Computer Vision and Image Understanding 151 (2016) 114–123

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

Combining multiple expert annotations using semi-supervised learning and graph cuts for medical image segmentation

Dwarikanath Mahapatra

Department of Computer Science, ETH Zurich, 8092 Zurich, Switzerland

ARTICLE INFO

Article history:
Received 1 March 2015
Accepted 13 January 2016

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

ABSTRACT

Generating consensus ground truth segmentation from multiple experts is important in medical imaging applications such as segmentation. We propose a novel approach to combine multiple expert annotations using graph cuts (GC) and semi supervised learning (SSL). Current methods use iterative Expectation–Maximization (EM) based approaches to estimate the final annotation and quantify annotator's performance. This poses the risk of getting trapped in local minimum and providing inaccurate estimates of annotator performance. A novel self consistency (SC) score quantifies annotator performance based on the consistency of their annotations in terms of low level image features. The missing annotations are predicted using SSL techniques that consider global features and local image consistency. The self consistency score also serves as the penalty cost in a second order Markov random field (MRF) cost function which is optimized using graph cuts to obtain the final consensus label. Graph cut optimization gives a global maximum and is non-iterative, thus speeding up the process. 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 those obtained by competing methods. It also highlights the effectiveness of self consistency in quantifying expert reliability and accuracy of SSL in predicting missing labels.

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 expertise 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 made on multiple expert annotations of Crohn's Disease (CD) patients on abdominal magnetic resonance (MR) images, and also on retinal fundus images. Fig. 1 shows an example with two consecutive slices of a patient affected with CD. In both slices red contour indicates 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

E-mail address: dwarikanath.mahapatra@inf.ethz.ch

http://dx.doi.org/10.1016/j.cviu.2016.01.006
1077-3142/

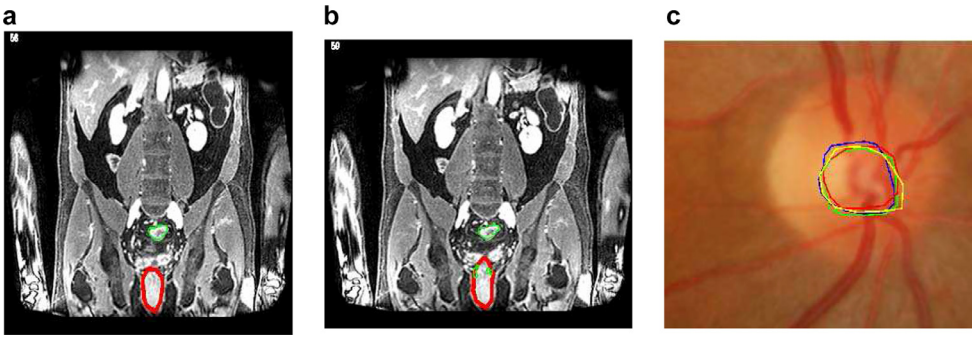

Fig. 1. (a) and (b) Illustration of subjectivity in annotating medical images. In both figures, red contour indicates diseased region as annotated by *Expert 1* while green contour denotes diseased region as annotated by *Expert 2*. (c) Outline of optic cup by different experts. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

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 strategies, which is prone to errors.

A widely used algorithm for label fusion is STAPLE [3] that uses Expectation–Maximization (EM) to find sensitivity and specificity values that maximize the data likelihood. These values quantify the quality of expert segmentations. However the performance varies depending 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 demonstrate the importance of having different expert labels in segmentation since multiple atlases reduce error over using a single training atlas [13,14].

Crohn's disease is a condition of the gastrointestinal (GI) tract that leads to bleeding, weight loss, diarrhea, ulcerations and in extreme cases blockage of the GI tract. Early detection of CD can help in rapid diagnosis, reduce the time and cost for therapy planning, and improve the quality of life of affected patients. In previous work we have proposed (semi-)automated machine learning (ML) algorithms to detect and segment CD tissues from abdominal magnetic resonance (MR) images [15–18]. The performance of these methods depends, to a large extent, on the quality of annotations provided by experts. Hence the motivation to develop a robust method to obtain consensus segmentations.

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 [19] 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. Disagreement between different experts is a common occurrence. However in some cases we find that one or more experts do not provided 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. Our work addresses the above limitations through the following novelties:

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 formation but without making any strong assumptions of the form of the data generating function.
2. A SC score quantifies the reliability and accuracy of each annotation by calculating visual features in a pixel neighborhood and comparing it with other pixels over a larger area. 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 pose minimal risk of being trapped in local minimum as in 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 [15,20] we have used these same set of features to design a fully automated system for detecting and