Pattern Recognition xx (xxxx) xxxx–xxxx

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Semi-supervised learning and graph cuts for consensus based medical image segmentation

Dwarikanath Mahapatra

IBM Research - Australia, Melbourne, Australia

ARTICLE INFO

Keywords:
Multiple experts
Segmentation
Crohn's disease
Retina
Self-consistency
Semi supervised learning
Graph cuts

ABSTRACT

Medical image segmentation requires consensus ground truth segmentations to be derived from multiple expert annotations. A novel approach is proposed that obtains consensus segmentations from experts using graph cuts (GC) and semi supervised learning (SSL). Popular approaches use iterative Expectation Maximization (EM) to estimate the final annotation and quantify annotator's performance. Such techniques pose the risk of getting trapped in local minima. We propose a self consistency (SC) score to quantify annotator consistency using low level image features. SSL is used to predict missing annotations by considering global features and local image consistency. The SC score also serves as the penalty cost in a second order Markov random field (MRF) cost function optimized using graph cuts to derive the final consensus label. Graph cut obtains a global maximum without an iterative procedure. Experimental results on synthetic images, real data of Crohn's disease patients and retinal images show our final segmentation to be accurate and more consistent than competing methods.

1. Introduction

Combining manual annotations from multiple experts is important in medical image segmentation and computer aided diagnosis (CAD) tasks such as performance evaluation of different registration or segmentation algorithms, or to assess the annotation quality of different raters through inter- and intra-expert variability [1]. Accuracy of the final (or consensus) segmentation determines to a large extent the accuracy of (semi-) automated segmentation and disease detection algorithms.

It is common for medical datasets to have annotations from different experts. Combining many experts' annotations is challenging due to their varying expertize levels, intra- and inter-expert variability, and missing labels of one or more experts. Poor consensus segmentations seriously affect the performance of segmentation algorithms, and robust fusion methods are crucial to their success. In this work we propose to combine multiple expert annotations using semi-supervised learning (SSL) and graph cuts (GC). Its effectiveness is demonstrated on example annotations of Crohn's Disease (CD) patients on abdominal magnetic resonance (MR) images, retinal fundus images, and synthetic images. Fig. 1 shows an example with two consecutive slices of a patient affected with CD. In both slices, the red contour indicates a diseased region annotated by *Expert 1* while green contour denotes diseased regions annotated by *Expert 2*. Two significant observations can be made: (1) in Fig. 1(a) there is no common region which is marked as diseased by both experts; (2) in Fig. 1(b) the area agreed by both experts as diseased is very small. Fig. 1 (c) illustrates the challenges in retinal fundus images where different experts have different contours for the optical cup. The challenges of intra- and inter-expert variability are addressed by a novel self-consistency (SC) score and the missing label information is predicted using SSL..

1.1. Related work

Fusing expert annotations involves quantifying annotator performance. Global scores of segmentation quality for label fusion were proposed in [2,3]. However, as suggested by Restif in [4] the computation of local performance is a better measure since it suits applications requiring varying accuracy in different image areas. Majority voting has also been used for fusing atlases of the brain in [5]. However, it is limited by the use of a global metric for template selection which considers each voxel independently from others, and assumes equal contribution by each template to the final segmentation. It also produces locally inconsistent segmentations in regions of high anatomical variability and poor registration. To address these limitations weighted majority voting was proposed in [6] that calculates weights based on intensity differences. This strategy depends on intensity normalization and image registration and is error prone.

A widely used algorithm for label fusion is STAPLE [3] that uses Expectation-Maximization (EM) to find sensitivity and specificity values maximizing the data likelihood. These values quantify the quality of expert segmentations. Their performance varies depending

E-mail address: dwarim@au1.ibm.com.

http://dx.doi.org/10.1016/j.patcog.2016.09.030
Received 1 February 2016; Received in revised form 2 September 2016; Accepted 21 September 2016
Available online xxxx
0031-3203/

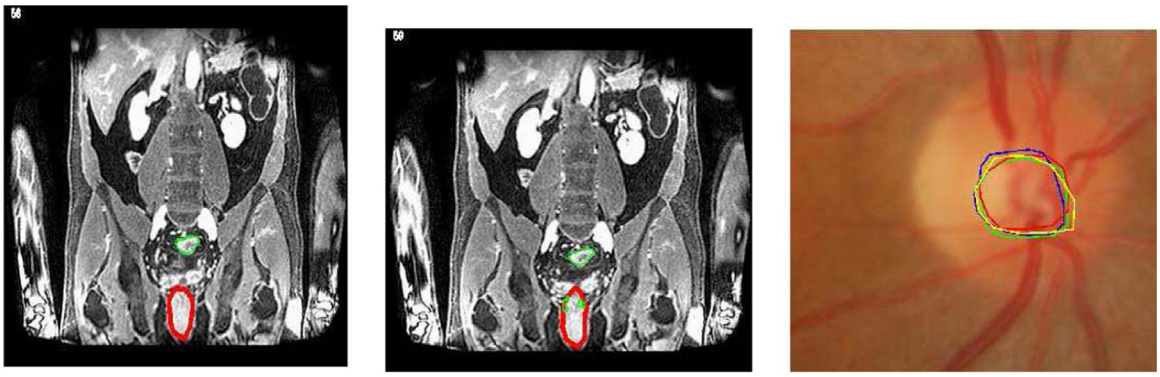

Fig. 1. (a)–(b) Illustration of subjectivity in annotating medical images. In both figures, red contour indicates diseased region as annotated by *Expert 1* while green contour denotes diseased region as annotated by *Expert 2*. (c) outline of optic cup by different experts.

upon annotation accuracy, or anatomical variability between templates [7]. Commowick et al. propose Local MAP STAPLE (LMSTAPLE) [8] that addresses the limitations of STAPLE by using sliding windows and Maximum A Posteriori (MAP) estimation, and defining a prior over expert performance. Wang et al. [9] exploit the correlation between different experts through a joint probabilistic model for improved automatic brain segmentation. Chatelain et al. in [10] use Random forests (RF) to determine most coherent expert decisions with respect to the image by defining a consistency measure based on information gain. They select the most relevant features to train the classifier, and do not combine multiple expert labels. Statistical approaches such as COLLATE [11] model the rating behavior of experts and use statistical analysis to quantify their reliability. The final annotation is obtained using EM. The SIMPLE method combines atlas fusion and weight selection in an iterative procedure [12]. Combining multiple atlases demonstrates the importance of anatomical information from multiple sources in segmentation tasks leading to reduced error compared to a single training atlas [13,14].

1.2. Our contribution

The disadvantage of EM based methods is greater computation time, and the risk of being trapped in local minimum. Consequently, the quantification of expert performance might be prone to errors. Statistical methods such as [15] require many simulated user studies to learn rater behavior, which may be biased towards the simulated data.

Another common issue is missing annotation information from one or more experts. It is common practice to annotate only the interesting regions in medical images such as diseased regions or boundaries of an organ and disagreement between experts is a common occurrence. However in some cases we find that one or more experts do not provide any labels in some image slices, perhaps due to mistakes or inattention induced due to stress. In such cases it is important to infer the missing annotations and gather as much information as possible since it is bound to impact the quality of the consensus annotation. Methods like STAPLE predict missing labels that would maximize the assumed data likelihood function, which seems to be a strong assumption on the data distribution.

Our work addresses the above limitations through the following contributions:

1. SSL is used to predict missing annotation information. While SSL is a widely used concept in machine learning it has not been previously used to predict missing annotations. Such an approach reduces the computation time since it predicts the labels in one step without any iterations as in EM based methods. By considering local pixel characteristics and global image information from the available labeled samples, SSL predicts missing annotations using global information but without making any strong assumptions of the form of the data generating function.
2. A SC score based on image features that best separate different training data quantifies the reliability and accuracy of each annotation. This includes both local and global information in quantifying segmentation quality.
3. Graph cuts (GC) are used to obtain the final segmentation which gives a global optimum of the second order MRF cost function and also incorporates spatial constraints into the final solution. The SC is used to calculate the penalty costs for each possible class as reference model distributions cannot be defined in the absence of true label information. GC also pose minimal risk of being trapped in local minima compared to previous EM based methods.

We describe different aspects of our method in Sections 2–5, present our results in Section 7 and conclude with Section 8.

2. Image features

Feature vectors derived for each voxel are used to predict any missing annotations from one or more experts. Image intensities are normalized to lie between $[0, 1]$. Each voxel is described using intensity statistics, texture and curvature entropy, and spatial context features, and they are extracted from a 31×31 patch around each voxel. In previous work [16] we have used this same set of features to design a fully automated system for detecting and segmenting CD tissues from abdominal MRI. These patches were used on images of different sizes, 400×400 and 2896×1944 pixels. Through extensive experimental analysis of the RF based training procedure we identified context features to be most important followed by curvature, texture and intensity. Our hand crafted features also outperformed other feature combinations [17]. Since the current work focuses on a method to combine multiple expert annotations, we refer the reader to [16] for details.

2.1. Intensity statistics

MR images commonly contain regions that do not form distinct spatial patterns but differ in their higher order statistics [18]. Therefore, in addition to the features processed by the human visual system (HVS), i.e., mean and variance, we extract skewness and kurtosis values from each voxel's neighborhood.

2.2. Texture entropy

Texture maps are obtained from 2-D Gabor filter banks for each