We propose a method for accurately identifying anatomical landmarks in 3D CT volumes based on dense matching of parts-based graphical models [1]. We propose novel methods for optimizing parameters of appearance models for landmark localization in 3D endoscopic images for which there is only a limited labeled training set, and investigate the trade-off between the number of model parameters and registration accuracy. Our method contributes to existing work in medical imaging (e.g. [3]) and computer vision (e.g. [2]) in that it is applicable to whole-body registration (i.e. skeletal and soft tissue landmarks) and improves registration accuracy in 3D with only limited training data.

### 1 Parts-based model for landmark registration

Our method is based on the Pictorial Structures (PS) model with dense matching [1]. Landmarks are specified by their \([x,y,z]\) coordinates. Local landmark appearance is modeled in terms of rectangular patches centered at landmark locations. The parameters of a generative appearance model are obtained directly from the training patches. The part-specific anisotropic patch scales, varied between 4 - 32 voxels, are free parameters that are optimized separately using the approach outlined in Section 2.

### 2 Quality measures for dense matching of landmarks

An ideal part descriptor used for dense matching of medical landmarks should have the following complementary properties:

- **Accuracy** - local maxima close to the ground-truth part location.
- **Compactness** - feature maps should be locally prominent, i.e. the descriptor response map is peaked around ground-truth location.
- **Consistency** - over different patients, the distribution of feature map local maxima with respect to the ground-truth is spatially well constrained and the patterns in the feature maps are locally similar.

Although the latter criteria have no direct connection to the performance objective, compact and consistent maps should help to better generalization compared to minimizing error directly with only limited training set.

We studied 15 measures and 120 combinations of triplets of complementary quality measures over increasing number of free parameters of the descriptor (i.e. shared and part-specific isotropic and anisotropic patch sizes), in order to optimize model parameters for landmark registration. For space reasons, we refer reader to the paper for full explanation.

**Registration accuracy** is measured by calculating the RMS distance between a maxima in the descriptor response map and the ground-truth. This is performed for each landmark and for each image \(k\).

\[
E_{\text{max}} = \sum_k d_2(\tilde{x}_k, \tilde{x}_{\text{GT}}),
\]

**Compactness** is estimated as the Jensen-Shannon divergence between the feature responses and an “ideal” response, considered as spatial histograms. This is performed for each landmark and for each image \(k\).

\[
\text{Comp} \sim \sum_{k=1..N} JSD(\|\alpha f_k\|^2_{\text{norm}}, \|\alpha f_{\text{ideal}}\|^2_{\text{norm}})
\]

The spatial histograms for the feature images \(f_k\) and for the ideal response \(f_{\text{ideal}}\) are obtained by taking the local maxima on a grid centered at the ground-truth part location. \(\alpha, \beta\) are selected empirically to assign more importance to voxels with higher posterior probability.

**Spatial consistency** is estimated for each landmark by calculating the covariance matrix \(\Sigma\) of error vectors pointing from the ground truth location to the best local match over all images, and taking its determinant

\[
\text{Spc}_{\text{max}} \sim \text{det}(\Sigma)
\]

**Appearance consistency** is estimated for each landmark as the pairwise Jensen-Shannon divergence between spatial histograms of feature responses (as in Eq. 2) over pairs of images \(k_1,k_2\).

\[
\text{AppC} \sim \sum_{k_1=1..N} \sum_{k_2=1..N} JSD(\|\alpha f_{k_1}\|^2_{\text{norm}}, \|\alpha f_{k_2}\|^2_{\text{norm}}, W_{\text{Gauss}})
\]

\(W\) is a Gaussian kernel that emphasizes bins closer to the ground-truth.

**Standard object detection measures**: We also investigate measures previously used in the literature (e.g. [2]), including the Area under ROC curve (AUC), Average precision (AP) and the F-measure (F).

### 3 Experiments and Results

We present results for the localization of 22 landmarks in clinical 3D CT volumes of lung cancer patients and optimization of part-specific patch scales. Over-fitting is likely due to high variability of the data and a limited labeled set (here, 83 patients) so we employ bootstrap analysis for validation. The average mean and maximum landmark registration error is reduced by 31% and 25% for the optimized model, compared to an empirically determined baseline scale (Figure 1). Additionally, we show improved performance over standard methods as the number of free parameters increases from an isotropic patch scale shared by all parts, to specific anisotropic patch scales learnt for each part.

![Figure 1: Summary statistics over bootstraps. Columns show errors based on part scales optimized using different measures. C1-3 are the best combinations of measures (triplets). Red line is the baseline error. Lower error and variance means more accurate and reliable descriptor.](image)

Here, we focus on optimizing patch size, but any other quantities may be optimized (e.g. number of features per part) using standard optimization techniques with quality measures presented here as cost functions.

