Organ Localization Using Joint AP/LAT View Landmark Consensus Detection and Hierarchical Active Appearance Models

Qi Song[1], Albert Montillo[1(✉)], Roshni Bhagalia[1], and V. Srikrishnan[2]

[1] General Electric Global Research, Niskayuna, NY, USA
[2] General Electric Global Research, Bangalore, India
{song,bhagalia,montillo}@ge.com

Abstract. Parsing 2D radiographs into anatomical regions is a challenging task with many applications. In the clinic, scans routinely include anterior-posterior (AP) and lateral (LAT) view radiographs. Since these orthogonal views provide complementary anatomic information, an integrated analysis can afford the greatest localization accuracy. To solve this integration we propose automatic landmark candidate detection, pruned by a learned geometric consensus detector model and refined by fitting a hierarchical active appearance organ model (H-AAM). Our main contribution is twofold. First, we propose a probabilistic joint consensus detection model which learns how landmarks in *either or both* views predict landmark locations in a given view. Second, we refine landmarks by fitting a joint H-AAM that learns how landmark arrangement and image appearance can help predict across views. This increases accuracy and robustness to anatomic variation. All steps require just seconds to compute and compared to processing the scouts separately, joint processing reduces mean landmark distance error from 27.3 mm to 15.7 mm in LAT view and from 12.7 mm to 11.2 mm in the AP view. The errors are comparable to human expert inter-observer variability and suitable for clinical applications such as personalized scan planning for dose reduction. We assess our method using a database of scout CT scans from 93 subjects with widely varying pathology.

Keywords: Automatic landmark localization · Organ localization · Image parsing · CT · Hierarchical active appearance model · Rejection cascade

1 Introduction

Many medical imaging protocols rely on 2-D radiographs for patient specific organ localization, which facilitates a variety of clinical applications including scanner set-up and scan planning, precise organ segmentation, semantic navigation and structured image search. Manual organ localization can be time consuming, impede workflow and often suffers from large operator errors. Automatic

Authors 'Q. Song' and 'A. Montillo' contributed equally.

B. Menze et al. (Eds.): MCV 2013, LNCS 8331, pp. 138–147, 2014.
DOI: 10.1007/978-3-319-05530-5_14,

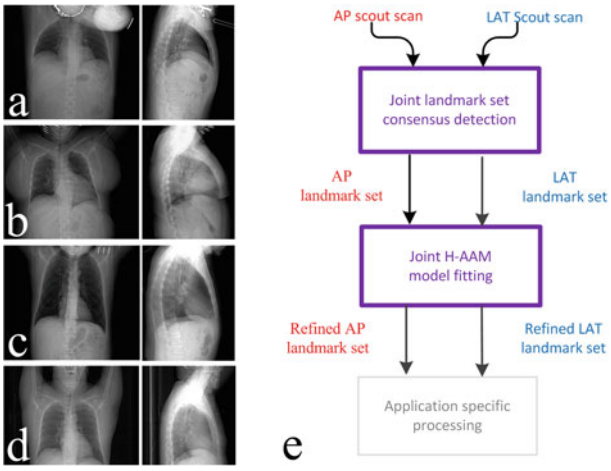

Fig. 1. Image analysis challenges and proposed solution. (a–d) Image paris consist of AP (left) and LAT (right) views. (e) Our method consists of two steps: joint landmark set detection followed by joint H-AAM organ localization.

localization from 2-D radiographs is therefore urgently needed. In this paper, we enable the automatic parsing of 2D radiographs from ubiquitous clinical CT scans. Such scans routinely include both a 2D anterior-posterior (AP) scout and a lateral (LAT) projection scout image. Automatic organ localization from 2-D scout images is a very challenging task due to low image quality from high noise level and low image contrast. Furthermore, scout images are 2-D projections of three dimensional data and as such have greatly reduced image information due to significant tissue overlap compared to volumetric scans. Representative 2D scout images are shown in Fig. 1a–d.

We hypothesize that an image analysis method combining information from AP and LAT views will afford the greatest localization accuracy. Our proposed solution (Fig. 1e) has two steps. First a set of landmarks delineating the boundaries of salient organs is extracted from the image pair though a joint consensus detector which removes outliers from the set of landmark candidates detected on AP *and* LAT views. This organ localization is further refined by fitting a hierarchical active appearance organ model (H-AAM) to the image pair.

Previous methods using landmark detection to parse radiographs include [5,8]. In [8] false negatives are not inferred nor are the detections refined with a joint H-AAM which we show substantially improve accuracy. In [5] the landmark detection uses only a single AP-only model and does not handle LAT images. It is essential to process both scouts because their orthogonal views provide complementary organ location information. Parsing 3D CT volumes using landmark detection has been presented [6], where the landmark detections are refined by searching exemplar cross-correlation maps. Active shape model (ASM) [3] and active appearance model (AAM) [1] have also been reported to combine with

landmark detection approach. In [7], an active shape model based refinement was applied after landmark detections. In [2], the shape model fitting is driven by a random forest regression voting. Neither of these methods directly applies to soft tissue localization in radiographs. This is because the projective image formation causes multiple structures to overlap making direct application of ASMs error prone and because the non-Hounsfield pixel intensities make cross-correlation maps problematic.

2 Methods

Our method (Fig. 1e) consists of two steps: (1) joint landmark set consensus detection for an initial organ localization, (2) refinement by joint H-AAM fitting. The following sections describe each step.

2.1 Joint Landmark Set Detection

Joint landmark set detection consist of two substeps. We begin with the input which consists of a pair of 2D scout images, one for the AP scout, denoted I_{AP}, and one for the LAT scout, I_{LAT} (Fig. 2a). These are processed separately using an *individual "sliding-window" patch detector* for each landmark. One set of detectors searches I_{AP} and outputs a set of candidate landmark locations C_{AP}, while another set searches I_{LAT} and outputs candidate locations C_{LAT}. Detectors are run in parallel. In general, the output candidate sets contain false positives and negatives. Both are corrected by applying a *joint landmark set consensus detector* in Fig. 2a (box 3). This employs a greedy approach that iteratively removes the least likely candidate, considering the set of candidates recovered from both views and the probabilistic anatomy (landmark constellation) model. The result after consensus detection is a consistent N-labeling of the N landmarks for each subject. These N labels consist of landmarks for the AP scout, $L_{AP,1}$, and for the LAT scout $L_{LAT,1}$.

Training and Applying Landmark Detectors. Each individual landmark detector is trained as a two-category rejection cascade classifier [9], Fig. 2e, using supervised learning. Each cascade stage is a Gentle Adaboost [4] classifier.

To train we need positive landmark patches and negative patches. Positives come from cropping a rectangular patch around each manually annotated landmark. As illustrated in Fig. 2b we manually label 21 *AP landmarks* including: heart-diaphragm intersection (1, 11), diaphragm peak (2), lung corners (3, 19), left most in left lung (15), lung sides at 1/3 and 2/3 the arc length to top (4, 5, 18, 17), top of lungs (6, 16), airway-lung intersections (7, 13), heart top (14), heart sides (8, 12) at 1/2 arc length to top, ends of diaphragm near heart (9, 10), lower rib cage beneath lungs (20, 21). As shown in Fig. 2c we use 13 *LAT landmarks*: ends of diaphragm (1, 13), spine-diaphragm intersection (2), top of lung (5), lung side (3, 12) at 2/3 the arc length to top, posterior of spine (4) at 2/3 the arc length to lung top, heart top (6), heart side (7), heart-diaphragm intersection

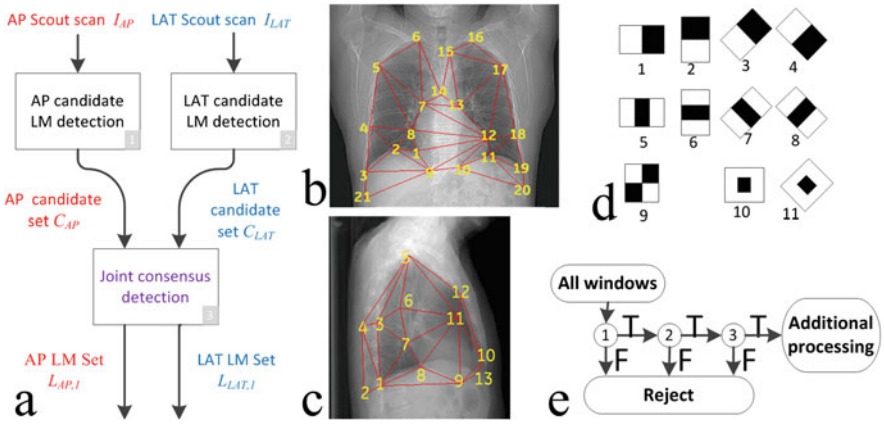

Fig. 2. Joint landmark set detection. (a) Landmark detectors scan input images producing landmark candidates; then a joint consensus detector corrects false positives and negatives. (b, c) Detectors are trained from positive and negative landmark patches dropped from images with manually labeled landmarks on lung and heart boundaries. (d) Haar image features (e) Rejection cascade based detectors.

(8), bottom of heart (9), right most of heart (10), heart side (11) at 2/3 the arc length bottom to top. These landmarks delineate lung and heart boundaries. The positive exemplars for each landmark are image patches large enough to include visible anatomical structure around the landmark. Negative exemplars are randomly cropped from the image that overlap the positive by <40 %. Haar image features (Fig. 2d) are computed efficiently using integral images [9]. Each cascade stage is trained to achieve a true positive rate of 99.7 % with a false positive rate of 50 % and stages are added until a desired overall true/false positive rate is reached or a maximum number of stages (15) is achieved.

Joint Landmark Set Consensus Detector. Applying the landmark detectors yields a set of candidate detections, $C = C_{AP} \cup C_{LAT}$. There can be multiple detections per landmark (false positives) and landmarks that were not detected but are present (false negatives). To correct for both cases we use a consensus detector to remove the false positives and infer the false negatives. There are two phases of consensus detection: training and application of the trained model which are described next.

Phase 1, Training: Training learns a probabilistic model of the global geometric arrangement of the landmarks in the N-landmark constellation. Given the manually labeled N-landmark set for each pair of training images, then for each **target** landmark i, and for each pair of **voting** landmarks from the remaining N-1, we learn the multivariate Gaussian distribution of the relative position of i to the location of the pair. Each distribution encodes the probability the target landmark i is at any location in the image plane, conditioned on the location of

the voting pair. Specifically, denoting the N-landmark set as $L = L_{AP} \cup L_{LAT}$, and the pair of distinct voting landmarks as $s_i \subset L$, we learn the parameters, $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ of the multi-variate Gaussian distribution for each target landmark, $q \in L$ and $q \notin s_i$ using maximum likelihood estimation (MLE).

To formulate the optimization, we begin by letting the coordinates of the AP image be (x, z); those of LAT image be (y, z). We use linear regression to model the dependency of the target landmark q's coordinates on the location of the two **voting** landmarks in s_i. The target's coordinates are (x_3, z_3) if from the AP image and (y_3, z_3) if from the LAT image. The voting landmarks can both be from the AP, both from the LAT or one from each. All possible cases of target and voting landmarks are modeled using Eqs. (1)–(6):

$$x_3 = \alpha_0 + \alpha_1 x_1 + \alpha_2 z_1 + \alpha_3 x_2 + \alpha_4 z_2 \quad z_3 = \beta_0 + \beta_1 x_1 + \beta_2 z_1 + \beta_3 x_2 + \beta_4 z_2 \tag{1}$$

$$x_3 = \alpha_0 + \alpha_1 x_1 + \alpha_2 z_1 + \alpha_3 y_2 + \alpha_4 z_2 \quad z_3 = \beta_0 + \beta_1 x_1 + \beta_2 z_1 + \beta_3 y_2 + \beta_4 z_2 \tag{2}$$

$$x_3 = \alpha_0 + \alpha_1 y_1 + \alpha_2 z_1 + \alpha_3 y_2 + \alpha_4 z_2 \quad z_3 = \beta_0 + \beta_1 y_1 + \beta_2 z_1 + \beta_3 y_2 + \beta_4 z_2 \tag{3}$$

$$y_3 = \alpha_0 + \alpha_1 y_1 + \alpha_2 z_1 + \alpha_3 y_2 + \alpha_4 z_2 \quad z_3 = \beta_0 + \beta_1 y_1 + \beta_2 z_1 + \beta_3 y_2 + \beta_4 z_2 \tag{4}$$

$$y_3 = \alpha_0 + \alpha_1 y_1 + \alpha_2 z_1 + \alpha_3 x_2 + \alpha_4 z_2 \quad z_3 = \beta_0 + \beta_1 y_1 + \beta_2 z_1 + \beta_3 x_2 + \beta_4 z_2 \tag{5}$$

$$y_3 = \alpha_0 + \alpha_1 x_1 + \alpha_2 z_1 + \alpha_3 x_2 + \alpha_4 z_2 \quad z_3 = \beta_0 + \beta_1 x_1 + \beta_2 z_1 + \beta_3 x_2 + \beta_4 z_2 \tag{6}$$

Using the voting pair s_i, we model the probability that the target is at any location $\boldsymbol{x}$ in the kth training image as a multivariate Gaussian:

$$p_k(\boldsymbol{x}) = \frac{1}{\sqrt{2\pi|\boldsymbol{\Sigma}_{ki}|}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_{ki})^T \boldsymbol{\Sigma}_{ki}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_{ki})}.$$

The unknown coefficients from the appropriate pair of linear regression equations (1)–(6) can be used to form a projection matrix:

$$\boldsymbol{A}_i = \begin{pmatrix} \alpha_0 & \beta_0 \\ \alpha_1 & \beta_1 \\ \alpha_2 & \beta_2 \\ \alpha_3 & \beta_3 \\ \alpha_4 & \beta_4 \end{pmatrix} \tag{7}$$

Similarly, given K total training LAT/AP image pairs, the coordinates of the voting landmarks can be expressed compactly as $\boldsymbol{P}_s$ (where x becomes x or y depending on AP or LAT) and the target coordinates as $\boldsymbol{P}_t$ using:

$$\boldsymbol{P}_s = \begin{pmatrix} 1 & x_{11} & z_{11} & x_{21} & z_{21} \\ 1 & x_{12} & z_{12} & x_{22} & z_{23} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1k} & z_{1k} & x_{2k} & z_{2k} \end{pmatrix} \tag{8}$$

$$\boldsymbol{P}_t = \begin{pmatrix} x_1 & x_2 & \ldots & x_k \\ z_1 & z_2 & \ldots & z_k \end{pmatrix} \tag{9}$$

We compute the projection matrix via MLE using $\boldsymbol{A}_i = (\boldsymbol{P}_s^T \boldsymbol{P}_s)^{-1}(\boldsymbol{P}_t \boldsymbol{P}_s)^T$. Then the mean and covariance parameterizing the Gaussian are computed from: $\boldsymbol{\mu}_i = \boldsymbol{P}_s \boldsymbol{A}_i$ and $\boldsymbol{\Sigma}_i = cov(\boldsymbol{P}_t^T - \boldsymbol{\mu}_i)$. Note that even if the AP and LAT scans are not aligned well our method *still works well* because our model learns the distribution of AP/LAT misalignments.

Phase 2, application of the trained model: First we iteratively prune false positives, similar to [7,8]. At each iteration we remove the candidate least likely to be valid. Candidate likelihood is the maximum probability of the candidate, computed from the Gaussian distributions given its relative position to all other pairs of landmark candidates. Lowest probability candidate is removed if its probability is $<\tau$, an empirically determined threshold. Iterations stop when the lowest probability $>\tau$.

Next we infer the location of false negative landmarks, which is unique to our method and not found in [7,8]. Given C our set of candidate detections, we let P be the set of landmarks spanned by C. The undetected landmarks are $U = L \setminus P$. *For each undetected landmark* $u \in U$, we infer its location, $\boldsymbol{x}$, using predictions from the detected candidates. We compute a location estimate for each subset $c_k \subset P$ of two candidates of distinct landmarks, using the mean offset, $\boldsymbol{\mu}$, from c_k learned in our training dataset. This forms a set of estimates, $E = \{\mathbf{e}_n\}$ where $\mathbf{e}_n = (x_n, z_n)$ for AP image. Our final estimate of $\mathbf{e}_n$ is formed from the trimmed mean of the central 50 % over all estimates in E.

2.2 Joint H-AAM Organ Localization

Joint H-AAM. Like the consensus detector, the active appearance model (AAM) is also a generative learning-based approach. Trained on labelled image data, the model learn both relative positions between different parts of the object and the expected textures inside the ROI. By incorporating both shape and appearance information, AAM-based interpretation leads to a robust solution even in the presence of serious image noise and large structure variation.

In this work, a joint H-AAM approach is introduced, encoding shape and appearance information from both AP and LAT views. Furthermore, a hierarchical pyramid is employed. At the coarse level, a single global joint model is trained on the manually-labelled radiographs of AP and LAT views. All landmarks used to train joint consensus detectors are included in the model. There are 21 landmarks in training image I_{AP} of AP scout and 13 landmarks in I_{LAT} of LAT scout. Through concatenation the shape of the training image pair is represented by a 34 dimensional vector $v = [L_{AP}, L_{LAT}]^T$, where $L_{AP} = [x_1^{AP}, z_1^{AP}, \cdots, x_{21}^{AP}, z_{21}^{AP}]$ is the set of 2D coordinates of landmarks in I_{AP} and $L_{LAT} = [y_1^{LAT}, z_1^{LAT}, ..., y_{13}^{LAT}, z_{13}^{LAT}]$ is the set for I_{LAT}. To obtain the associated appearance information, we construct two triangulated meshes based on these landmarks, one on AP view and one on LAT view (see Fig. 3a). The region inside the mesh is taken as the ROI. A global AAM model is then trained from the v and the ROI of the training images, which encodes the intensity texture from both AP and LAT scouts. Figure 3(b and c) show the constructed mean

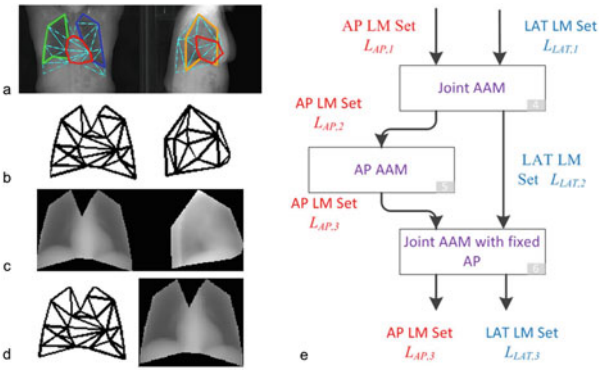

Fig. 3. (a) Triangulated meshes from manually-annotated landmarks give rough locations of lungs (green, blue, orange) and heart (red). (b) Mean joint shape model. (c) Mean joint appearance model. (d) Mean shape and appearance AP sub-model. (e) Joint H-AAM fitting workflow (Color figure online).

shape and the mean appearance of the joint model, respectively. The joint global model captures the probabilistic correlation between structures in both views, which helps infer obscured shapes from other parts and is less sensitive to initialization errors though less flexible than two individual scout models.

In subsequent finer levels of the pyramid, sub-models are trained using scout specific vertices from the global model, allowing better description of local structures and reducing the chance of over-constrained by learning variations in a single view. Figure 3d shows the constructed AP only sub-model. The following section shows how hierarchical model fitting helps localize organs in AP and LAT views.

Hierarchical Model Fitting. Our model fitting workflow is illustrated in Fig. 3e. Initialized to landmark consensus detection results, a joint model incorporating feature points from both AP and LAT scout images is simultaneously fitted to the AP/LAT image pair (Fig. 3e, box 4). Next the localization result on the AP image is refined by applying a sub-model learned from AP images, which is initialized by previous joint model fitting results (Fig. 3e, box 5). We only apply the sub-model for AP scouts because in practice, AP images have more reliable features since the projection image is formed from less tissue overlap than LAT images. Since LAT images have greater structure occlusion more constraints are required to infer organ locations. To further refine LAT locations, we fit a joint model again, during which we leave the AP landmarks fixed. These points serve as reliable "anchor" points, enforcing contextual constraints for LAT landmark refinement.

3 Experiments and Results

We evaluate our approach on 93 subjects from whom both AP and LAT scout images were acquired using four-fold cross validation, i.e. 70 subjects for training and the remainder for testing in each fold. The image size ranges from 888×660 pixels to 888×1026 pixels for AP scout, and 888×660 pixels to 888×935 pixels for LAT scout. The resolution is 0.60×0.55 mm for both scouts. The subjects vary in age, gender, and pathology including obesity (Fig. 1b), lung cancer, and cardiomyopathy. Additional variability includes metallic implants: cardiac stents, hip implants, and jewelry (Fig. 1a, c). Acquisition protocol variations include large variation in the Z range and patient positioning, e.g. arm position (Fig. 1a, d).

Qualitative Evaluation. In Fig. 4a, we compare landmark detection results using separate consensus detection, (top), with those from joint consensus detection, (bottom). True landmarks are shown in dark blue X's, those detected by the method are shown as green and yellow X's, while those inferred using these detections are light blue. Differences are highlighted in yellow; the detection and the corresponding true location are enclosed by a yellow ellipse. We observe these ellipses are much smaller using joint consensus detection than separate detection, indicating higher landmark accuracy. In further analysis we found that every LAT landmark has improved mean accuracy. Figure 4b–d show comparative organ localization results. The fitting results are shown in cyan dashes with right lung (green), left lung (blue), chest cavity (orange) and heart (red). The ground-truth is marked by yellow dash. We observe the joint model yields significant improvement (Fig. 4c) over the single view processing (Fig. 4b) and is further improved by enforcing a joint hierarchical model fitting structure (Fig. 4d).

Quantitative Evaluation. The mean landmark distance error between computed and manually labelled landmarks across all 93 test images is shown in Fig. 5a. Compared to separate view consensus detection, our proposed joint view approach reduces distance error from 12.7 mm to 11.2 mm for AP view and from 27.3 mm to 15.7 mm for LAT view. Joint consensus detection without AAM fitting maintains AP landmarks at 22.3 mm while dramatically reducing error (*by* >14 mm) for LAT view from 32.0 mm to 17.3 mm. Joint *hierarchical* AAM reduces overall distance error for *AP and LAT*, including from 14.0 mm to 11.2 mm for AP and from 17.3 mm to 15.7 mm for LAT compared to joint model fitting only.

A potential application of our method is to determine the bounding box of the heart for cardiac scan range planning. To evaluate method suitability we compare the smallest rectangle containing all landmarks along the heart boundary to heart bounding boxes manually defined by physicians. The unsigned distance errors of the box sides are shown in Fig. 5b–c. Our method improves bottom and all four sides in AP and LAT scouts respectively. Improvement of the bottom side is particularly noteworthy given the high organ occlusion there.

Processing the images at full resolution, landmark detection and joint consensus detection takes about 25 s while joint H-AAM requires about 30 s with a modern desktop computer (4–8 core, 8 GB RAM). Further speedup is achievable through multi-resolution processing.

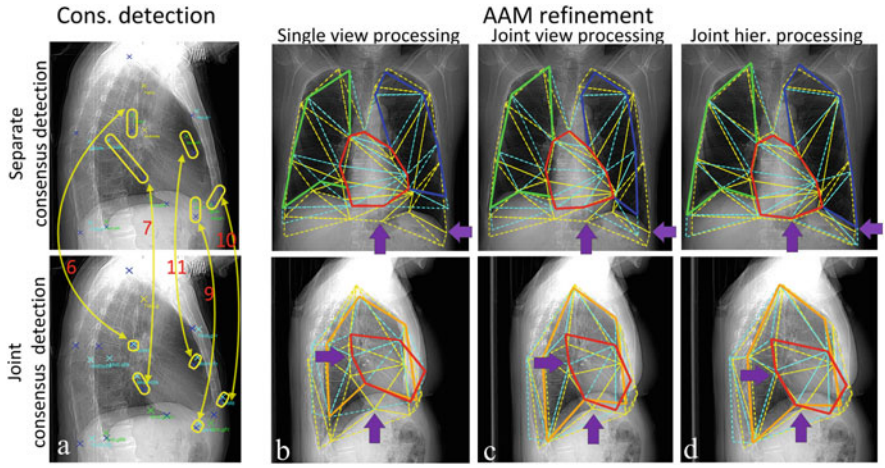

Fig. 4. Impact of joint AP/LAT view processing. (a) Separate detection results (top) have larger landmark errors than joint consensus detection (bottom) for landmarks (red #). (b)–(d) Each step in which we fit our AAM model (shown with red, blue, green, cyan lines) improves fidelity to ground-truth (yellow dash). Compare fitting improvements (purple arrows) among (b) single view AAM; (c) joint view AAM and (d) joint H-AAM (Color figure online).

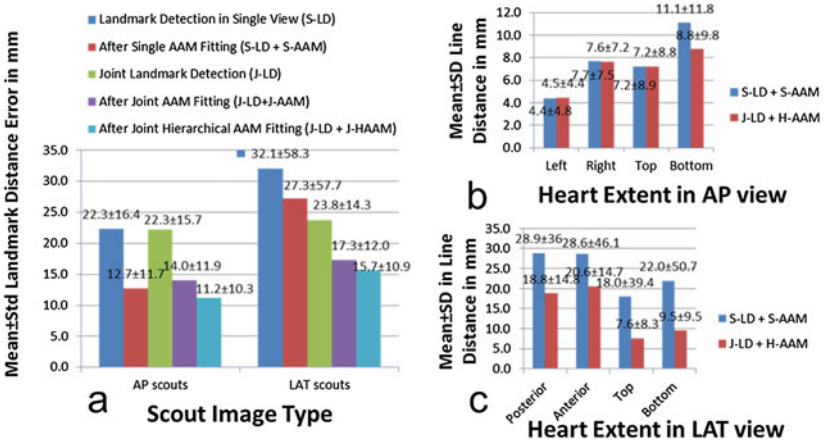

Fig. 5. Cross-validation results from 93 subjects. Proposed joint AP/LAT view approach achieves lowest distance error (a) and heart bounding-box distance error (b, c).

4 Discussion and Conclusions

In this work we address the challenging task of parsing AP and LAT radiographs into salient anatomic structures. To the best of our knowledge, *this work is the first to jointly leverage information from AP and LAT scouts to delineate the*

heart and lungs. We demonstrate that fitting a coarser initial joint hierarchical AAM across AP/LAT views reliably refines the consensus landmark detection results. Further, finer single-view-only models can be subsequently applied for a final round of refinement. Using joint landmark detection and joint H-AAM fitting reduces mean distance error in LAT landmarks from 27.3 to 15.7 mm. This is an improvement of over 40 percent compared to using only LAT scout scans, where features are inherently more difficult to localize due to greater overlap of structures. Lastly, compared to separate view processing, our joint view approach reduces overall mean landmark distance error from 12.7 mm to 11.2 mm in the AP view and from 27.3 mm to 15.7 mm in LAT view. For the AP scout our error of 11.2 mm compares well to the mean human expert inter-observer variability of 10.2 mm while our error for the LAT scout of 15.7 mm compares to the human inter-observer error of 14.3 mm. These inter observer errors were computed using manual landmark estimates obtained from two independent observers. Our algorithm achieves a level of accuracy sufficient to enable clinically relevant tasks such as reducing radiation for the patient through personalized scan planning and to facilitate consistent longitudinal scanning in the clinic, and such clinical productization has already begun.

References

1. Cootes, T., Edwards, G., Taylor, C.: Active appearance models. IEEE Trans. Pattern Anal. Mach. Intell. **23**(6), 681–685 (2001)
2. Cootes, T.F., Ionita, M.C., Lindner, C., Sauer, P.: Robust and accurate shape model fitting using random forest regression voting. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part VII. LNCS, vol. 7578, pp. 278–291. Springer, Heidelberg (2012)
3. Cootes, T., Taylor, C., Cooper, D., Graham, J.: Active shape models-their training and application. Comput. Vis. Image Underst. **61**(1), 38–59 (1995)
4. Freund, Y., Schapire, R.: Experiments with a new boosting algorithm. In: Saitta, L. (ed.) ICML, pp. 148–156. Morgan Kaufmann, San Francisco (1996)
5. Montillo, A., Song, Q., Liu, X., Miller, J.: Parsing radiographs by integrating landmark set detection and multi-object active appearance models. SPIE Medical Imaging (2013)
6. Potesil, V., Kadir, T., Platsch, G., Brady, M.: Personalization of pictorial structures for anatomical landmark localization. In: Székely, G., Hahn, H.K. (eds.) IPMI 2011. LNCS, vol. 6801, pp. 333–345. Springer, Heidelberg (2011)
7. Seifert, S., Barbu, A., Zhou, S., Liu, D., Feulner, J., Huber, M., Suehling, M., Cavallaro, A., Comaniciu, D.: Hierarchical parsing and semantic navigation of full body CT data. SPIE Med. Imaging **7259**, 02:1–8 (2009)
8. Tao, Y., Peng, Z., Krishnan, A., Zhou, X.: Robust learning-based parsing and annotation of medical radiographs. IEEE Trans. Med. Imaging **30**(2), 338–350 (2011)
9. Viola, P., Jones, M.: Robust real-time face detection. IJCV **57**(2), 137–154 (2004)