# Improved Anatomical Landmark Localization in Medical Images Using Dense Matching of Graphical Models

Vaclav Potesil1

Timor Kadir<sup>1</sup>

Günther Platsch<sup>2</sup>

Mike Brady<sup>1</sup>

- <sup>1</sup> University of Oxford Dept. of Engineering Science Oxford, UK
- <sup>2</sup> Siemens AG (Healthcare Sector) Molecular Imaging Oxford, UK

#### Abstract

We propose a method for reliably and accurately identifying anatomical landmarks in 3D CT volumes based on dense matching of parts-based graphical models. Such a system can be used to establish reliable correspondences in medical images which can be useful on their own or as part of more complex processing e.g. atlas building. We propose and investigate novel methods for efficiently optimizing parameters of appearance models for landmark localization in 3D images. We also investigate the trade-off between the number of model parameters and registration accuracy. We present results for the localization of 22 landmarks in clinical 3D CT volumes of cancer patients and optimization of part-specific patch scales. Over-fitting is likely due to an intrinsically high variability of the data and a limited labeled training and test set, here 83 scans, so we employ a rigorous bootstrap analysis to validate the results. The average mean and maximum registration error over all landmarks is reduced by 31% and 25% for the optimized model, compared to an empirically determined baseline. Additionally, we show a significantly 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 in the model.

### 1 Introduction

Inter-subject matching and registration of whole-body oncology CT images is a challenging problem due to the intrinsically high variability of normal subjects and of pathological structures. The motivation for our work is to develop algorithms for improved inter-subject registration in medical imaging, primarily in whole-body PET/CT oncology applications. In this work, we report a method to spatially register corresponding structures of multiple patients for which there is only a very limited labeled training dataset, reflecting clinical reality and limited clinician time. Our approach is based on a parts-based graphical model and dense matching  $[\mbox{\sc d},\mbox{\sc d}]$ .

Conventional approaches to determining correspondences in medical imaging typically rely on registration methods. However, despite considerable advances in deformable registration, there still do not exist reliable methods for aligning whole-body images of differ-

ent subjects. The performance of inter-subject registration could be improved by informative priors that capture the wide variability of structures. Active shape/appearance models (ASM/AAM) have addressed this problem [5]. However, several authors have drawn attention to the limitations of such global models when applied to clinical images in which there are significant local abnormalities [5, 15]. Graphical Models, developed in computer vision (CV), offer an alternative approach to modeling flexible objects, one which does not impose explicit global priors, and which is generally considered to be potentially more robust to local abnormalities than global AAMs.

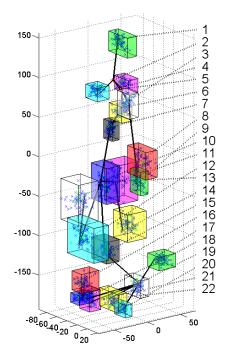
This paper proposes a method for reliably and accurately identifying landmarks in CT images (Table 1) based on dense matching of parts-based graphical models. We develop methods for efficiently optimizing the parameters of appearance models for improved landmark registration and report results with an increasing number of parameters in order to investigate the trade-off between the number of model parameters and registration accuracy. Our method improves dense matching of highly variable landmarks in clinical 3D images of cancer patients with a limited labeled training set (tens of exemplars). Over-fitting is likely to occur, and so we employ a rigorous bootstrap analysis to validate the results. In this paper, we report results for optimizing appearance model parameters in parts-based graphical models (part-specific patch scales). While we investigate a specific set of parameters in a generative model, the core method can be used within standard discriminative learning algorithms to optimize other model parameters for part localization.

### 1.1 Related Work

In medical image analysis, the attractive properties of parts-based graphical models have led to them being adopted in spine labelling [6, 11]. Whole-body registration in clinical oncology involves considerably more complex variability compared to organ-specific applications, and necessitates the localization of both skeletal and soft tissues with widely different patterns of variability. Optimal localization may demand specific settings for each part in the model. For more than a few parameters, it quickly becomes intractable to evaluate all combinations empirically.

The problem of quantifying performance of local descriptors for standard computer vision applications has been studied extensively e.g. [1]. Recently, authors investigated local appearance descriptors optimized for specific tasks [1] and appearance models trained with standard boosting approaches were shown to provide improved parts-based localization [12]. In general, these methods use detection success (e.g., threshold on the bounding box overlap) as a surrogate for localization quality and maximize a score derived from the ROC or precision-recall curve [11]. The suitability of such measures for dense localization of medical landmarks with limited training data in 3D has however not been investigated.

Our application has a number of novel aspects compared to previous work, both in terms of performance objectives and the image content. The variability of appearances in oncology images is high, due to natural variability, aging-related changes, as well as the presence of diseases and surgical interventions. The number of labeled training examples is low (in tens). To make things worse, the uncertainty of labels is high due to differences of opinions between radiologists as well as variability of repeated readings by the same expert (typically 30% plus in each case). Despite this, physicians demand low registration error within units of voxels. Importantly, many landmarks are intrinsically poorly localized and it is impossible to unambiguously declare where a landmark 'ends'. Simplifying landmark localization into a bounding box classification problem as in [2] may not be the ideal solution.



1	C2 vertebra				
2	C7 vertebra				
3	top of the sternum				
4	top right lung				
5	top left lung				
6	aortic arch				
7	carina				
8	lowest point of sternum (ribs)				
9	lowest point of sternum (tip)				
10	Th12 vertebra				
11	top right kidney				
12	bottom right kidney				
13	top left kidney				
14	bottom left kidney				
15	L5 vertebra				
16	right spina iliaca anterior superior				
17	left spina iliaca anterior superior				
18	right head of femur				
19	left head of femur				
20	symphysis				
21	os coccygeum				
22	center of bladder				

Table 1: Anatomical constraints. Colored shapes show valid constraint ranges, overlaid on one training patient (black lines). Blue and green points represent the training and test points respectively, plotted in the local coordinates of the neighboring landmarks.

## 2 Parts-Based Graphical Model for Inter-Patient Registration

Our method is based on the Pictorial Structure (PS) model, which we briefly summarize here, but refer reader to  $[\square, \boxtimes]$  for an introduction. We describe specifics of our implementation in Section 2.1. In Section 3, we outline a set of free parameters of the part appearance models and introduce method to select their values to optimize landmark registration.

### 2.1 Pictorial Structure Representation

All landmarks share the same representation, which comprises a unary energy term for local part appearance and a set of pair-wise terms for spatial compatibility. It takes the form of a tree-structured undirected graphical model.

$$P(L \mid I, \Theta) \propto \left( \prod_{i=1}^{n} p(I \mid l_i, u_i)^{\alpha} \prod_{\nu_i, \nu_j \in G} p(l_i, l_j \mid c_{ij}) \right)$$
(1)

In this equation, the first product term represents the cost of each landmark  $v_i$  at location  $l_i$ , given the quality of match of patch appearance model  $u_i$  to the image I. The second term is the spatial compatibility cost of connected pairs of landmarks  $v_i$ ,  $v_j$  placed at  $l_i$ ,  $l_j$ , given anatomical constraints  $c_{ij}$ . This equation can be rewritten as an energy minimization problem

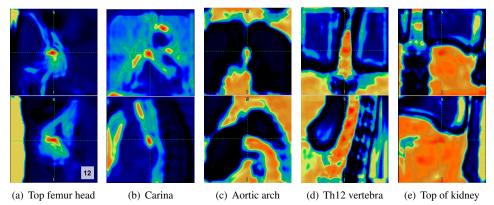


Figure 1: Example landmark posterior probability maps. Top: Coronal, Bottom: Sagittal views. Images show a 16x16x16 cm region centered at the ground-truth location. Bright colors correspond to high probability landmark placements (1). Dark colors correspond to low probability landmark placements (0). White box indicates a patch scale s = 12 vox.

by taking its negative logarithm.  $\alpha = 1$  is a normalization constant, selected empirically and fixed for all parts and experiments.

**Local Appearance.** We model local tissue appearance in terms of local rectangular image patches centered at landmark locations  $l_i$  with a scale  $S_i(x,y,z)$ . Patches are projected onto a set of n = 17 features that includes the mean template and the top 16 principal eigenpatches obtained from the ground truth patches. Additional positive examples were generated from the annotated patches by rotations within anatomically plausible ranges. Negative examples of non-landmark tissue ("background") are sampled from unlabelled tissue from spherical regions within the segmented body outline. The variability of part appearance is represented by the distribution of the coefficients,  $u_i \sim N(M, \Sigma)$  for each part i, where  $M, \Sigma$  are diagonal matrices.  $u_{bi} \sim N(M, \Sigma)$  is the model for the local "background".

**Pair-wise Anatomical Constraints.** Pair-wise spatial compatibility terms  $c_{ij}(l_i,l_j)$  penalise part placements outside of the degree of anatomical variability exhibited within the training database. The functional form of this spatial model is approximated as a distribution over  $\vec{l}_{ij,k} = l_{i,k} - l_{j,k}$  (the relative position vector of two parts i,j in training patient k).  $P(l_i \mid l_j) \sim U\left(\min \vec{l}_{ij,k}, \max \vec{l}_{ij,k}\right)$ ; (range-minimum search within a bounding cube).

### 2.2 Learning Pictorial Structure Parameters from Training Images

The parameters of the PS are learnt from the training data, following the approach of  $[\Box]$ . In more detail: graph structure is a minimum spanning tree of spatially compact ("rigid") constraints. The pair-wise spatial model is fitted by directly calculating the MLE estimate for the vector  $\vec{l}_{ij,k} = l_{i,k} - l_{j,k}$  (i.e. box-ranges).

Landmarks are specified by their [x,y,z] coordinates i.e. click-points in annotated training images. The appearance model parameters  $u_i,u_{bi}$  are obtained by direct calculation of the mean and variance of the feature coefficients from the training patches. The part-specific anisotropic patch scales  $S_i(x,y,z)$ , varied between 4 - 32 voxels, are free parameters that are optimized separately (Section 3).

### 2.3 Matching Model to Unseen Patient Images

To find the best placements for the parts in a new image, we apply the appearance models to the whole image using a sliding window approach. The log-likelihood ratios are combined using a Naive Bayes classifier to obtain the posterior probability for each landmark and sliding window placements  $l_i$ . For computational reasons, the dense descriptors were evaluated in a cube-shaped region of interest centered around the ground-truth landmark placement ([65 x 65 x 65 voxels]) for parameter optimization (Figure 1).

Min-sum belief propagation algorithm is used for an exact, globally optimal fitting using the (dense) appearance posterior probability maps and the pair-wise compatibility terms. Pair-wise terms are evaluated using fast box range-minimum search.

# 3 Improving Part Localization by Optimizing Model Parameters

In this section, we list free parameters of the appearance model introduced in Section 2.1 and then present existing and novel methods to select their values such that part registration accuracy is improved.

**Free parameters of appearance:** The model has 6 key parameters per part that influence its localization: patch scale  $S_i(x,y,z)$ , number of features per part  $n_i$  and the weight factor  $\alpha$  for each part i. In order to reduce the dimensionality of the problem given limited training examples, we focus on learning the part scale  $S_i(x,y,z)$  and keep the values n and  $\alpha$  constant for all parts (i.e. up to 66 free parameters in total).

Indeed, we could try to optimize the part scales by fitting the whole model and evaluating all combinations of scales for all 22 parts. However, this would be computationally infeasible for more than a few parameters. Therefore, we seek measures of descriptor goodness for our application that can serve as a surrogate for the ultimate goal of improving the accuracy and detection rate of the system. Generalization is an important criteria: Our training set exhibits high variability, yet we have fewer labeled training images than parameters.

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

- A. Accuracy local maxima close to the ground-truth part location.
- B. **Compactness** feature maps should be locally prominent, i.e. the descriptor response map is peaked around ground-truth location, while weak signal is present elsewhere.
- C. Consistency the spatial distribution of local maxima over different patients with respect to the part ground-truth location is spatially well constrained (i.e., low spatial variability) and the intensity patterns in the feature maps are locally similar over different patients (i.e., low quantitative variability).

Although the latter two criteria have no direct connection to the desired performance objective (i.e., registration accuracy), compact and consistent maps should lead to better generalization compared to minimizing error directly, when only limited training set is available. Measures of descriptor quality for localization are described next.

### 3.1 Measuring descriptor quality for patch scale optimization

A measure should capture the complementary qualities of a descriptor (A-C). In previous work in object recognition and localization that addresses this kind of problem, most authors pose the task as binary classification. This means that a success threshold is imposed on detections (e.g. area overlap of a bounding box) and a score that is derived from the ROC or precision-recall curve is maximized [II, II].

Here, we evaluate these existing and new measures by applying them locally to dense feature maps obtained from validation images with known ground-truth part locations. The pseudo-algorithm for patch scale optimization with a quality measure Q works as follows:

FOR each landmark i

FOR patch scale  $S_i(x, y, z)$  in a range  $(MinScale, MaxScale)_{xyz}$ 

Generate patch examples from the training images

**Learn** appearance models  $u_i, u_{bi}$ 

**Apply**  $u_i, u_{bi}$  to set of validation images  $I_k$ .

**Create** a stack of local feature response maps  $f_{i,S_{x,y,z},k}$ 

**Estimate** a quality score  $Q_{i,S_{x,y,z}}$  from the stack of all  $f_{i,S_{x,y,z},k}$ 

Find best scale  $S_{i\_best}(x, y, z) \sim \operatorname{argmin} Q_{i.S(x, y, z)}$ 

**Record** the appearance models  $u_{i best}$ ,  $u_{bi best}$  for the best scale.

**Combinations of Measures:** The quality measures  $Q_{i,S_{x,y,z}}$  are estimated independently and also as linear combinations of all possible triplets of measures. Our aim was to ascertain if there was any value in combining scores, rather than trying to optimally combine them. In more detail, for the combinations, the resulting score was obtained by the following simple fusion. For each landmark and measure (in a triplet), we binned measure scores from all scales  $Q_{i,S_{x,y,z}}$  into a vector and sorted it in an ascending order (best to worst). The ranking order was used to linearize the measure scores to [0,1]. The combined score was obtained as the geometric mean of the three linearized scores i.e. all measures were given equal weight.

### 3.2 Quality measures for dense matching

We studied 15 measures and 120 combinations of triplets of complementary quality measures over different search space parameterizations (shared and part-specific isotropic and anisotropic patch sizes). Overall, we performed 320 experiments using bootstrap analysis with a total of over 7 million localized landmarks. For space reasons, we only report those methods that performed the best.

**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.

$$E_{max} = \sum_{k} d_2(\vec{x}_k, \vec{x}_{GT}), \qquad (2)$$

Since it cannot be determined which maxima the spatial model will ultimately select among the many in the region of interest around the ground-truth, we investigated two variations: the highest maxima in the region of interest,  $E_{max}$ , and the local maxima closest to the ground-truth,  $E_{nearest}$ .

**Compactness** is estimated as the Jensen-Shannon divergence between the feature responses and an "ideal" response modeled as a smoothed "Dirac" signal centered at the ground truth ( $\delta = 4$  voxels). The responses are considered as spatial histograms. This is performed for each landmark and for each image

$$Comp \sim \sum_{k=1}^{N} JSD\left(\left(\alpha f_{k}\right)^{\beta}_{norm} \mid \left(\alpha f_{ideal}\right)^{\beta}_{norm}\right) \tag{3}$$

The spatial histograms for the feature images  $f_k$  and for the ideal response  $f_{ideal}$  are obtained by taking the local maxima on a 5x5x5 grid centered at the ground-truth part

location. $\alpha = 125$ ,  $\beta = 6$  are selected empirically so as to assign more importance to voxels with higher posterior probability. After the scaling, the histograms  $f_k$ ,  $f_{ideal}$  are normalized.

**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

$$SpC_{max} \sim \det(\Sigma)$$
 (4)

Similar to registration accuracy, we investigated two variations: the covariance of the highest maxima in the region of interest,  $SpC_{max}$ , and the covariance of the local maxima closest to the ground-truth,  $SpC_{nearest}$ .

**Appearance consistency** is estimated for each landmark as the pair-wise Jensen-Shannon divergence between spatial histograms of feature responses over pairs of images  $k_1, k_2$ 

$$AppC \sim \sum_{k_1=1..N} \sum_{k_2=1..N} JSD\left( (\alpha f_{k_1})^{\beta}_{norm} \mid (\alpha f_{k_2})^{\beta}_{norm}, W_{Gauss} \right)$$
 (5)

The spatial histograms are obtained and normalized following the same procedure as in Eq. 3. *W* is an isotropic Gaussian kernel, which confers more weight to the consistency of spatial bins close to the ground-truth location.

Standard object detection measures: We also investigate methods previously used for performance evaluation or as optimization criteria for parts-based object localization algorithms in the literature  $[\Box, \Box]$ . We investigated the **Area under ROC curve** (**AUC**), **Average interpolated precision** (**AP**) and the **F-measure** (**F**). Positive and negative examples are defined analogically to the compactness measure. This means, for each landmark, positive examples are all training image feature response map within 4 voxels RMS error from the ground-truth. Negative examples are surrounding voxels in the region of interest, from all images. Additionally, we calculated **Detection Failure**( $D_{fail}$ ) measure, defined as % detections with higher RMS error than 10 voxels from the ground truth across all images.

### 4 Data and Experiments

Our training and test set comprises 83 patients from a database of lung cancer PET/CT cases. Contrast-enhanced, diagnostic-quality CT scans of the torso for each patient were acquired using Siemens Biograph 16 and re-sampled to 2mm isotropic resolution.

22 clinical anatomical landmarks were selected by an expert nuclear medicine physician with 20 years experience, according to their utility as reference points for whole-body PET/CT. Ground-truth landmark positions were annotated by a trainee medic, who followed interactive guidance by the expert. The expert subsequently validated a subset of the annotations, including a) All placements flagged as uncertain or abnormal b) All where the medic disagreed with another trainee medic c) All in 30% of randomly selected patients.

The database was split into a training and validation set of 63 patient images and an unseen test set of 20 images. For all images and 22 parts in the model (Table 1), feature maps were generated by applying the appearance descriptors in a [65x65x65] local region of interest around the ground truth locations. This was repeated for a range of isotropic and anisotropic patch scales between [MinScale = 4, MaxScale = 32 voxels].

The quality measures were evaluated on validation image feature maps, in a local region of interest sized [33x33x33] vox centered at ground-truth part locations. 20 maps were randomly sampled from the training set with replacement and results averaged over n=50 bootstraps. For each part and measure, the patch size that minimized the quality measure

was selected, following the algorithm in Section 3. Finally, the graphical model with spatial priors and optimized part sizes was matched to the test set and localization errors recorded.

To assess the statistical reliability of the results, we computed summary statistics over all bootstraps. This included the mean RMS error from the ground-truth part location (incl. ranges and variance of this estimate) and the mean maximum RMS error. The baseline reference is a fixed isotropic patch size shared by all parts, empirically selected in previous development of the landmark detection system (s = 16 voxels).

### 5 Results

First, we present results which compare the various descriptor quality measures applied to the problem of selecting the best part-specific anisotropic patch scale. The results show the average localization errors, evaluated on the unseen test set, which arise from using patch scales which are optimized on the training set using our algorithm. Both the best single measures and combinations are shown.

Results are presented as statistics of the localization error over 50 bootstraps in Figure 2 and in Table 2. Several combinations achieved similar results, but here, for brevity, we only include the best three  $C1\{E_{max},SpC,F\}$ .  $C2\{E_{max},SpC,AP\}$ ,  $C3\{E_{max},AppC,Comp\}$ .

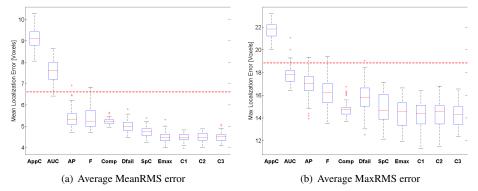


Figure 2: Summary statistics (over bootstraps). The columns show mean and max RMS errors, based on part scales optimized using different measures and the best combinations of measures. Red dashed line is the baseline reference error. Lower error means better descriptor. Lower variance means more reliable descriptor.

Next, in Table 3, we show the effect of varying the number of parameters of the optimization; from shared isotropic scale through to part-specific anisotropic scale. Results are shown for the patch scales learned using the best of standard measures (F), best of all  $(E_{max})$  and the best combination  $(C1\{E_{max}, SpC, F\})$ .

### 6 Discussion and Conclusion

We aim to establish correspondences for a set of clinical anatomical landmarks which are guaranteed to be present in whole-body oncology images. Our method can compensate for weak appearance descriptors with a high level of false positives and can disambiguate repeated self-similar structures by mobilizing relations between landmarks in the image. In order to achieve accurate and reliable localization, free parameters of part descriptors need to be set appropriately. However, an exhaustive optimization of all part descriptors

Table 2: **Summary Statistics.** The table shows statistics on error (columns) resulting from optimizing anisotropic patch sizes using different quality measures (rows). Lower mean and maximum error means better descriptor. Lower variance of error means more reliable descriptor. Statistics are over all landmarks and bootstraps. CoV is the coefficient of variation.

	Avg	Max	CoV	Avg
	(MeanRMS)	(MeanRMS)	(MeanRMS)	(MaxRMS)
Baseline	6.6	-	-	18.82
AppC	9.13	10.26	0.05	21.7
AUC	7.60	8.63	0.07	17.83
AP	5.37	6.92	0.08	16.90
F	5.37	6.56	0.09	16.20
Comp	5.24	5.66	0.03	14.78
$D_{fail}$	5.02	5.81	0.06	15.82
SpC	4.75	5.39	0.05	14.78
$E_{max}$	4.5	5.3	0.06	14.4
C1	4.48	4.83	0.04	14.27
C2	4.50	4.88	0.05	14.32
C3	4.51	5.07	0.05	14.23

in a graphical model presents a challenging problem. It is difficult to optimize so many free parameters at once, especially considering we have less training examples at hand than the free parameters. In this paper, we provide an algorithm that automatically selects scale for each part in the model by measuring complementary aspects of descriptor quality i.e. registration accuracy, compactness and consistency.

We show clear improvement in landmark localization with part scales optimized using our algorithm. The average mean and maximum registration error over all landmarks is reduced by 31% and 25% for the optimized model compared to the baseline. Registration accuracy ( $E_{max}$ ) provides a significantly lower error compared to other measures. However, the best combination of measures achieves even slightly lower mean error compared to ( $E_{max}$ ), with a significantly reduced coefficient of variation (4% vs. 6%) and lower maximum mean error (4.83 vs. 5.3 voxels). This suggests the combined measures provide an improved generalization compared to minimizing the mean error directly. We observe that some measures are not useful on their own, but improve generalization when combined with others (e.g.  $E_{max}$  and F). Several other combinations performed nearly as well as the best, in which  $E_{max}$  was complemented by a Cons measure and either Comp or  $D_{fail}$  measure.

Some landmarks remain relative poorly localized. These include bottom sternum, a highly variable structure with ambiguous ground-truth placements and the center of bladder, an intrinsically poorly localizable structure, due to variable fullness and appearance of any contrast agent. Finally, the Th12 vertebra cannot be accurately localized as the available constraints are too broad to disambiguate from unlabeled L1 and Th11.

There are a number of possibilities for future work. We did not attempt to combine optimally the measure scores. Clearly, some qualities are more important for localization and learnt weights would make the improvement of combined measures even more pronounced. Although we used a brute-force search over the parameters, standard optimization techniques may be deployed by adopting the quality measures presented here as cost functions. Importantly, here we focus on optimizing patch size, but any other quantities may be optimized, such as the number of features per part. Our approach takes no account of the spatial prior the graphical model structure and pair-wise constraints remain fixed. Indeed, if many more confidently labeled training samples were available, once could contemplate optimizing the

Table 3: **Summary statistics**. The table shows localization error as we increase the number of free parameters (rows: from shared isotropic to part-specific anisotropic scale, for methods F,  $E_{max}$  and C1). Columns show the mean and maximum error in voxels averaged over all parts, images and bootstraps. "Improvement" is the average % decrease in the mean and max error w.r.t. the baseline.

	Avg	Improvement	Avg	Improvement
	(MeanRMS)	(% per parts)	(MaxRMS)	(% per parts)
Baseline	6.6	-	18.82	-
Best shared scale	5.63	11	15.89	8
Standard measures (F)				
Best isotropic part	5.63	10	15.86	11
Best anisotropic part	5.41	11	16.21	16
All measures ( $E_{max}$ )				
Best isotropic part	5.02	16	15.58	12
Best anisotropic part	4.5	25	14.4	23
Combinations (C1)				
Best isotropic part	4.96	16	15.34	14
Best anisotropic part	4.48	26	14.27	24

part descriptors and the graphical model structure jointly, using measures presented here.

### References

- [1] S. Agarwal, A. Awan, and D. Roth. Learning to detect objects in images via a sparse, part-based representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1475–1490, 2004.
- [2] M. Andriluka, S. Roth, and B. Schiele. Pictorial structures revisited: People detection and articulated pose estimation. In *Computer Vision and Pattern Recognition*, 2009.
- [3] R. Beichel, H. Bischof, F. Leberl, and M. Sonka. Robust active appearance models and their application to medical image analysis. *IEEE Transactions on Medical Imaging*, 24(9):1151–1169, 2005.
- [4] M. Brown, G. Hua, and S. Winder. Discriminative learning of local image descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010.
- [5] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):681–685, 2001.
- [6] J. J. Corso, RS Alomari, and V. Chaudhary. Lumbar disc localization and labeling with a probabilistic model on both pixel and object features. In *MICCAI*, volume 11, page 202, 2008.
- [7] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *International Journal of Computer Vision*, 61(1):55–79, 2005.
- [8] MA Fischler and RA Elschlager. The representation and matching of pictorial structures. *Computers, IEEE Transactions on*, 100(22):67–92, 1973.
- [9] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(10):1615–1630, 2005.
- [10] M. Bergtholdt et al. S. Schmidt, J. Kappes. Spine detection and labeling using a parts-based graphical model. *LECTURE NOTES IN COMPUTER SCIENCE*, 4584: 122, 2007.
- [11] M. Toews and T. Arbel. A statistical parts-based model of anatomical variability. *Medical Imaging, IEEE Transactions on*, 26(4):497–508, 2007.