NON-RIGID REGISTRATION ASSESSMENT ERRORS

17th November 2005

Abstract

In a recent document that was entitled **Error Propagation**, detailed explanation was provided, which culminated in the generation of sensitivity plots. These sensitivity plots gave insight into the strengths and validity of non-rigid registration (NRR) assessment methods. We dealt with two such assessment methods, which were rather independent and begged for some means of comparison. Ultimately, a criterion was devised, which could reliably distinguish between the two and show how one method surpassed the other.

A step was missed in that explanation, however, which ought to have fully covered the derivation of some numbers. Primarily, there needed to have been a discussion on how error bars get assigned to values of our assessment methods. Moreover, there ought to be a clearer explanation as to what they practically represent and where they all come from.

This document explains, to some degree, how such values are computed both in the case of model-based assessment, where an appearance model gets evaluated, as well as the case of overlap-based assessment. Since both methods are quite different in nature, calculation of the errors should be treated separately, each in turn.

The closing section briefly describes the steps which follow, but for a fuller, more contextual review, the previous related document ought to be looked up in conjunction.

Contents

1 INTRODUCTION

We embark on the task of assessing non-rigid registration (NRR) and we have a group of methods for doing so. For each dataset that we are given, we are able to assign a measure, which is representative of registration quality. That measure, however, is subjected to some level of uncertainty, thus it must be treated as a value with a certain variance. This document outlines the way in which we derive information about the standard errors. This should include inter-instantiation error, which is present when repeated experiments are performed and averaged over. We also deal with some error that is associated with the calculation of registration quality, which is either model-based or overlap-based in our case.

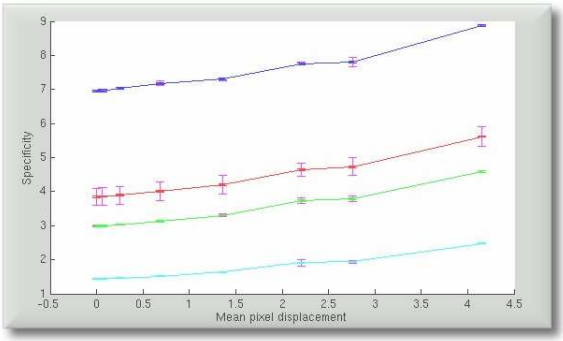
2 MODEL-BASED ASSESSMENT

Model-based NRR assessment is closely correlated (if not analogous) to the evaluation of appearance models. In order to carry out such evaluation, many synthetic images are drawn from the model. A finite number of such synthetic images get generated and then compared against the training set of the same model. In that respect, data which is representative of the model is 'locked onto' that model's seminal data, i.e. the training dataset. Since the number of synthetic images is finite, there remains some susceptibility to an error. The estimate of the model-data match suffers if the number of instantiations,

upon which the evaluation is based, remains overly low.

Distances are calculated between pairs of images, namely images in the training set and synthetic images derived from the model. These distances are computed in accordance with the shuffle distance principles, whereby each pixel is compared against a corresponding neighbourhood of pixels in another image. Since there are never enough such distances to attain a stable measure, errors should be associated with the size of the set under consideration. Having got a collection of distances for a large number of possible image pairings, the standard deviation must be computed. Subsequently, the standard error can be computed as well.

Below lies a figure where the registration quality (horizontal axis) affects the measured quality as perceived by the model-based assessment method.



Inter-set error is somewhat of a simpler case. Since we repeat our experiments 10 times and obtain different values each time, there is a certain error associated with the measure. That error is calculated simply by taking the standard error over corresponding values across the instantiations. The errors are bound to be rather large if the measured quantity varies greatly among datasets or if the number of instantiations is fairly low.

3 OVERLAP-BASED ASSESSMENT

Overlap-based assessment relies on labels (anatomical mark-up in our case), embedded in a group of images. The principal idea is that of computing overlap between corresponding labels in image sets which have been registered. Such labels, which are transformed along with the images that embody them, often reflect rather well on the correspondence among the images themselves. Thus, labels can infer registration quality in the images are less prone to error in the case of coarse anomalies.

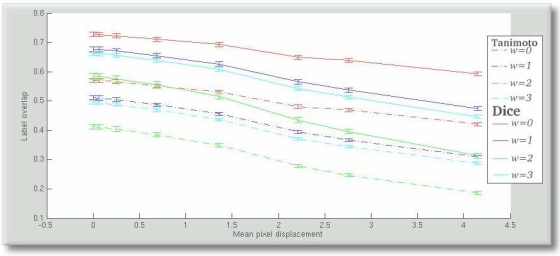
For each displacement value, we take the error associated with overlap, which has been accumulated over all labels. To get a value that is representative of all 10 instantiations, we take the average over the 10 instantiations. This accounts for the error (uncertainty) in the measure, which in this case is label overlap.

Another error we must consider is due to varying values of the measured overlap. Each instantiation gives us a slightly different value. We derive the standard deviation of the overlap quantities for each displacement value and divide by $\sqrt{N-1} = 9$, i.e. divide by 3 to get the errors.

The way these two errors should 'blended' is somewhat ill-comprehended. There are arguments to suggest that the two errors are entirely independent, but contrary arguments have been raised as well.

The evaluation of the error bars is a very crucial detail as it eventually leads to the derivation of sensitivity plots. *Correct* error bars will be a pre-requisite for impartial comparison between NRR assessment methods.

To associate a standard deviation with the overall overlap we assume that each pair of registered images represents a sample from a normal distribution. Therefore N pairwise comparisons gives a standard deviation associated with the N intersections and unions, accumulated over all the labels. The standard deviation and standard error for the total overlap are then estimated using standard error propagation formulas. The figure below depicts the effect of degrading the quality or registration on the overlap-based assessment method.



4 SUBSEQUENT PROCESSING STEPS

In order to compare a variety of NRR assessment method, a benchmark criterion needs be identified. From the plots described in previous section, a measure of "sensitivity" can be derived. More on the process in the document titled **Error Propagation**.