DISCRIMINATIVE LEARNING OF DEFORMABLE CONTOUR MODELS

Haithem Boussaid, Iasonas Kokkinos, Nikos Paragios

Center for Visual Computing, Ecole Centrale de Paris, France
Galen Team, INRIA Saclay, France

ABSTRACT

In this work we propose a machine learning approach to improve shape detection accuracy in medical images with deformable contour models (DCMs). Our DCMs can efficiently recover globally optimal solutions that take into account constraints on shape and appearance in the model fitting criterion; our model can also deal with global scale variations by operating in a multi-scale pyramid.

Our main contribution consists in formulating the task of learning the DCM score function as a large-margin structured prediction problem. Our algorithm trains DCMs in an joint manner - all the parameters are learned simultaneously, while we use rich local features for landmark localization.

We evaluate our method on lung field, heart, and clavicle segmentation tasks using 247 standard posterior-anterior (PA) chest radiographs from the Segmentation in Chest Radiographs (SCR) benchmark. Our learned DCMs systematically outperform the state of the art methods according to a host of validation measures including the overlap coefficient, mean contour distance and pixel error rate.

1. INTRODUCTION

Precisely localizing shapes in medical images is of paramount importance in a host of medical image applications, involving organ segmentation, tracking, registration and atlas building. In our work we consider the problem of localizing a set of landmarks that are strung together along a contour; our task is to recover this landmark sequence by exploiting both the local appearance information around the individual landmarks, as well as their ordering constraints.

Deformable contour models (DCMs) constitute a main workhorse for detecting such 1D structures in medical images - starting from the seminal works of Snakes [1], Deformable Templates [2] and Active Shape/Appearance Models [3, 4], DCMs have been thriving in problems involving shapes for more than two decades. One of the most desirable properties of DCMs is that they allow to cast tasks such as segmentation or tracking in terms of optimization by incorporating the desirable properties of the envisioned solution in the form of a merit function. One can then optimize this function with off-the-shelf techniques, such as Dynamic Programming (DP) [5], Gradient Descent [4], or more dedicated techniques such as curve evolution with Level Sets [6].

We focus on learning the merit function being optimized so as to improve the shape localization performance of DCMs. Earlier research has involved enhancing the geometric terms in DCMs, including their formulation in intrinsic geometric terms [7], the incorporation of better regularization terms [8] and the introduction of shape priors [9] in curve evolution. Current works in medical imaging estimate the model parameters in a two-stage training manner, using e.g. maximum likelihood (ML) estimation for the pairwise terms,
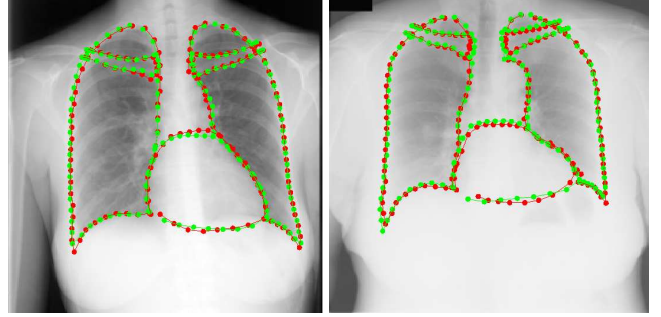
Fig. 1. Segmentation results on lungs, heart and clavicles. Ground truth contours are shown in green, our results are shown in red. The detection is accurate across variations and the scale change among cases.

and potentially other combinations of boosting/eigenspaces for the landmark appearance models ([10]/[11] respectively). In our experience, the resulting unary and pairwise terms can often be uncommensurate, and hand-tuning the relative contribution of the resulting terms may be needed. The minimal intensity and shape cost path (MISCP) algorithm [12] uses exhaustive leave-one out experiments to calibrate unary and pairwise terms.

Our work advances the state of the art in DCM-based medical image analysis primarily along the following lines:

First, **we learn the merit function of DCMs in an joint manner**; we express it as the inner product of a weight vector with appropriately formed appearance and geometric features (Sec. 2) and then estimate the optimal weight vector by casting the training problem as structured prediction (Sec. 4). This comes with the guarantees of optimality for the task at hand, in the sense that we can directly exploit the structure of the space of solutions. Our joint training tackles the estimation of all model parameters in terms of a single objective that directly reflects the performance in the task being solved. In particular, we give as input to our training algorithm, as appearance feature, dense (i.e. computable at every point) image descriptors as opposed to convolution features, which further boost shape detection performance.

Second, **we use a model that efficiently recovers a globally optimal configuration** without requiring initialization (as e.g. ASM/AAM-based models do [3]) or a shortlist of candidate landmark locations (as e.g. [13, 11] do). We use a chain-structured contour representation, hence DP [5] can recover the global optimum efficiently - and for the particular form of our geometric terms, the solution is obtained in a few seconds with Generalized Distance Transforms (GDTs) [14, 15]. Furthermore we accommodate scale changes through a multi-scale detection algorithm that uses an image pyramid.

In order to assess the merit of our contributions, we consider the task of segmenting the lung fields, the heart, and the clavicles in posterior-anterior (PA) chest radiographs. An accurate segmentation of these organs is essential for computer-aided diagnosis/analysis tasks, such as measuring the cardiothoratic ratio (CTR) for cardiomegaly diagnosis [16].We report results on the Segmentation in Chest Radiographs (SCR) benchmark [16, 17] which are largely superior to the current state-of-the-art MISCP algorithm [12] in terms of overlapping coefficient, mean contour distance and pixel error rate evaluation measures. Our implementation will be soon made available from www.centrale-ponts.fr/personnel/boussaid/code.

2. MERIT FUNCTION FORMULATION

We represent a 2D shape as an open contour consisting of a sequence of K anatomical landmarks: $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_K\}$, where every landmark is a 2D position vector $\mathbf{x}_i = (h_i, v_i)$. Given an image I we score a contour $\mathbf{X}$ with a merit function s formed as the sum of unary and pairwise terms:

$$s_I(\mathbf{X}) = \sum_{i=1}^{K} \mathcal{U}_{I,i}(\mathbf{x}_i) + \sum_{i=1}^{K-1} \mathcal{V}_i(\mathbf{x}_i, \mathbf{x}_{i+1}). \quad (1)$$

The unary terms capture local appearance information around $\mathbf{x}_i$ and the pairwise terms enforce geometric consistency among consecutive landmarks $(\mathbf{x}_i, \mathbf{x}_{i+1})$. More precisely, the unary terms capture the local fidelity of the image observations at $\mathbf{x}_i$ to a landmark-specific appearance model $\mathcal{U}_i$, in terms of an inner product between a weight vector $\mathbf{u}_i$ and image features extracted around $\mathbf{x}_i$: $\mathcal{U}_{I,i}(\mathbf{x}_i) = \langle \mathbf{u}_i, \mathbf{f}_I(\mathbf{x}_i) \rangle$. We denote by $\mathbf{f}_I(\mathbf{x}_i) : R^2 \to R^D$ a 'dense' mapping from any image coordinates to a D-dimensional feature; as detailed in Sec. 5, we experiment with state-of-the-art dense features, including dense SIFT [18], and a multi-scale convolution baseline. The elements of the weight vector $\mathbf{u}_i$ are learned discriminatively as detailed in Sec. 4.

The pairwise term constrains the location $\mathbf{x}_{i+1}$ with respect to its predecessor $\mathbf{x}_i$ in terms of a quadratic expression of the form: $\mathcal{V}_i(\mathbf{x}_i, \mathbf{x}_{i+1}) = -(\mathbf{x}_{i+1} - \mathbf{x}_i - \mu_i)^T C_i (\mathbf{x}_{i+1} - \mathbf{x}_i - \mu_i)$, where $C_i = \text{diag}(\nu_i, \eta_i)$ is a diagonal concentration matrix and μ_i is the nominal displacement between $\mathbf{x}_i$ and $\mathbf{x}_{i+1}$. We set μ_i equal to the mean displacement of a procrustes-aligned training set, but estimate ν_i, η_i discriminatively as in Sec. 4. The diagonal form of C_i allows us to write the pairwise term as a function separable in h and v:

$$\mathcal{V}_i(\mathbf{x}_i, \mathbf{x}_{i+1}) = -(h_{i+1} - h_i - \bar{h})^2 \nu_i - (v_{i+1} - v_i - \bar{v})^2 \eta_i \quad (2)$$
$$= \langle \mathbf{v}_i, \mathbf{p}(\mathbf{x}_i, \mathbf{x}_{i+1}) \rangle, \quad \text{where } \mathbf{v}_i = (\nu_i, \eta_i), \quad (3)$$
$$\mathbf{p}(\mathbf{x}_i, \mathbf{x}_{i+1}) = (-(h_{i+1} - h_i - \bar{h})^2, -(v_{i+1} - v_i - \bar{v})^2) \quad (4)$$

Having written the pairwise terms as the inner product between a weight and a feature vector, and given that the unary terms are also inner products between weights and features, it follows that Eq. 1 is the inner product between two vectors:

$$s_I(\mathbf{X}) = \langle \mathbf{w}, \mathbf{h}_I(\mathbf{X}) \rangle \quad \text{where}$$
$$\mathbf{w} = (\mathbf{u}_1, \ldots, \mathbf{u}_K, \mathbf{v}_1, \ldots, \mathbf{v}_{K-1}),$$
$$\mathbf{h}_I(\mathbf{X}) = (\mathbf{f}_I(\mathbf{x}_1), \ldots, \mathbf{f}_I(\mathbf{x}_K), \mathbf{p}(\mathbf{x}_1, \mathbf{x}_2), \ldots, \mathbf{p}(\mathbf{x}_{K-1}, \mathbf{x}_K))$$

We will henceforth denote $s_I(\mathbf{X})$ as $s_I(\mathbf{X}, \mathbf{w})$ to explicate its dependence on $\mathbf{w}$.

3. EFFICIENT GLOBAL OPTIMIZATION WITH GDTS

Given an image, the optimal configuration is found as $\mathbf{X}_I^* = \arg\max_{\mathbf{X}} s_I(\mathbf{X}, \mathbf{w})$. This optimization runs over a huge number

of possible contour configurations - for an image with N pixels N^K configurations are possible. But by virtue of the chain-structured form of s (Eq. 1) we can recover the optimal configuration efficiently with $O(KN^2)$ operations through Dynamic Programming as in [12]. However for the separable quadratic form of the pairwise terms in Eq. 2 we can employ GDTs [14, 15] to obtain the $\mathbf{X}^*$ with $O(KN)$ operations. We accommodate global scale changes through multi-scale optimization; from our original image I we construct an image pyramid by resampling at a set of scales $\mathcal{S} = (1, r, r^2, \ldots, r^S)$ and compute:

$$\mathbf{X}_{I,\mathcal{S}}^* = \arg\max_{\mathbf{X},i} s_{I(r^i)}(\mathbf{X}, \mathbf{w}), \quad (5)$$

where $I(r^i)$ denotes the image resampled with a ratio r^i. For notational convenience we will drop the $\mathcal{S}$ subscript from now on and it will be implied that the result is obtained through a multi-scale optimization.

4. LEARNING THE DCM MERIT FUNCTION

Having laid out the merit function driving the optimization of DCMs we now turn to learning its parameters. We assume that we have a training set of images and associated ground-truth contour locations, which we will denote as $\mathcal{X} = \{(I_i, \hat{\mathbf{X}}_i)\}, i = 1 \ldots N$. Our goal is to use this training set to learn a merit function such that on new, unseen, images the optimal contour configuration will be close to the respective ground truth configuration.

More specifically, we can see the optimization problem in Eq. 5 as defining a mapping $\Gamma : \mathcal{I} \to \mathcal{X}$ between the space of images and the space of contours: $\Gamma_{\mathbf{w}}[I] = \mathbf{X}_I^*$. Namely, given a parameter vector $\mathbf{w}$ we have a mapping ('operator') that takes an image as input and outputs a contour. Our goal is to estimate $\mathbf{w}$ so that this mapping will deliver contours close to the desired ones.

We consider that we have a loss function $\Delta(\mathbf{X}, \hat{\mathbf{X}})$ which indicates the discrepancy between two elements of the output space; for simplicity, we use the 0-1 loss, which is zero only when $\mathbf{X}$ and $\hat{\mathbf{X}}$ are identical and one otherwise.

To learn our merit function, we use *structured prediction* [19, 20] which amounts to requiring that for a training image I_i any configuration $\mathbf{X}$ other than the ground truth $\hat{\mathbf{X}}_i$ should score below $\hat{\mathbf{X}}_i$ by a certain margin; we can write this requirement concisely as:

$$s_{I_i}(\hat{\mathbf{X}}_i, \mathbf{w}) + \xi_i \geq s_{I_i}(\mathbf{X}, \mathbf{w}) + \Delta(\mathbf{X}, \hat{\mathbf{X}}_i), \forall \mathbf{X}, \xi_i \geq 0 \quad (6)$$

where we introduced a slack variable ξ_i associated with the i-th training example to relax this set of *constraints*. To avoid this relaxation in cases where it is unnecessary, we penalize the sum of slack variables through our training criterion. In particular we cast training our merit function as the optimization of the following quadratic program (QP):

$$\text{minimize} \quad C(\mathbf{w}, \xi) = \|\mathbf{w}\|_2^2 + \lambda \sum_{i=1}^{N} \xi_i \quad (7)$$

$$\text{subject to } Eq. 6$$

The first term in the training criterion regularizes the solution and guarantees good generalization, while the second term penalizes the amount by which the constraints are relaxed. The λ parameter controls the relative importance of each of the terms. We assess λ with cross validation. The cost function is quadratic in $\mathbf{w}$ and linear in ξ, while the set of constraints is linear in $\mathbf{w}$, since $s_{I_i}(\mathbf{X}, \mathbf{w})$, $s_{I_i}(\mathbf{X}, \hat{\mathbf{X}}_i)$ are linear in $\mathbf{w}$; as such, a single global optimum exists and can be found in principle with any QP solver. In particular, we use cutting plane optimization [20]. This amounts to solving the QP iteratively by appending at each iteration the most violated constraint. to the set of constraints already entertained.
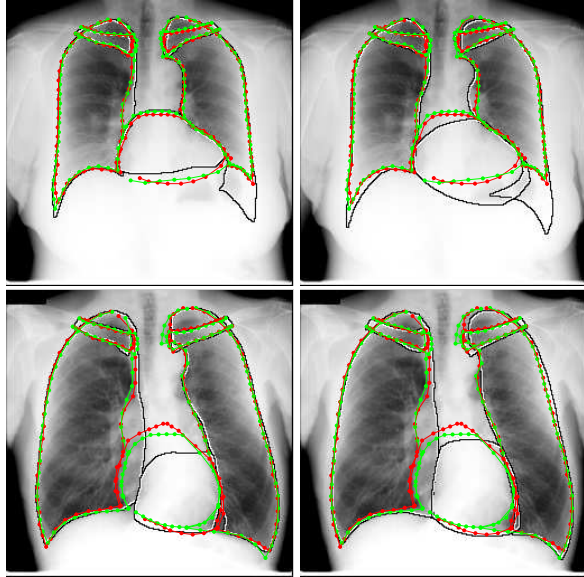
Fig. 2. Left: Our segmentations (red) superimposed on the results of the MISCP algorithm [12] (black contours). Right: Our segmentations (red) superimposed on the results of ASM based method of [17] (black contours). The ground truth segmentations are shown in green. Each row represents the same patient chest.

Table 3. Pixel error results on the SCR database [16, 17]. Proposed method scores better than the state-of-the art approaches.

method	pixel error
Our	**0.022±0.006**
MISCP [12]	0.033±0.017
ASM tuned [17]	0.044±0.014

5. EXPERIMENTAL SETTINGS AND RESULTS

We systematically evaluate our method on the publicly available **dataset** of [16, 17] which contains 247 standard posterior anterior chest radiographs of healthy and non-healthy subjects (presenting nodules). The database contains segmentations from radiologists, which provide a delineation of the lung fields, the heart and the clavicles. Gold standard segmentation masks are hence available as well as corresponding landmark positions lying on the contour. In all of our experiments we use the standard evaluation setup described in [16], using 123 images for training and a separate set of 124 images for testing; the reported results are on the test set. We test our algorithm on rescaled images of size 256 X 256.

We **evaluate** the performance of our method by comparing our segmentations to the ground truth segmentations by means of a range of validation measures; these include Jaccard, Dice coefficients, mean contour distance (MCD) and pixel error measures. The Jaccard coefficient between two objects is defined as the area of their intersection divided by the area of their union. The Dice coefficient between two objects is defined as two times the area of their intersection divided by the sum of their areas. The Dice coefficient is an equivalent measure to Jaccard index. We include it for completeness. These overlap coefficients are high for large, simple objects and low for complex, small objects. The MCD between two contours A and B is defined in terms of the average distance from a point on the contour A to the nearest point on the contour B. It is

obtained by averaging the distance A B to the distance B A. All the reported MCD measures are in millimeter. The pixel error index is defined as the proportion of pixels for which any of the five object labels (lungs, heart, clavicles) is not in agreement with the ground truth segmentation.

To validate the merit of our contributions, we assess the performance of our work against existing baselines

Training: two-stage parameter estimation vs. joint training: We compare the performance of our DCMs in the following two settings: (a) using, as in [12], parameters estimated through Maximum Likelihood estimation for the pairwise terms (means and standard deviations) and classifiers trained separately per every landmark and (b) using parameters jointly trained through our joint training objective.

Appearance features: Existing works in medical imaging either use potentially rich, but sparse features [13, 11] -which requires recovering from detector failures- or dense, but less discriminative features using convolution with filterbanks [10]. We require that the image features used to construct the unary terms should be (a) dense to ensure we are not ruling out potential landmark locations and (b) informative and discriminative, so that they can potentially distinguish among different landmarks. To meet these two requirements we use dense SIFT descriptors [18]. Our results indicate that our dense descriptors yield a systematic boost in performance, when compared to simpler baselines. In our comparisons we use as a baseline the responses of multi-scale steerable filterbank, implemented along the lines of [21].

Quantitative evaluation of different design choices: In Table 1 we report validation measures for the different design choices considered. Our very first observation is that our simplest baseline (convolution filters with two-stage learning) reaches similar performance to the state-of-the-art MISCP algorithm [12], which in its turn outperforms ASMs [17] as extensively demonstrated in [12]. We further verify that (a) joint training boosts performance when compared to two-stage training; and (b) dense appearance descriptors have a clear edge over standard convolution features in both joint training and two-stage training.

The results in Table 1 are complemented by the results in Table 2 where we provide validation measures for the heart and clavicles segmentation results and ASMs [17] on the same dataset (results of [12] on the heart and clavicles segmentation tasks were not reported in [12]). Furthermore, Table 3 shows that our method has the lowest mean pixel error (0.022) compared to all the available methods evaluated on the SCR database. We refer to the SCR website [22] for a detailed overview of the performance of all available methods described in [17]. Finally, an extensive side-by-side comparison of our proposed method and the state-of-the-art MISCP algorithm [12] as well as the ASM based method [17] is provided in Fig 2, which qualitatively demonstrates the higher accuracy attained by our model on challenging areas with poor low-level information and the substantial improvement over the state-of-the art MISCP [12] and ASMs [17]. The segmentations of [12, 17] were retrieved from the SCR benchmark website [22].

6. CONCLUSION

In this work we have introduced a discriminative method for training DCMs; by virtue of being chain-structured our DCMs allow for globally optimal solutions with limited computational demands (less than 8 seconds on a standard PC), while our joint training, clearly outperform the current state-of-the-art for the task of finding anatomical object shapes in chest radiographs. The use of rich appearance features improves further the results.

Table 1. Lung segmentation Performance measures including Dice and Jaccard coefficients (larger is better) and means contour distance (smaller is better). We compare the performance of the previous state-of-the-art, MISCP [12], ASM [17], and different choices for our method, involving dense SIFT at a resolution of 4 pixels per bin and a convolution baseline (CONV) with steerable-scalable filters. The suffix TS indicates two-stage and JT indicates joint training. The proposed method corresponds to the use of SIFT descriptors and joint training

| method	Right Lung (44 landmarks)			Left Lung (50 landmarks)		
	Dice	Jaccard	M.C.D	Dice	Jaccard	M.C.D
Proposed method	**97.85±1**	**95.8±3**	**1.82±0.3**	**97.52±1**	**95.2±5**	**1.96±0.4**
SIFT+TS	97.17±2	94.6±2	1.96±0.7	96.65±1	93.6±5	2.52±1.1
CONV+JT	97.26±2	94.5±3	1.82±0.7	96.8±1	93.8±5	2.66±1.3
CONV+TS	96.84±1	93.9±1	1.96±0.7	95.81±1	92.0±5	2.38±0.4
MISCP [12]	N/A	94.0±3	2.1±0.8	N/A	92.0±4	2.38±1.4
ASM tuned [17]	N/A	92.1±3	2.66±1	N/A	88.6±7	3.78±3.5

Table 2. Heart and clavicles segmentation Performance measures including Dice and Jaccard coefficients (larger is better) and means contour distance (smaller is better). We compare the performance of the previous state-of-the-art, ASM [17], and our best performing method involving dense SIFT at a resolution of 4 pixels per bin and joint training.

method	Heart (26 landmarks)			Clavicles (23 landmarks)		
	Dice	Jaccard	M.C.D	Dice	Jaccard	M.C.D
Proposed method	**94.8±2**	**90.1±1**	**3.03±1.3**	**90.4±5**	**80.1±9**	**1.4±0.7**
ASM tuned [17]	N/A	81.4±8	5.96±2.73	N/A	73.4 ±1	2.04 ±1.36

In future work we intend to further pursue this research direction for tasks involving more complex energy functions, and 3D data, as well as to incorporate contour closedness in the problem constraints, which cannot be encoded through chain structured graphs. This can be done efficiently with dual decomposition techniques [23].

7. REFERENCES

[1] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active Contour Models," in *ICCV*, 1987.

[2] A. L. Yuille, P. W. Hallinan, and D. S. Cohen, "Feature Extraction from Faces Using Deformable Templates," *IJCV*, 1992.

[3] T.F. Cootes, G. J. Edwards, and C.J. Taylor, "Active Appearance Models," in *ECCV*, 1998.

[4] Timothy F. Cootes, Christopher J. Taylor, David H. Cooper, and Jim Graham, "Active shape models-their training and application," *CVIU*, 1995.

[5] Davi Geiger, Alok Gupta, Luiz A. Costa, and John Vlontzos, "Dynamic programming for detecting, tracking, and matching deformable contours," *PAMI*, 1995.

[6] Ravi Malladi, James A. Sethian, and Baba C. Vemuri, "Shape modeling with front propagation: A level set approach," *PAMI*, 1995.

[7] V. Caselles, R. Kimmel, and G. Sapiro, "Geodesic Active Contours," *IJCV*, 1997.

[8] Benjamin B. Kimia, Ilana Frankel, and Ana-Maria Popescu, "Euler spiral for shape completion," *IJCV*, 2003.

[9] M. Rousson and N. Paragios, "Shape Priors for Level Set Representations," in *ECCV*, 2002.

[10] Vaclav Potesil, Timor Kadir, Gunther Platsch, and Mike Brady, "Improved anatomical landmark localization in medical images using dense matching of graphical models," in *BMVC*, 2010.

[11] Ahmed Besbes and Nikos Paragios, "Landmark-based segmentation of lungs while handling partial correspondences using sparse graph-based priors," in *ISBI*, 2011.

[12] Dieter Seghers, Dirk Loeckx, Frederik Maes, Dirk Vandermeulen, and Paul Suetens, "Minimal shape and intensity cost path segmentation," *IEEE Trans. Med. Imaging*, 2007.

[13] Stefan Schmidt, Jorg H. Kappes, Martin Bergtholdt, Vladimir Pekar, Sebastian P. M. Dries, Daniel Bystrov, and Christoph Schnorr, "Spine detection and labeling using a parts-based graphical model," in *IPMI*, 2007.

[14] Pedro F. Felzenszwalb and Daniel P. Huttenlocher, "Distance transforms of sampled functions," Tech. Rep., Cornell CS, 2004.

[15] P. Felzenszwalb and D. Huttenlocher, "Pictorial Structures for Object Recognition," *IJCV*, 2005.

[16] J. Shiraishi, S. Katsuragawa, T. Matsumoto J. Ikezoe, T. Kobayashi, K. Komatsu, M. Matsui, H. Fujita, Y. Kodera, and K. Doi, "Development of a digital image database for chest radiographs with and without a lung nodule," *Am. J. Roentgenology*, 2000.

[17] B. van Ginneken, M.B. Stegmann, and M. Loog, "Segmentation of anatomical structures in chest radiographs using supervised methods: a comparative study on a public database," *Medical Image Analysis*, 2006.

[18] Brian Fulkerson, Andrea Vedaldi, and Stefano Soatto, "Localizing objects with smart dictionaries," in *ECCV*, 2008.

[19] Sebastian Nowozin and Christoph H. Lampert, "Structured learning and prediction in computer vision," *Foundations and Trends in Computer Graphics and Vision*, 2011.

[20] Thorsten Joachims, Thomas Finley, and Chun-Nam John Yu, "Cutting-plane training of structural svms," *Machine Learning*, 2009.

[21] William T. Freeman and Edward H. Adelson, "The design and use of steerable filters," *PAMI*, 1991.

[22] "SCR database," http://www.isi.uu.nl/Research/Databases/SCR /results/browser.php, Accessed: 2013-10-01.

[23] Nikos Komodakis, Nikos Paragios, and Georgios Tziritas, "Mrf optimization via dual decomposition: Message-passing revisited," in *ICCV*, 2007.
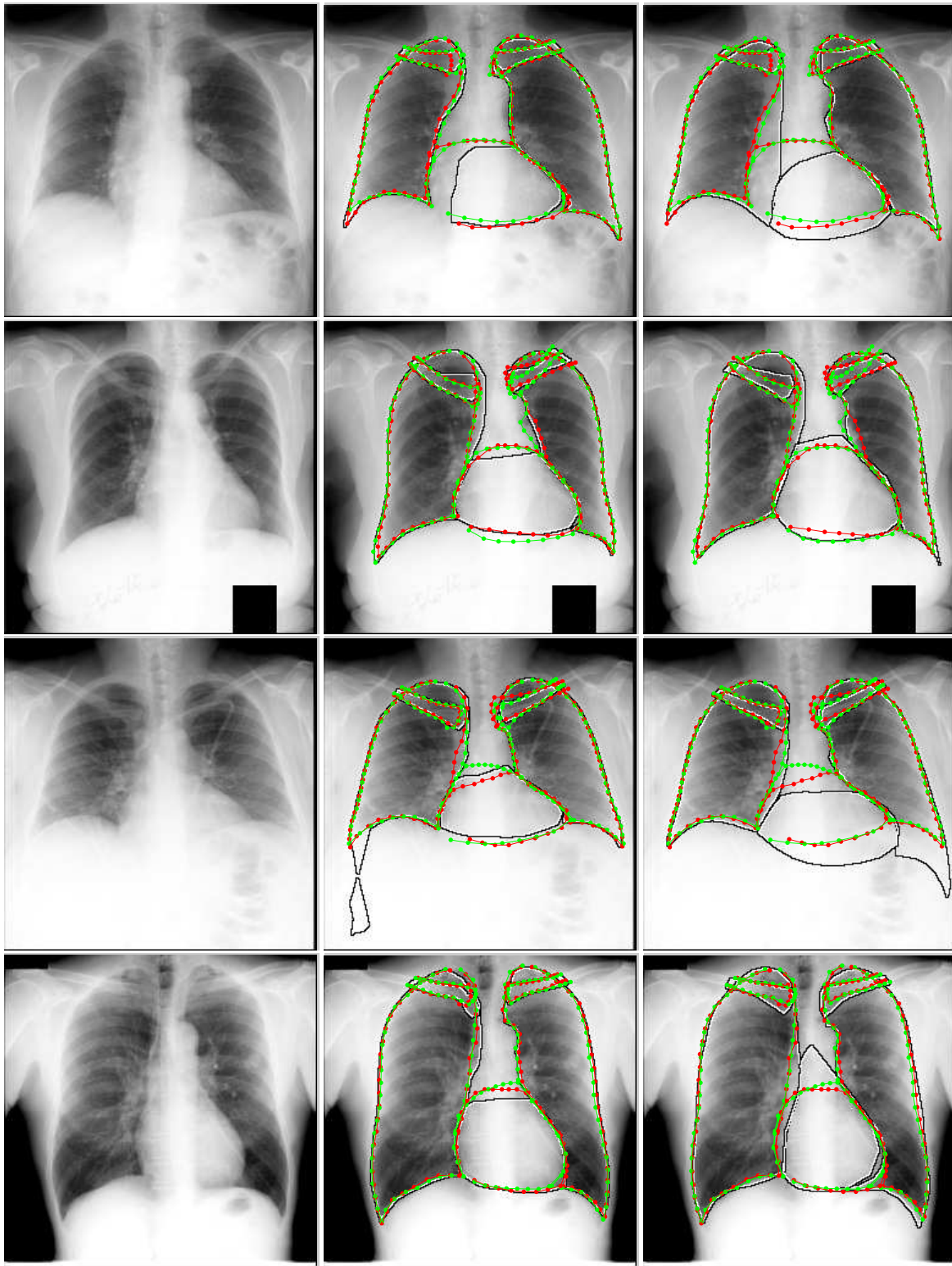
Fig. 3. Left: patient chest radiograph. Middle: our segmentations (red) superimposed on the results of the MISCP algorithm [12] (black contours). Right: Our segmentations (red) superimposed on the results of ASM based method of [17] (black contours). The ground truth segmentations are shown in green. Each row represents the same patient radiograph.