Accurate 3D Bone Segmentation in Challenging CT Images: Bottom-up Parsing and Contextualized Optimization (Supplementary)

Le Lu Dijia Wu Nathan Lay David Liu Isabella Nogues Ronald M. Summers
National Institutes of Health le.lu@nih.gov Google Inc. Siemens Healthcare

Supplementary Material

We create our ground-truth bone masks by first assigning (0,1) labels to supervoxels. A specific GUI displaying supervoxel mesh models overlaid in 3D CT volume rendering is included to assist with visual assessment. Only $\sim 0.1\%$ of the supervoxels may cross the bone versus non-bone surface boundary. The final bone mask merges all positively labeled supervoxels and is edited with manual touch-ups on voxels by means of an interactive graph-cut system. Direct bone voxel painting based annotation is too time consuming.

We first describe an optional, mean-atlas based probabilistic spatial prior filter to reject non-bone supervoxels. Our module is computationally efficient and differs widely from the conventional "atlas $\rightarrow$ deformable registration $\rightarrow$ label fusion" pipeline [9, 21]. The atlases are three weighted 8-bit 3D volumes (as shown in Fig. 1), derived from Affinity Propagation (AP) [6] based bone mask clustering and the volume uniform-scaling and shifting from landmark alignment [22, 7]. For a new CT image, simple 3D spatial shifting and uniform rescaling parameters are deterministically computed. Hence, each atlas can be aligned with the current volume. This alignment is both efficient in computation and rough in registration accuracy. We do not intend to provide a highly precise registration for the complete set of bone structures in a full-body or arbitrary field-of-view CT image, because this task is very difficult for the "testing in-the-wild" setup (even with high computational resources, e.g., several hours per volume on a modern desktop).

The current registration methodology works best for spatially-confined anatomies without substantial articulations or shape variation, such as the brain and the lung. As a comparative study in the literature [4], the registration based multi-organ localization has a $\sim$ 30% gross failure rate. Learning methods work significantly better for such tasks. The initial atlas-to-volume registration results are noisy but informative. They are robustly encoded with six descriptive features per supervoxel as spatial prior votes received by the atlas alignment. Furthermore, results in Table 1 indicate that classification on these voting features trained from the (+/-) labeled supervoxels improves performance on the segmentation task. For a limited FOV target volume, the atlases (after alignment) can possibly occupy space beyond its image dimensions. The "outside" parts of registered atlases are simply removed or ignored, since no supervoxels will receive their votes.

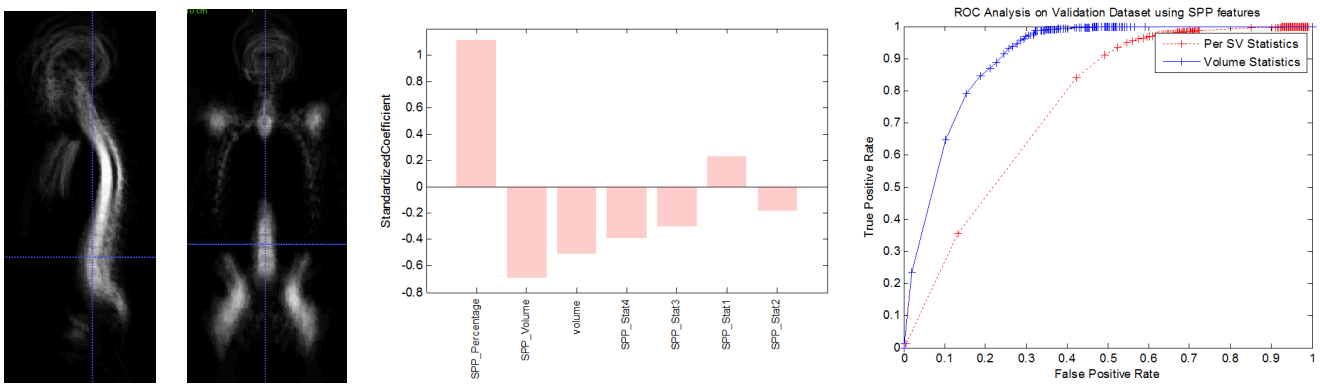
1. Soft Spatial Prior

From 57 training volumes, we do the following: **1)**, we isotropically resample the corresponding annotated label mask into a coarse spatial resolution and align the mask and the image using the centroid position of the T-6 (thoracic) vertebra. The resolution unit is patient-specific. It is set to $1/5$ of the T-6 vertebra height, and can be automatically provided when a spine labeling tool (similar to [22, 7]) is available. The T-6 vertebra is chosen because it is located near the center of the body and appears in full-body scans with high probability. Other landmarks are also applicable. Hence, the alignment only involves patient-normalized scaling and shifting and can be computed efficiently. **2)**, After mapping the mask volumes into the same coordinate system, we compute their pairwise affinity matrix A by

$$A_{ij} = exp(-\frac{1 - Dice(M_i, M_j)}{\sigma}), \qquad (1)$$

where $Dice(M_i, M_j)$ is the Dice coefficient between two normalized bone mask volumes M_i, M_j. **3)**, The Affinity Propagation (AP) algorithm [6] is applied to A in two recursive rounds, in which 17, then 3 clusters are formed. The cluster size, selected by the AP algorithm, decreases from 17 to 3 for computational efficiency purposes. **4)**, Finally, we project the 57 mask volumes M into three coordinate spaces according to their assignments from clustering. The resulting atlas volumes $\widehat{M}_{1,2,3}$ receive and add votes from Ms' projections, in order to form 8-bit intensity-valued 3D volumes after histogram equalization. An atlas example is shown in Fig. 1 (a,b).

During run-time, an atlas selection procedure is required for unseen data. We simply register each of the 3 atlases via spatial shifting and rescaling for an incoming volume. Their



(a) (b) (c) (d)

Figure 1. Spatial prior atlas construction and its influence on supervoxel pruning. (a) one of three atlases in sagittal view; (b) in coronal view; (c) plot of standardized feature weights in classifier; (d) ROC curve of the trained classifier on validation datasets.

element-wise dot-products $\widehat{Z}_{1,2,3}$ are then computed. We select the atlas with the maximal value among $\widehat{Z}_{1,2,3}$ (analogous to recovering higher intensity bone voxels). For 57 training volumes, this atlas selection process yields 100% accuracy. We integrate the spatial prior (SP) in a soft way, by moving to supervoxel features for classification. Six SP features per V_i are calculated:

- the percentage of voxels having non-zero votes from registered $\widehat{M}$,
- the indicated bone volume,
- and the 1st through 4th order moments of $\widehat{M}$ readings within V_i.

As expected, in any supervoxel, the classification method used to obtain $\mathbf{C}_{prior}$in section 3 weights the first feature as the purity of $\widehat{M}$ coverage most significantly. From Fig. 1 (c,d), negative supervoxels occupying $\sim 65\%$ of volumes (among the volumes preserved after voxel thresholding) can be rejected, without sacrificing volume sensitivity.

Note that if the T-6 vertebra is presented and detected in a CT image (as the only requirement), this spatial prior can be directly applied to arbitrary FOV bone segmentation, with no need for model retraining. In some cases, this will cause a few sections of the spatial prior template to fall outside of the CT scan coordinate system. Fortunately, this will not affect supervoxel classification. The inner portion of the spatial template behaves in its normal fashion for supervoxels to receive their weights, which are encoded as features for the classification of $\mathbf{C}_{prior}$. For a very rare patient pose distribution, this averaging model is expected to contribute less. Our data-driven, mode-seeking clustering procedure based on the AP algorithm successfully captures the major modes in the patient pose space of 57 training scans (Refer to Fig. 1).

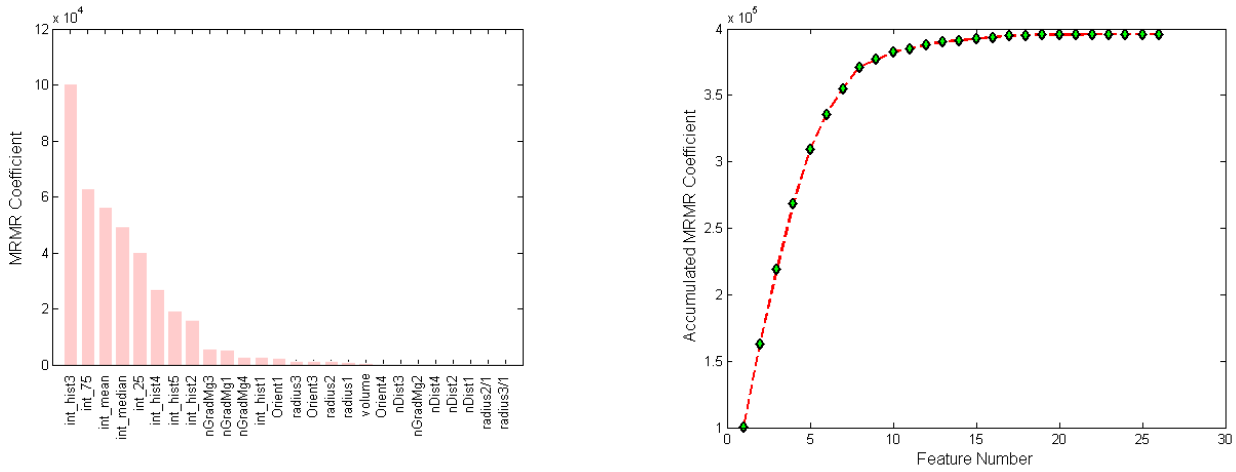
The standalone bone segmentation accuracy of soft spatial prior is far from satisfactory. We incorporate it as an effective pre-filtering stage (before the main feature computation and superpixel classification via CRF) to reject supervoxels early on, in order to maximize both computational efficiency and overall system performance.

2. Sparse Feature Selection

To further evaluate our sparse feature selection and weighting optimization (Sec. 3), we compare it with the well known feature selection method described in [17]. This method employs minimal-redundancy-maximal-relevance mutual information based criteria (MRMR). It can be trained efficiently, and it returns a compact feature set. The MRMR feature selection results are provided in Fig. 2. We show MRMR feature weights or coefficients of 26 sorted features. Both algorithms allow for understanding the relative weights and highlight the importance of features, in the joint optimization objectives of classification or feature selection respectively. By comparing Fig. 2 (a) to Fig. 5 (b) (in the main submission), one can see that both methods weight intensity features more heavily than the other two feature groups. Boundary (contrast) features also seem to be more informative than shape features. This indicates that both algorithms are consistent. The main difference is that our feature weights are more balanced, i.e., more evenly distributed over a certain number of features, than those from the MRMR feature selection algorithm. The volume feature also plays a more significant role in our sparse linear classification of $\mathbf{C}_1$.

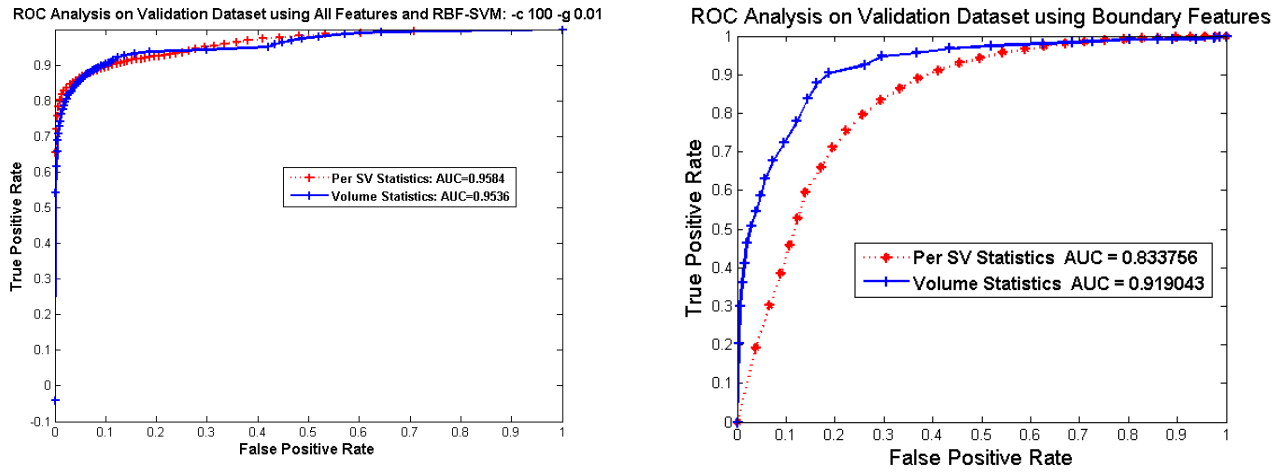
2.1. Comparative Analysis on Classifier and Feature Performance

To evaluate a classifier, training time and test accuracy are the most important factors, as mentioned in [15]. Training RBF or linear kernel SVM classifiers using LibSVM



(a) (b)

Figure 2. Results of feature selection by minimal-redundancy-maximal-relevance criterion [17]. (a) MRMR feature weights or coefficients of 26 sorted features; (b) the plot of accumulated coefficients.



(a) (b)

Figure 3. (a) Plot of validation ROC curve of per-V_i count or volume mm^3 statistics; using RBF kernel SVM [3] with $C = 100$ and kernel size $g = 0.01$. (b) ROC curves of per-V_i or volume mm^3 statistics when only 12 boundary statistics features are used.

[3] requires $10 \sim 12$ hours on a single workstation. On the same training dataset, our optimization process in Sec. 4 takes 1 minute. On the other hand, when optimized properly, the linear classifier can match or outperform nonlinear classifiers in test accuracy [15, 8]. In our case, we have a reasonably large dataset of $78,997$ training samples. However, they are not i.i.d. (independently and identically distributed), but are highly correlated (extracted from 57 training CT scans). 3D volumetric structures or supervoxels may also have higher intrinsic dimensions [13] to be represented. To avoid the curse-of-dimensionality issue and optimize generality on unseen data, a constrained linear classification model can be more desirable (e.g., faster with comparable or higher accuracy) than nonlinear models. This is more clear in medical imaging based applications. Our empirical observation and experimental evaluation coincides with the recent findings in [15, 8], where linear models achieve better trade-offs in training time, testing time and testing accuracy levels.

We compare the classification results of the base Bayesian linear classifier $\mathbf{C}_1$) to those of the linear classifiers, RBF kernel SVM classifiers and random forest ($TreeBagger()$ in Matlab), in terms of accuracy and testing speeds. On the validation datasets, $\mathbf{C}_1$ marginally outperforms other classifiers ($3 \sim 7\%$ higher recall at at 10% FP rate), and AUC (Area Under Curve) values on the complete range of curves are similar. In practice, only the sensitivity levels at low false positive rates (e.g., at most $\leq 15\%$)

are problem-relevant. The best performing SVM classifier uses an RBF kernel with the optimized parameters $C = 100$ and kernel size $g = 0.01$, through grid search on C and g and cross-validation. The ROC plot is shown in Fig. 3 (a). This obtained RBF-SVM has a total of $11,681$ nonlinear support vectors, which will need a significantly longer evaluation time (i.e., $> 10,000$ times) in testing than a single linear dot-product and sigmoid execution by our model $\mathbf{C}_1$. The testing time requirement for a random forest using 50 trees is also greater than $\mathbf{C}_1$'s, but is lesser than that of the RBF kernel SVM classifier. Thus, the optimized sparse linear classifier $\mathbf{C}_1$ may offer better generality than the highly nonlinear models, especially the nonlinear SVMs recently used in the winning PASCAL challenge studies [2, 1, 5] (not including the SVM from [15]), in our under-sampled learning scenario.

Using all three types of supervoxel derived image features for training does achieve the best ROC curve, compared to the results that only use individual groups. The ranks of group importance are *Intensity features*, *Boundary features* and *Shape features*, in descending order. In Fig. 3 (b), ROC curves of per-V_i and volume mm^3 statistics from only 12 boundary statistics features are illustrated. Its AUC values drop from 0.9765 to 0.9190 , unlike the case where all features are included for sparse training. All features are normalized to zero mean and unit standard deviation before training.

3. Zoned Piecewise Linear Classification

We experiment with training zoned classification for CT body ranges as upper and lower parts (i.e., upper or lower regions from the vertical position of the center of cervical vertebra C-6). This forms a piecewise linear classification architecture, rather than a single linear separation boundary in feature space. As shown in Fig. 4 (a,c), the volume-wise validation AUC improves to 0.9924 for $\mathbf{C}_{1U}$ and 0.9681 for $\mathbf{C}_{1L}$. These results indicate that the upper skull and neck zone can achieve nearly perfect classification results, while the lower torso and limb zones are harder to classify. Furthermore, interesting observations are found by comparing the standardized feature weights in Fig. 4 (b,d). In the upper zone, the intensity feature group mostly dominates the relative importance in classification. Intensity, volume, shape and boundary features more evenly contribute to classification in the lower zone. This divide-and-conquer scheme indicates that the discriminative feature distributions of positive and negative data samples vary spatially. When both $\mathbf{C}_{1U}$ and $\mathbf{C}_{1L}$ are used instead of $\mathbf{C}_1$ alone, the AUC value increases from 0.9765 to 0.9846 in mm^3 volume statistics. Though this difference seems subtle, it is significant, as it can increase recall levels in the low FP range.

4. More Visual Examples & Others

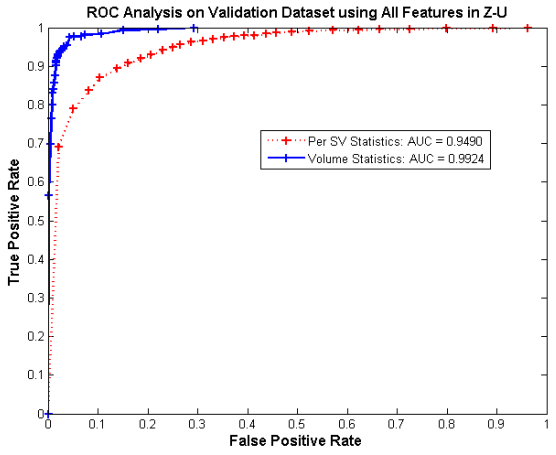
Our method runs in $2 \sim 3$ minutes per CT volume, under all tested conditions for data resolution, fields-of-view and possible pathologies. To the best of our knowledge, this paper presents the first complete solution for precise full or arbitrary field-of-view bone masking in 3D CT images. Previous works focus on segmenting specific bones, mainly the spine [18, 14, 10, 19], the knees [20], and the hip [16]. [18] segments only the spine columns (not including more complexly-shaped vertebra processes) in 20 CT studies and depends on a spine detection procedure. The technique used in [14] is evaluated only on thoracic spines for 40 datasets and requires pre-alignment initialization from the trained deformable model to the testing image data. The main application of [10] is spine disk detection under an iterative geometric constraint in 30 CT volumes, which may heavily limit its success on scoliosis cases. [14, 10] treat individual vertebrae or vertebra disks as separate objects to segment. [16] performs hip segmentation for 12 CT volumes via voxel labeling and graph-cut approaches. [20] uses 23 CT scans to validate and enhance their bone segmentation technique by shape modeling and matching on the knees. Our method considers all bone structures as the single-labeled foreground and all remaining voxels as background. No statistical shape model based constraints are applied. Qualitative examples of bone segmentation or masking in different views are provided in Fig. 5. Finally, two 3D rendering views of a segmented skeleton mask for a patient with moderate scoliosis are shown in Fig. 6.

Our main quantitative evaluation metric is based on Dice coefficient as a type of volume overlap ratio between the pair of ground truth and segmented bone masks. The reason is as follows. Computing geometry consistent 3D mesh surfaces of any full or arbitrary FOV bone mask is non-trivial. Any **degenerate scenario** may occur. For instance, if the cortical bones are perfectly segmented but marrow tissues are missed (a common situation for the bone removal software), the converted 3D bone mesh will contain both exterior (bone vs. non-bone) and interior (cortex vs. marrow) surfaces. Thus, the final surface-to-surface metrics can be good while recall and Dice measurements are undesirably low. Our 137 patient datasets represent a wide variety of body shapes, body mass indices, pathologies, fields-of-view and body postures, which may cause severe challenges for these model based top-down methods [23, 12, 14, 10, 11].

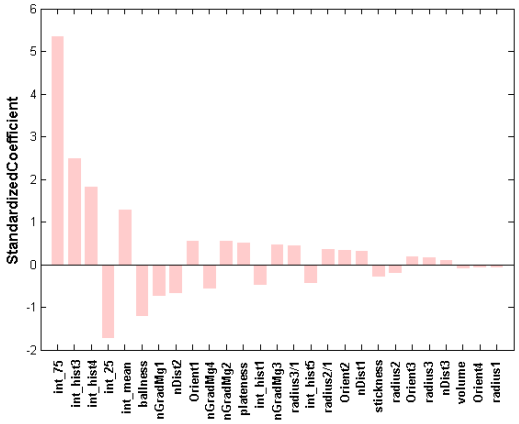
References

[1] X. Boix and et al. Harmony potentials - fusing global and local scale for semantic image segmentation. *Int. J. Comp. Vis.*, pages 83–102, 2012. 4

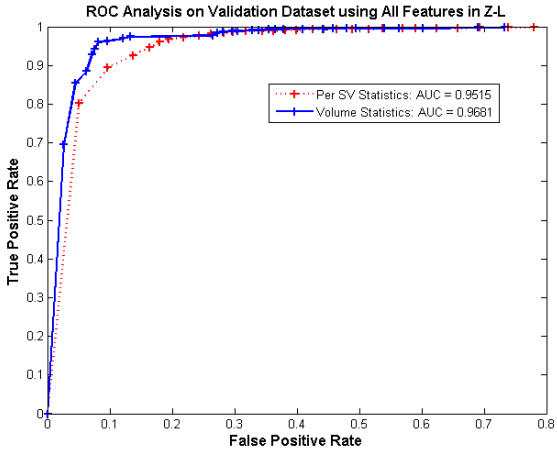
[2] J. Carreira, F. Li, and C. Sminchisescu. Object recognition by sequential figure-ground ranking. *IJCV*, 2012. 4



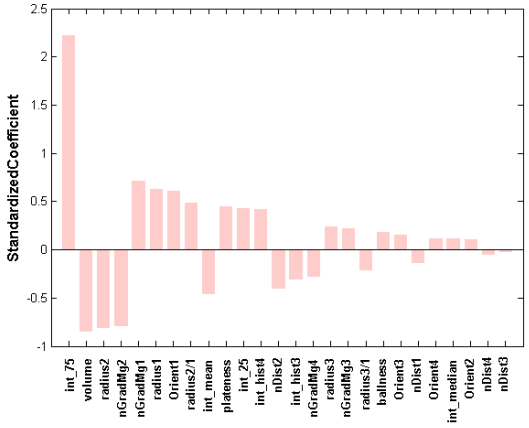
(a)



(b)



(c)



(d)

Figure 4. Plots of zoned classification on supervoxels. (a,c) validation ROC curves of per-V_i count or volume mm^3 statistics; (b,d) plots of standardized feature weights in for $\mathbf{C}_{1U}$ or $\mathbf{C}_{1L}$, respectively.

[3] C.-C. Chang and C.-J. Lin. Libsvm : a library for support vector machines. *ACM Trans. Intell. Sys. Tech.*, 2:1–27, 2011. 3

[4] A. Criminisi, D. Robertson, E. Konukoglu, J. Shotton, S. Pathak, S. White, and K. Siddiqui. Regression forests for efficient anatomy detection and localization in computed tomography scans. *Medical Image Analysis*, 17(8):1293–1303, 2013. 1

[5] M. Everingham, L. V. Gool, C. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge (voc2012) results. 2012. 4

[6] B. Frey and D. Dueck. Clustering by passing messages between data points. *Science*, pages 972–976, 2007. 1

[7] B. Glocker, D. Zikic, and et al. Vertebrae localization in pathological spine ct via dense classification from sparse annotations. *MICCAI*, 2013. 1

[8] B. Hariharan, J. Malik, and D. Ramanan. Discriminative decorrelation for clustering and classification. *ECCV*, 4:459–472, 2012. 3

[9] I. Isgum, M. Staring, and et al. Multi-atlas-based segmentation with local decision fusionapplication to cardiac and aortic segmentation in ct scans. *IEEE Trans. Med. Imaging*, 28(7):1000–1010, 2009. 1

[10] M. Kelm, M. Wels, and et al. Spine detection in ct and mr using iterated marginal space learning. *Med Image Anal.*, 17(8):1283–92, 2012. 4

[11] T. Klinder, J. Ostermann, M. Ehm, A. Franz, R. Kneser, and C. Lorenz. Automated model-based vertebra detection, identification, and segmentation in ct images. *Medical Image Analysis*, 13:471–481, 2009. 4

[12] H. Ling and et al. Hierarchical, learning-based automatic liver segmentation. *CVPR*, 2008. 4

[13] L. Lu, A. Barbu, and et al. Accurate polyp segmentation for 3d ct colonography using multi-staged probabilistic binary learning and compositional model. *CVPR*, 2008. 3

[14] J. Ma and L. Lu. Hierarchical segmentation and identification of thoracic vertebra using learning-based edge detection

Figure 5. Qualitative examples of bone segmentation or masking in different views. Datasets demonstrate large variations in body shape, body mass index and pathology with the patient population under study. (a,b,c) are sagittal views and (c) shows a mild scoliosis case; (d,e,f) are coronal views and (d) illustrates a small false positive bone segment near femur caused by a metal implant; (g,h,i) are axial views in different body sections (abdomen spine, pelvis and neck).

and coarse-to-fine deformable model. *CVIU*, 117(9):1072–83, 2013. 4

[15] S. Maji. Linearized smooth additive classifiers. *ECCV Workshop on Web-scale Vision and Social Media*, 2012. 2, 3, 4

[16] D. Malan, C. Botha, and E. Valstar. Voxel classication and graph cuts for automated segmentation of pathological

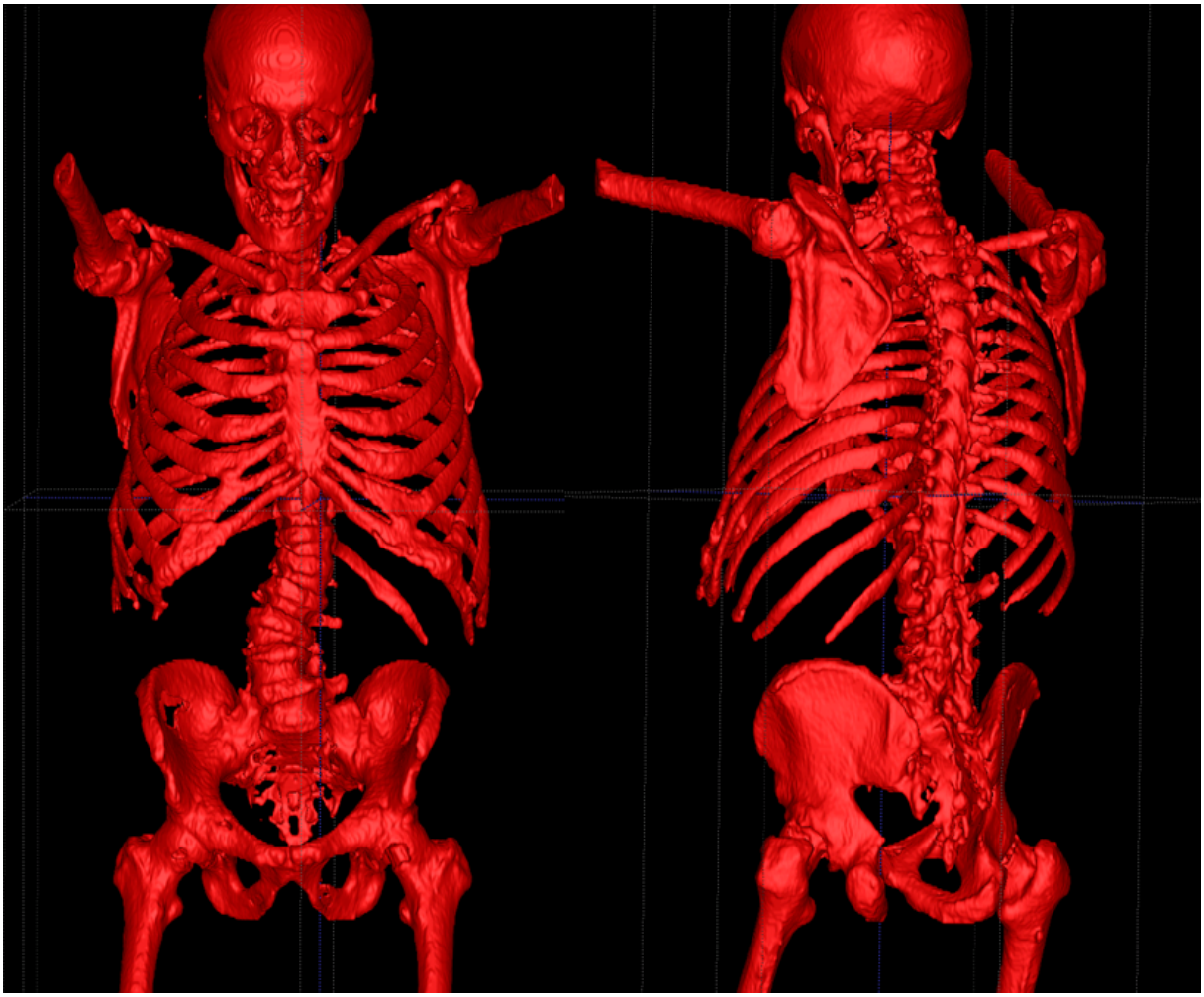


Figure 6. A segmented 3D full-body skeleton bone mask shown from two different view points, for a patient with moderate scoliosis. Different 3D shapes and appearance characteristics over various parts of skeleton, from the skull to the spine, ribs, pelvis and limbs are apparent.

periprosthetic hip anatomy. *Int. J. CARS*, 8:63–74, 2013. 4

[17] H. Peng, F. Long, and C. Ding. Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and minredundancy. *IEEE TPAMI*, pages 1226–1238, 2005. 2, 3

[18] M. Schwier, T. Chitiboi, T. Hulnhagen, and H. Kahn. Automated spine and vertebrae detection in ct images using object-based image analysis. *Int. J. Numer. Meth. Biomed. Engng.*, 29(9):938–63, 2013. 4

[19] S. Tan, J. Yao, and et al. Computer aided evaluation of ankylosing spondylitis using high-resolution ct. *IEEE Trans. Med. Imaging*, 27(9):1252–67, 2008. 4

[20] L. Wang, M. Greenspan, and R. Ellis. Validation of bone segmentation and improved 3-d registration using contour coherency in ct data. *IEEE Trans. Med. Imaging*, 25(3):324–334, 2006. 4

[21] R. Wolz, C. Chengwen, K. Misawa, M. Fujiwara, K. Mori, and D. Rueckert. Automated abdominal multi-organ segmentation with subject-specific atlas generation. *IEEE Trans. Med. Imaging*, 32(9):1723–1730, 2013. 1

[22] Y. Zhan, M. Dewan, M. Harder, and X. Zhou. Robust mr spine detection using hierarchical learning and local articulated model. *MICCAI*, 2012. 1

[23] Y. Zheng, A. Barbu, and et al. Four-chamber heart modeling and automatic segmentation for 3-d cardiac ct volumes using marginal space learning and steerable features. *IEEE Trans. Med. Imaging*, 27(11):1668–1681, 2008. 4