Assessing Ground Truth of Glandular Tissue

Christina Olsén and Fredrik Georgsson

Department of Computing Science,
Umeå University, SE-901 87 Umeå Sweden
{colsen, fredrikg}@cs.umu.se

Abstract. In medical image analysis a ground truth to compare results against is of vital importance. This ground truth is often obtained from human experts. The aim of this paper is to discuss the problem related to the use of markings made by an expert panel. As a partial solution, we propose a method to relate markings to each other in order to establish levels of agreement. By using this method we can assess the performance of, for instance, segmentation algorithms.

1 Background

Performance evaluation is essential for providing a scientific basis for image analysis in general and for medical image analysis in particular. In order to evaluate the performance of an image analysis system the output of the system has to be correlated to a true value. This true value is often referred to as ground truth, golden truth, golden standard, etc. In some cases it can be difficult and highly controversial for a layman to assess the true value that the image analysis system is supposed to achieve. In these cases the solution often involve human domain experts who define the true value.

Several researchers in image analysis have studied the evaluation of segmentation algorithms based on ground truth obtained from a group of experts. Many of them emphasize the importance of an objective ground truth. Zou et al. [13] presented systematic approaches to validate the accuracy of automated image segmentation. Based on the Expectation-Maximization algorithm for computing a probability estimate of the ground truth segmentation from a group of expert segmentations presented in [12], the authors modeled the probabilistic segmentation results using a mixture of two beta distributions with different shape parameters for the interpretation of the tumor class. Furthermore, Warfield et al. [12] present a simultaneous measure of the quality of each expert, which enables the assessment of an automated image segmentation algorithm, and direct comparison of expert and algorithm performance. Smyth [11] and Bromiley et al. [3] have also addressed the problems related to algorithm evaluation based on uncertain ground truth. Olsén and Georgssson [9] and Olsén [8] addressed these problems in relation to segmentation methods concerning mammography.

The aim of this paper is to discuss the problem of assessing ground truth and to provide a novel method of estimating ground truth in the case of binary markings in $\mathbb{Z}^n$.

Susan M. Astley et al. (Eds.): IWDM 2006, LNCS 4046, pp. 10–17, 2006.

2 Theoretical Background

We assume that we have K different domain experts who all marked some properties p regarding anatomical landmarks depicted in L images. A measure of agreement can be defined as

$$\Lambda_i^p = \frac{\mu(A_1^p \cap A_2^p \cap \ldots \cap A_K^p)}{\mu(A_1^p \cup A_2^p \cup \ldots \cup A_K^p)}. \tag{1}$$

where $i = 1, 2, \ldots, L$ and $\mu(\cdot)$ is a measure of the set (i.e. the numbers of points if A is discrete).

In the general case, a ground truth can be any subset $A_i \subset \mathbb{Z}^n$, where n is the spatial dimensionality of the media, i.e. $n = 1$ for a signal, $n = 2$ for an image, $n = 3$ for a volume etc. In passing, it is noted that the dimensionality of the set A_i may be lower than n. Examples of this are; marking a line in an image, a point in a volume etc.

We define the distance from a point $\mathbf{p}$ to a set $S \subset \mathbb{Z}^n$ to be $D(\mathbf{p}, S) = \inf\{d(\mathbf{p}, \mathbf{q}) \mid \mathbf{q} \in S\}$, where $d(\cdot, \cdot)$ is some metric defined on $\mathbb{Z}^n$. We note that $D(\mathbf{p}, S) = 0$ if $\mathbf{p} \in S$. The distance can be estimated efficiently by using a distance transformation [1].

A distance between two discrete sets S and U can then be defined as

$$\mathcal{D}(S, U) = \sum_{\mathbf{p} \in S} D(\mathbf{p}, U) + \sum_{\mathbf{q} \in U} D(\mathbf{q}, S). \tag{2}$$

The distance between a set S and an ensemble of sets $\mathbf{A} = \{A_1, \ldots, A_K\}$ is given by

$$\mathbf{D}_\mathbf{A}(S) = \sum_{A_i \in \mathbf{A}} \mathcal{D}(S, A_i). \tag{3}$$

It is easily seen that if we have different overlapping sets S_i, then the set with the smallest measure contained in the intersection of the sets (i.e. the set marking the smallest *area* if the underlying dimensionality of S_i is 2) is likely to minimize $\mathbf{D}_\mathbf{A}(S_i)$. This property makes a distance measure such as the one defined in Eq. 3 unsuitable for comparing a marking to that of a set of experts. It can be said that the distance defined in Eq. 3 penalizes sets that fill the plane and thus we need to add a measure of how well a set S "fills" the ensemble $\mathbf{A}$. In order to construct such a measure we define an occurrence operator

$$\phi_S(\mathbf{p}) = \begin{cases} 1 & \text{if } \mathbf{p} \in S \\ 0 & \text{otherwise} \end{cases}$$

for every set S of points. By using $\phi(\cdot)$ the ensemble $\mathbf{A}$ now gives rise to a measure of the subsets S of $\mathbb{Z}^n$, which is defined by

$$\mathbf{M}_\mathbf{A}(S) = \sum_{A_i \in \mathbf{A}} \sum_{\mathbf{p} \in S} \phi_{A_i}(\mathbf{p}). \tag{4}$$

Given an ensemble of ground truths $\mathbf{A}$, any set S that maximizes $\mathbf{M}_\mathbf{A}(S)$ whilst simultaneously minimizing $\mathbf{D}_\mathbf{A}(S)$ is said to be in good agreement with the ensemble.