing with a pattern recognition system that is able to fuse different image patches into a single coherent region, based on the individual classification of each patch. Our segmentation question can now be seen as: Is there a small set of image patches that, when merged, provide us with a good approximation of a clinician's manual annotation?



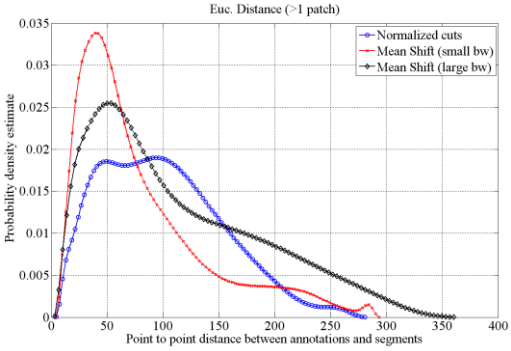
a) PI (Single Patch Analysis)



b) EPC (Single Patch Analysis)



c) PI (Multiple Patch Analysis)



d) EPC (Multiple Patch Analysis)

Fig. 2. Comparison of segmentation methods using single patch and multiple patch analysis

For this purpose, let us consider the following:

- We will only consider patches with PI larger than an arbitrary threshold, assuming that our pattern recognition system will still be able to correctly classify patches even when faced with slightly corrupted visual characteristics. For the purposes of this study, we considered 70% to be a reasonable expectation from a robust statistical classifier.
- All considered patches are merged into a single one, which is then compared with the annotated region using the measures described in Section 3.1.

Table II shows the average values of all comparison measures for both segmentation algorithms. If we assume that the number of merged patches is not significant for our choice of method, mean shift with small kernel bandwidths seems to perform better when compared to normalized cuts, mainly due to significant differences between *PI* and *EPC*. An interesting observation in Fig. 2c is the shift of density from higher to lower values of *PI* for mean shift. With increasing kernel bandwidth, average *PI* for mean shift approaches that of normalized cuts (Table II). We can also observe in Fig. 2d that an increase in kernel bandwidth induces a strong impact on the *EPC* peak which lies in low value ranges when using mean shift.

Given all the presented results, we can conclude that if our objective is to obtain the best approximation possible to the manual annotation of a clinician, we should use mean shift with low kernel bandwidths and use multiple patch analysis. If this additional computational pressure cannot be handled by the posterior pattern recognition stage, then a simple but yet effective solution is to use normalized cuts.

Measures	MS (low BW)	MS (med BW)	MS (high BW)	NC
Merg. Segs.	2.15	1.53	1.29	1.14
PI	0.693	0.651	0.601	0.567
AAC	0.808	0.768	0.795	0.834
EPC	79.8	99.1	111.5	99.6
DSC	0.6823	0.60	0.565	0.6245

Table II. Average values of comparison measures using multiple patch analysis (MS – mean shift; NC – normalized cuts; BW – kernel bandwidth).

5. DISCUSSION

In this paper, our objective was to compare the annotations made by specialist physicians with the image patches, obtained using different segmentations algorithms for a specific in-body imaging scenario. The aim was to investigate if the currently existing algorithms can be used to approximate the annotations or not?

As conclusions to this study, the algorithms investigated in this paper give us a reasonably good approximation of the manual annotations, however, some specific scenarios can affect our choice amongst them. If our objective is to simply obtain the most relevant image patch which mimics the manual annotations, it is more appropriate to use normalized cuts. However, if we have an advanced classification layer, which has the capability to fuse together a number of patches based on their semantic relevance, mean shift is an automatic choice.

Future work includes a deeper study on kernel parameters of mean shift. Although we have tested a set of different bandwidth parameters, which we considered reasonable for nearly all endoscopic images from the available dataset, the results for mean shift when doing single patch analysis could theoretically improve given an optimal automatic choice of kernel's bandwidth for each image.

REFERENCES

[1] Hermann Brenner, "Long term survival rates of cancer patients achieved by the end of the 20th century: a period analysis", *The Lancet, 360,* pp. 1131-1135, Oct 2002.

[2] M.D. Ribeiro, "Clinical, Endoscopic and Laboratorial Assessment of Patients with Associated Lesions to Gastric Adenocarcinoma", *Faculdade de Medicina da Universidade do Porto,* Phd Thesis, 2005.

[3] A.M. Sousa, "Analysis of Colour and Texture Features of Vital-Stained Magnification-Endoscopy images for Computer-assisted Diagnosis of Precancerous and Cancer Lesions", *Faculdade de Engenharia da Universidade do Porto, Master Thesis, 2008.*

[4] A.M. Sousa, M.D. Ribeiro, M. Areia, M. Correia, M. Coimbra, "Towards more adequate colour histograms for in-body images", *International Conference of the IEEE Engineering in Medicine and Biology Society,* Vancouver 2008.

[5] R. Rodriguez, P. J. Castillo, V. Guerra, A. G. Suarez, E. Izquierdo, "Two robust techniques for Segmentation of Biomedical Images", *Computacion Sistemas*, vol. 9, no. 4, pp. 355-369, Mexico 2006.

[6] K. Fukunaga and L.D. Hostetler, "The Estimation of the Gradient of a Density Function, with Applications in Pattern Recognition", *IEEE trans. Information Theory,* vol. 21, pp. 32-40, 1975.

[7] Y. Cheng, "Mean Shift, Mode Seeking, and Clustering", *IEEE trans. Pattern Analysis and Machine Intelligence,* vol. 17, no. 8, pp. 790-799, Aug. 1995

[8] Z. Wu and R. Leahy, "An Optimal Graph Theoretic Approach to Data Clustering: Theory and Its Applications to Image Segmentation", *IEEE trans. Pattern Analysis and Machine Intelligence,* vol. 15, no. 11, pp. 1101-1113, Nov. 2003

[9] J. Shi and J. Malik, "Normalized cuts and Image Segmentation", *IEEE trans. Pattern Analysis and Machine Intelligence,* vol. 22, no. 8, pp. 888-905, Aug. 2000.

[10] S. Gold, A. Rangarajan, C.P. Lu, S. Pappu and E. Mjolsness, "New Algorithms for 2D and 3D Point Matching: Pose Estimation and Correspondence", *Pattern Recognition,* vol. 31, no. 8, pp. 888-905, Aug. 2000.

[11] H. Chui and A. Rangarajan, "A New Algorithm for Non-Rigid Point Matching", *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 44-51, June 2000.

[12] Van Rijsbergen, C.J.:: Information Retrieval. Butterworth-Heinemann, Newton, MA, USA, 1979.