Published in final edited form as:

Proc SPIE Int Soc Opt Eng. 2017 February 11; 10133: . doi:10.1117/12.2254412.

Marginal Shape Deep Learning: Applications to Pediatric Lung Field Segmentation

Awais Mansoor^a, Juan J. Cerrolaza^a, Geovanny Perez^b, Elijah Biggs^a, Gustavo Nino^b, and Marius George Linguraru^{a,c}

^aSheikh Zayed Institute for Pediatric Surgical Innovation, Childrens National Medical Center, Washington DC

^bDivision of Pulmonary and Sleep Medicine, Childrens National Health System, Washington, DC

^cSchool of Medicine and Health Sciences, George Washington University, Washington DC

Abstract

Representation learning through deep learning (DL) architecture has shown tremendous potential for identification, localization, and texture classification in various medical imaging modalities. However, DL applications to segmentation of objects especially to deformable objects are rather limited and mostly restricted to pixel classification. In this work, we propose marginal shape deep learning (MaShDL), a framework that extends the application of DL to deformable shape segmentation by using deep classifiers to estimate the shape parameters. MaShDL combines the strength of statistical shape models with the automated feature learning architecture of DL. Unlike the iterative shape parameters estimation approach of classical shape models that often leads to a local minima, the proposed framework is robust to local minima optimization and illumination changes. Furthermore, since the direct application of DL framework to a multi-parameter estimation problem results in a very high complexity, our framework provides an excellent run-time performance solution by independently learning shape parameter classifiers in marginal eigenspaces in the decreasing order of variation. We evaluated MaShDL for segmenting the lung field from 314 normal and abnormal pediatric chest radiographs and obtained a mean Dice similarity coefficient of 0.927 using only the four highest modes of variation (compared to 0.888 with classical ASM¹ (p-value=0.01) using same configuration). To the best of our knowledge this is the first demonstration of using DL framework for parametrized shape learning for the delineation of deformable objects.

Keywords

deep learning; shape learning; statistical shape models; lung field; chest radiograph

1. INTRODUCTION

The use of image feature representation has been successfully evaluated for the applications of classification and detection;²⁻⁴ however, its use for object segmentation has been rather

limited. The primary reason may be that the architecture of representation learning methods favors solving classification problems. Although detection can easily be formulated as a classification hypothesis (absence/presence of object), the effective representation of a recognition task is not trivial and relies heavily on human ingenuity.^{2,3} Hand crafting learning features may not be straightforward in cases such as recognition of deformable objects that involves accurate shape modeling and deformation estimation. Recently, representation learning through deep learning (DL) has shown promising results in expanding the scope of machine learning algorithms to automated feature-crafting. DL framework relies on a hierarchical representation model that learn features from low-level image data and successively builds more comprehensive features in a hierarchical layer-by-layer manner. Specific to the medical imaging domain, DL frameworks are being used primarily for organ detection⁵ and classification tasks.⁴ The DL-based segmentation methods proposed in the literature are focused mostly on classification of pixels. Pixel classification-based methods although have shown promising potential in texture analysis and as a refinement step to initial segmentation, the cost of performing hypothesis testing at pixel level prohibits their use over more developed region and model-based segmentation methods. Moreover, since pixel classification-based segmentation is generally obtained as a concatenation of independent hypothesis, object shape specificity cannot be guaranteed. To our best knowledge, DL for object segmentation, outside of pixel classification, has not been attempted in the literature.

In this work, we extend the applicability of the DL framework to estimating shape deformation for model-based segmentation (detection and recognition). Over years, statistical shape models (SSM) have established themselves as a robust mechanism for modeling deformable objects. SSM learn patterns of shape deformation (known as *modes*) from a training data of annotated images. The learned model is subsequently deformed to fit a test image by estimating its shape deformation modes that are consistent with the training data through an iterative optimization. The optimization is however not robust to initialization, local minima, and illumination intensities; henceforth, accurate shape initialization⁶ and various refinements⁷ remain topics of active research to date. To address these challenges, we utilized the DL architecture to estimate the modes of shape deformation. Unlike the iterative deformation approach of classical SSM, learning-based methods tend to have a greater neighborhood awareness beyond pixel level which makes them less susceptible to local minima optimization. Furthermore, in order to create an efficient run-time solution, our framework extends the concept of marginal space learning^{2,3} to shape learning by marginally learning the parameters of shape deformation in the eigenspace of the largest variation and then gradually increasing the dimensionality of eigenspaces by including the next largest variation and so on. We evaluate the proposed marginal shape deep learning (MaShDL) framework for the automatic segmentation of the lung field from pediatric posterior-anterior chest radiographs (CXRs).

2. METHODS

The MaShDL framework presented in this manuscript is aimed at automatically segmenting deformable objects using DL. The framework is generically applicable to 2D and 3D images; therefore, without the loss of generality, in this manuscript, we demonstrate the

application of our framework to 2D pediatric CXR images. Fig. 1 shows the flow diagram describing the proposed MaShDL framework.

The success of SSM in the segmentation of deformable objects is owed to their capability to capture major shape deformation modes using very few parameters.¹ To build an SSM, N training shapes represented by L points with anatomical correspondence are generally used. In a simple scenario, the training shapes are aligned using the generalized Procrustes analysis to remove translation ($\mathbf{T}$), scaling ($\mathbf{S}$), and rotation ($\mathbf{R}$). The valid shape space traversed by N aligned shapes can be represented as a linear combination of K most representative eigenvectors $v_1, v_2, \dots, v_K$ based on principal component analysis (PCA). Using SSM, a new shape Y in the aligned shape space can be represented as

$Y = \overline{M} + \sum_{k=1}^K c_k v_k + e$, where $\overline{M}$ is the mean shape, c_k is the PCA coefficient of the k^{th} eigenvector v_k (also known as the k^{th} mode of shape deformation), and e is the residual error between the estimated landmark position and the true landmark position. Subsequently, a non-rigid deformation in n -dimensional image can be represented using the following shape parameter vector:

$$\Omega = \left\{ \underbrace{\mathbf{T} \in \mathbb{R}^n, \mathbf{R} \in \mathbb{R}^{n \times (n-1)/2}, \mathbf{S} \in \mathbb{R}^n}_{\text{Object Detection Parameters}}, \underbrace{\overline{M}}_{\text{Mean Shape}}, \underbrace{\mathbf{c} \in \mathbb{R}^K}_{\text{Object Recognition Parameters}} \right\}, \quad (1)$$

Using this parametric representation, the deformation estimation problem can be converted to a feature estimation problem. Specifically, given a shape parameter vector Ω , we can train a classifier that can distinguish positive hypothesis (correct parameters vector) from the negative hypothesis (incorrect parameter vector). Since $\overline{M}$ is calculated as the mean of aligned shapes and therefore constant; given an image f , the object segmentation problem can be reduced to estimating features that can maximize the following probability over the valid parameter range,

$$\hat{\Omega} = \left\{ \underbrace{\hat{\mathbf{T}}, \hat{\mathbf{R}}, \hat{\mathbf{S}}}_{\Omega_{\text{Detection}}}, \underbrace{\hat{\mathbf{c}}}_{\Omega_{\text{Recognition}}} \right\} = \underset{\Omega}{\operatorname{argmax}} p(\Omega | f). \quad (2)$$

Due to the large increase in testing hypotheses with the number of parameters (see eq. (1)), an exhaustive search of the target parameters using eq. (2) is impractical. Therefore, we introduced splitting the parameter space into marginal subspaces:

$$\begin{aligned} p(\Omega | f) &= p(\Omega_{\text{Object Detection}} | f) \cdot p(\Omega_{\text{Object Recognition}} | \Omega_{\text{Object Detection}}, f) \\ &= p(\mathbf{T} | f) \times p(\mathbf{R} | \mathbf{T}, f) \times p(\mathbf{S} | \mathbf{T}, \mathbf{R}, f) \times p(\mathbf{c} | \mathbf{S}, \mathbf{T}, \mathbf{R}, f), \end{aligned} \quad (3)$$

2.1 Stacked Autoencoder Hierarchical deep Learning Architecture

The stacked autoencoder (SAE), used in the proposed MaShDL framework, is a deep neural network consisting of multiple layered sparse autoencoders (AE) in which the output of every layer is wired to the input of successive layer. An AE consists of two components: the encoder and the decoder. The encoder fits a non-linear mapping that can project the higher dimensional observed data ($f(x_n) \in \mathbb{R}^n$) into its lower dimensional feature representation.

The decoder recovers the observed data ($\hat{f}(x_n)$) from the feature representation with minimal recovery error. For a detailed description of AE, SAE, and related DL concepts, the reader is encouraged to review.⁸

2.2 Object Detection

In our 2D example, the object detection problem was posed as a marginal approach to estimating three parameters that define the bounding box enclosing the object of interest: position $\mathbf{T} = \{t_x, t_y\}$, rotation $\mathbf{R} = \{r_\theta\}$, and scaling $\mathbf{S} = \{s_x, s_y\}$. DL classifiers are trained to evaluate the hypothesis of a bounding box centered at position $\mathbf{T}$ at orientation $\mathbf{R}$ and size $\mathbf{S}$ enclosing the object of interest. From eq.(2)

$$\hat{\Omega}_{\text{Object Detection}} = \{\hat{\mathbf{T}}, \hat{\mathbf{R}}, \hat{\mathbf{S}}\} = \underset{\Omega_{\text{Object Detection}}}{\operatorname{argmax}} p(\Omega_{\text{Object Detection}} | f). \quad (4)$$

Position Estimation—For position estimation, given a pixel position hypothesis $\mathbf{T} = (t_x, t_y)$, the classification problem is formulated as finding the object of interest centered at that position. Accordingly, a positive hypothesis is formulated as

$$\max(|t_x - \tilde{t}_x|, |t_y - \tilde{t}_y|) \leq 1 \text{ pixels}, \quad (5)$$

where $\tilde{t}_x$ and $\tilde{t}_y$ denote the centroid positions of the bounding-box enclosing the manually segmented ground truth along the x- and the y-directions respectively. A negative sample satisfies $\min(|t_1 - \tilde{t}_1|, |t_2 - \tilde{t}_2|) \geq 2 \text{ pixels}$. The gap between positive and negative thresholds is intended for a clean separation between the two classes during training. Moreover, since the number of positive hypotheses is smaller than the number of negative hypotheses, we adopted the data augmentation approach presented in⁹ to increase the number of training samples and prevent over-fitting along with random sampling to balance the samples prior to training. During the testing stage, each pixel in the image is tested for the object position using the trained classifier and a small number (100 in our experiments) of candidates having the highest score is retained.

Position-Orientation Estimation—The orientation classification problem is formulated as finding an object of interest with centroid at (t_x, t_y) and orientation θ . The orientation hypothesis is regarded as positive if, in addition to satisfying eq. (5), it also satisfies,

$$|\theta - \tilde{\theta}| \leq 0.1\text{rad}, \quad (6)$$

where $\tilde{\theta}$ denotes the orientation of the bounding box encapsulating the ground-truth label. A negative sample satisfies $|\theta - \tilde{\theta}| \geq 0.2\text{rad}$. During the testing stage, the retained samples from the position estimator are rotated with a step-size of 0.017 radians. Subsequently, trained SAE classifiers⁸ are used to calculate the similarity scores of the joint position and rotation estimation and top (100) candidates are retained.

Position-Orientation-Scale Estimation—To estimate anisotropic scale along x-dimension, a hypothesis (t_x, t_y, θ, S_x) is considered positive if, in addition, to satisfying eq. (5) and (6) it also satisfies,

$$|S_x - \hat{S}_x| \leq 1 \text{ pixels}, \quad (7)$$

where $\hat{S}_x$ is the x -dimension length of the bounding box encompassing the ground truth label. A negative hypothesis will satisfy eq. (5) and (6) as well as $|S_x - \hat{S}_x| \geq 2$ (pixels). The estimation of scale S_y along y -dimension is processed similarly for the retained candidates with highest score.

It is important to mention here that, in our experiments, we did not find any significant difference from the relative order in which orientation and anisotropic scale were estimated.

2.3 Object Recognition

Once the object of interest is detected, the recognition was performed by starting with a flexible template mean shape $(\overline{M})$. Subsequently, we estimate shape deformation modes ($\mathbf{c}$), deforming $\overline{M}$ to match the test shape, using a set of trained SAE DL classifiers. A recent application used hand-crafted features with a sampling pattern for shape estimation;² however, crafting an accurate sampling pattern and extracting hand-crafted features for shape hypothesis is non-trivial. In MaShDL, the representation learning through DL architecture automatically identifies, extracts, and learns distinguishable features directly from the low-level image data.

For object recognition, each hypothesis corresponds to a synthesized shape with specific shape deformation modes. In testing the hypothesis, we propose to estimate the coefficients of the K largest eigenvectors that maximize the posterior probability

$$\hat{\Omega}_{\text{Object Recognition}} = \underset{\mathbf{c}}{\text{argmax}} p(\mathbf{c} | \mathbf{S}, \mathbf{T}, \mathbf{R}, f), \quad (8)$$

where $\mathbf{c} = \{c_k\}_{k \in K}$ is the set of coefficients of the K largest eigenvectors in descending order. Again, the direct application of DL to estimating multiple shape parameters together would result in increased complexity. Therefore, we split the eigenspace into marginal shape

subspaces in the decreasing order of shape variation and begin by learning a DL classifier only for the mode representing the largest shape variation. Subsequently, we address the mode representing the second largest variation and so on. Similarly to detection, the marginal recognition approach can be expressed as the factorization of posterior probabilities:

$$p(\mathbf{c}|\mathbf{S}, \mathbf{T}, \mathbf{R}, f) = p(c_1|\mathbf{S}, \mathbf{T}, \mathbf{R}, f) \times \cdots \times p(c_K|c_{K-1}, \dots, c_1, \mathbf{S}, \mathbf{T}, \mathbf{R}, f), \quad (9)$$

However, unlike in object detection (eq. (4)), the posterior probability of the marginal shape deformation parameters is not cumulative, i.e., the parameters of the eigenvectors must be estimated in a specific order, namely the descending order of variation. To train a DL classifier for the parameter estimation of the k^{th} largest mode of variation, we generate a set of negative and positive hypothesis by (i) creating a set of synthesized shapes by varying only the k^{th} largest mode of variation and, (ii) extracting patches around the L points representing the synthesized shape. For any two deformation hypotheses to be resolvable using image-based feature representation, the minimum landmark-to-landmark distance between them should be at least 1 pixel. In our experiments, using this criterion, the minimum resolvable shape deformation was found to be 0.15. Subsequently, a positive hypothesis should satisfy eq. (4), as well as:

$$|c_k - \tilde{c}_k| \leq 0.15, \quad (10)$$

where $\tilde{c}_k$ is the ground truth parameter. In order to synthesize only the valid shapes for hypothesis, k assumes values within the valid space $c_k \leq |3\sigma_k|$ ¹ where σ_k is the square-root of the k^{th} eigenvalue. A negative hypothesis should satisfy eq. (4) and

$$|c_k - \tilde{c}_k| \geq 0.75, \forall c_k \in \{-3.0, \dots, +3.0\} \sigma_k, \quad (11)$$

This translates to a minimum landmark-to-landmark distance of 5 pixels. Fig. 2 shows examples of positive hypotheses (shown in green) generated using eq. (10) and negative hypotheses (shown in red) for the four highest modes of variation. A step-size of 0.02 (corresponds to maximum landmark-to-landmark distance of 1 pixel) was used within the valid space to generate positive and negative hypothesis. The extracted image patches were subsequently stacked together to train the SAE classifier for each deformation mode. To increase the robustness to outliers, the average value of the top candidates (100) are chosen as the estimated value for the deformation mode. The process was repeated sequentially by estimating the next highest variation mode in a similar fashion until a cumulative energy of 90%, which is equivalent to using the first four modes of variation (Fig. 2) or until the total number of modes are included.

3. EXPERIMENTS AND RESULTS

To demonstrate the performance of MaShDL framework, we chose the task of segmenting the lung field from pediatric CXRs. Lung field is a non-rigid structure and SSM have been used numerous in the literature for its segmentation. Indeed, due to the high mesenchymal components needed for growth and development, determining the shape of pediatric organs is a major challenge for deep learning approaches. After Internal Review Board approval, we collected retrospective CXRs from 314 subjects at our institution. The images have spatial resolution of $0.1 \text{ mm} \times 0.1 \text{ mm}$ and digital resolution of 12-bits. The intensity of the images was normalized using the method described in¹⁰ and the ground truth labels for lung field were annotated by experts. The dataset was split randomly at patient level to 167 training and 167 testing sets. The parameters for SAE architecture were adjusted empirically using the guidelines described in.⁸ Three hidden layers were used with 400, 200, and 100, layers respectively. The batch size was set to 100 with 100 epochs. The size of the image patch around each landmark was empirically set to 11×11 pixels (sizes from 5×5 to 13×13 were tested). Identical configurations were used for training all SAE classifiers. An average error of 2.67 mm in translation, 0.19 rad in rotation, and 4.38 mm in scaling was obtained in the testing set using MaShDL. For the deformation parameter estimation, an average error of 0.30, 0.40, 0.23, and 0.33 was observed for the first four modes respectively. To quantify the lung field segmentation results, we calculated Dice similarity coefficient (DSC) and mean landmark distance (MLD) of point-to-ground-truth-contour. Fig. 3 shows comparative DSC and MLD values for first four modes (90% cumulative energy) for MaShDL and classical ASM.¹ The detection module presented in the manuscript (Section 2.2) was used to initialize both approaches. By effectively capturing the shape variation using DL features, MaShDL outperformed ASM for accuracy using fewer parameters. We obtained a significant improvement by MaShDL with a mean DSC value of 0.927 compared to 0.888 by classical ASM (Wilcoxon, $p=0.01$). Similarly, gain in mean MLD was observed using MaShDL (0.171 mm) compared to classical ASM (0.423 mm) using the first four modes. Using the DL features, MaShDL makes much efficient use of information compared to classical ASM: significantly sophisticated ASM is needed to match MaShDL performance (eight modes of variation in our experiments). MaShDL is also found to be computationally faster ($\approx 12.5\%$) than classical ASM. In our experiments, average time per scan to perform object recognition on a test image was 10.97 seconds using MaShDL compared to 12.52 seconds using classical ASM on the same hardware. Importantly, as demonstrated in Fig. 3, unlike ASM, the performance of MaShDL increased monotonically with the number of modes. ASM's non-monotonic performance was due primarily to local minima.

4. CONCLUSION

In this work, we introduced the marginal shape deep learning framework (MaShDL). MaShDL combines the strength of statistical shape models with the automated feature learning architecture of DL. Unlike the iterative shape parameters estimation approach of classical shape models that often leads to a local minima; the proposed framework is robust to parameter optimization and illumination changes. Our work offers a fast and robust approach to estimate the pose and shape parameters of deformable models. We exemplified MaShDL on the segmentation of 2D pediatric CXR images, but the framework is applicable

to the estimation and optimization of other general model-based segmentation. In future work, we plan on extending and evaluating MaShDL for the shape segmentation problems in 3D applications.

Acknowledgments

This project was funded by NIH grants UL1TR000075/KL2TR000076 and a gift from the Government of Abu Dhabi to the Children's National Health System.

References

1. Cootes TF, Taylor CJ, Cooper DH, Graham J. Active shape models-their training and application. *Computer Vision and Image Understanding*. 1995; 61(1):38–59.
2. Zheng Y, Barbu A, Georgescu B, Scheuering M, Comaniciu D. Four-chamber heart modeling and automatic segmentation for 3D cardiac CT volumes using marginal space learning and steerable features. *Medical Imaging, IEEE Transactions on*. 2008; 27(11):1668–1681.
3. Ghesu, FC., Georgescu, B., Zheng, Y., Hornegger, J., Comaniciu, D. *Medical Image Computing and Computer-Assisted Intervention–MICCAI*. Vol. 2015. Springer; 2015. Marginal space deep learning: Efficient architecture for detection in volumetric image data; p. 710-718.
4. Anthimopoulos M, Christodoulidis S, Ebner L, Christe A, Mougiakakou S. Lung pattern classification for interstitial lung diseases using a deep convolutional neural network. *IEEE Transactions on Medical Imaging*. 2016; PP(99):1–1.
5. Shin H-C, Orton MR, Collins DJ, Doran SJ, Leach MO. Stacked autoencoders for unsupervised feature learning and multiple organ detection in a pilot study using 4d patient data. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*. 2013; 35(8):1930–1943.
6. Cosfo FA. Automatic initialization of an active shape model of the prostate. *Medical Image Analysis*. 2008; 12(4):469–483. [PubMed: 18346931]
7. Zhang S, Zhan Y, Dewan M, Huang J, Metaxas DN, Zhou XS. Towards robust and effective shape modeling: Sparse shape composition. *Medical image analysis*. 2012; 16(1):265–277. [PubMed: 21963296]
8. Larochelle, H., Erhan, D., Courville, A., Bergstra, J., Bengio, Y. *Proceedings of the 24th International Conference on Machine Learning*. ACM; 2007. An empirical evaluation of deep architectures on problems with many factors of variation; p. 473-480.
9. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*. 2012:1097–1105.
10. Mansoor, A., Linguraru, MG. 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI). IEEE; 2016. Generic method for intensity standardization of medical images using multiscale curvelet representation; p. 1320-1323.

Figure 1.

The proposed marginal shape deep learning (MaShDL) framework for model-based segmentation using deep learning.

Figure 2.

Four highest modes of variation in training data are shown from left to right. For each mode, we show the superimposed synthesized shapes (step size=0.05): Positive hypotheses (eq. 10) for the mode is shown in green while the negative hypotheses (eq. 11) are shown in red.

Figure 3. Box-plots of (a) Dice similarity coefficient (DSC) and (b) mean landmark distance (MLD) for MaShDL and classical ASM.