Supervised Action Classifier: Approaching Landmark Detection as Image Partitioning

Zhoubing Xu[1(✉)], Qiangui Huang[2], JinHyeong Park[1], Mingqing Chen[1], Daguang Xu[1], Dong Yang[3], David Liu[1], and S. Kevin Zhou[1]

[1] Medical Imaging Technologies, Siemens Healthineers Technology Center,
Princeton, NJ 08540, USA
`zhoubing.xu@siemens.com`
[2] Department of Computer Science, University of Southern California,
California, LA 90089, USA
[3] Department of Computer Science, Rutgers University, Piscataway, NJ 08854, USA

Abstract. In medical imaging, landmarks have significant clinical and scientific importance. Clinical measurements, derived from the landmarks, are used for diagnosis, therapy planning and interventional guidance in many cases. Automatic algorithms have been studied to reduce the need for manual placement of landmarks. Traditional machine learning techniques provide reasonable results; however, they have limitation of either robustness or precision given complexities and variabilities of the medical images. Recently, deep learning technologies have been emerging to tackle the problems. Among them, a deep reinforcement learning approach (DRL) has shown to successfully detect landmark locations by implicitly learning the optimized path from a starting location; however, its learning process can only include subsets of the almost infinite paths across the image context, and may lead to major failures if not trained with adequate dataset variations. Here, we propose a new landmark detection approach inspired from DRL. Instead of learning limited action paths in an image in a greedy manner, we construct a global action map across the whole image, which divides the image into four action regions (left, right, up and bottom) depending on the relative location towards the target landmark. The action map guides how to move to reach the target landmark from any location of the input image. This effectively translates the landmark detection problem into an image partition problem which enables us to leverage a deep image-to-image network to train a supervised action classifier for detection of the landmarks. We discuss the experiment results of two ultrasound datasets (cardiac and obstetric) by applying the proposed algorithm. It shows consistent improvement over traditional machine learning based and deep learning based methods.

Keywords: Landmark detection · Deep learning · Image partition · Machine learning · Ultrasound

© Springer International Publishing AG 2017
M. Descoteaux et al. (Eds.): MICCAI 2017, Part III, LNCS 10435, pp. 338–346, 2017.
DOI: 10.1007/978-3-319-66179-7_39

1 Introduction

Landmarks are commonly used to represent anatomical features in medical imaging. Clinicians use landmarks to derive measurements (e.g., width, length, size, etc.) of organs for diagnosis, while radiologists and scientists register two images using corresponding sets of landmarks for further analyses. Ultrasound imaging is a widely used clinical procedure because it is safe, cost-effective, and non-invasive, where landmarks in a certain plane are used to provide diagnostic references. In cardiac ultrasound scans, landmarks are typically defined to measure the width at the intersections between heart chambers, for example, the annulus points of mitral valves; in obstetric ultrasound scans, landmarks at the anterior and posterior end of the fetal head are considered important. Manual localization of the landmark points, however, is tedious and time consuming. In an ultrasound machine, user needs to use the track ball to adjust the caliper to the desirable location, which makes the work even more complex. Furthermore, the reliability of the measurements can be suffered from the subjective disagreement across users. Automating the landmark detection can substantially reduce the manual efforts, and make the clinical procedure more efficient; however, this is a very challenging task given the (1) noisy signal, (2) low contrast, and (3) variations in shapes, orientations, and respiration phases throughout ultrasound images (Fig. 1).

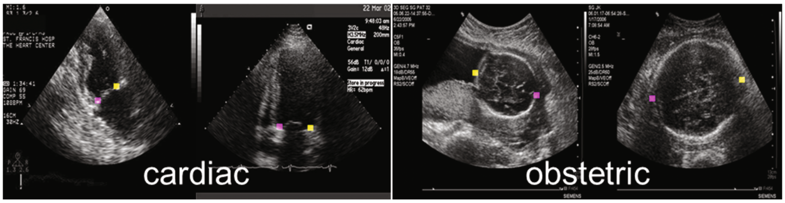

Fig. 1. Examples of cardiac and obstetric ultrasound scans. The magenta and yellow dots indicate the 1st and 2nd landmark, respectively.

The landmark detection problem has been studied using machine learning algorithms with reasonable outcomes. A bootstrapped binary classifier, e.g., probabilistic boosting-tree (PBT [1]), can be trained to distinguish landmark and non-landmark locations [2]; this approach can be biased due to the highly unbalanced positive and negative samples. Alternatively, landmark locations can be learned in a regression manner through aggregating pixel-wise relative distances to the landmark [3]; it provides more robustness, but less precision than the classification-based approach due to the complexity and variation of the image context. Recently, deep learning technologies have been adapted to medical imaging problems, and demonstrated promising performances by leveraging features trained from convolutional neural networks as opposed to hand-crafted features used in traditional machine learning approaches [4, 5]. For landmark detection, a deep reinforcement learning (DRL) approach has been shown successful to detect annulus points in cardiac ultrasound images [6]. The DRL algorithm designs an artificial agent to search and learn the optimized path from any location towards target by maximizing an action-value function. Its greedy searching strategy allows the agent to walk through only a subset of the almost infinite paths across the image instead of scanning exhaustively; however, this may lead to major failures if not trained with adequate dataset variations.

Here, we propose a new landmark detection approach inspired from DRL with the motivation of covering the entire searching space. We find that the optimal path can be broken down into optimal action steps at every pixel, while the pixel-wise optimal action steps can be derived given the landmark location based on Euclidian distances to generate an action map. Therefore, we can train a supervised action classifier (SAC) by explicitly learning the action steps across image instead of learning the actions implicitly along the searching path in DRL. The generation of action map effectively translates landmark detection into an image partitioning problem, where the highly unbalanced positive/negative sampling in PBT can be prevented. This also enables us to leverage a fully convolutional image-to-image neural network to train the SAC for estimating the action map. Furthermore, we design a robust aggregative approach to derive the landmark location from the estimated action map (Fig. 2), where our action-based aggregation is more precise than distance-based aggregation. To the best of our knowledge, we are the first to address landmark detection in the way of image partitioning. In this paper, we apply the proposed approach to a cardiac and an obstetric ultrasound dataset for landmark detection and compare the results with other learning-based methods.

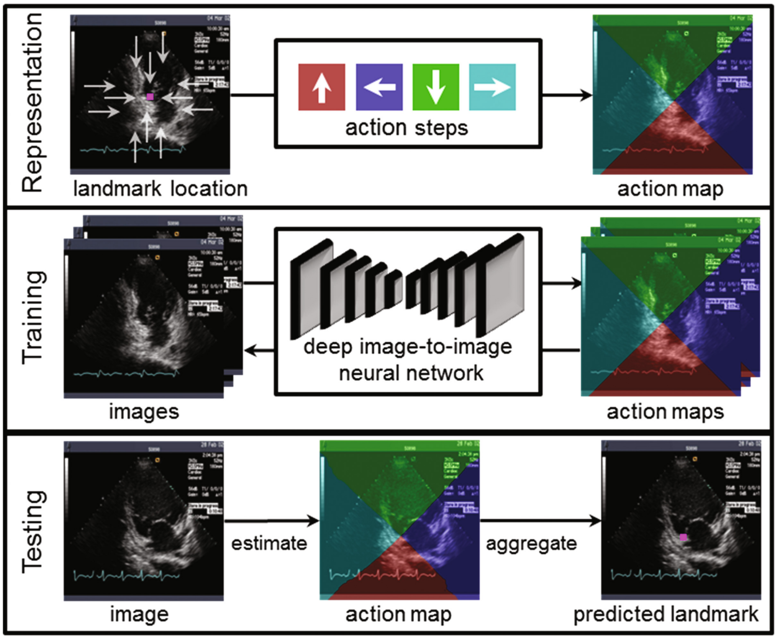

Fig. 2. Workflow of the proposed SAC approach. The action maps are generated based on the given landmark locations, then considered as the ground truths during the process of training a DI2IN. During testing, the trained DI2IN is applied to an unseen image to estimate an action map, based on which the predicted landmark location is aggregated.

2 Theory

2.1 Landmark Representation Based on Action Map

For the purpose of landmark detection, a landmark can be represented by an action map in terms of the pixel-wise optimal action step toward the landmark. Consider an optimal action path from any location at (x, y) towards landmark t at $\left(x_t, y_t\right)$ is composed of optimal action steps at pixels along the path on an image I. At each pixel, we define a unit movement $d_x^{(a)} = \{-1, 0, 1\}$ and $d_y^{(a)} = \{-1, 0, 1\}$. With the constraint of $d_x^{(a)} + d_y^{(a)} = 1$, we allow four possible action types $a \in \{0, 1, 2, 3\}$, i.e., up $\left(d_x^{(0)} = 0, d_y^{(0)} = -1\right)$, right $\left(d_x^{(1)} = 1, d_y^{(1)} = 0\right)$, down $\left(d_x^{(2)} = 0, d_y^{(2)} = 1\right)$, and left $\left(d_x^{(3)} = -1, d_y^{(3)} = 0\right)$, respectively. The optimal action step $\hat{a}$ is selected as the one with minimal Euclidian distance to landmark t after its associated movement,

$$\hat{a} = \operatorname{argmin}_a \sqrt{\left(x - x_t + d_x^{(a)}\right)^2 + \left(y - y_t + d_y^{(a)}\right)^2} \tag{1}$$

After cancelling out the common term, i.e., $\left(x - x_t\right)^2 + \left(y - y_t\right)^2 + 1$, $\hat{a}$ is shown to be dependent on the pixel location (x, y), where

$$\hat{a} = \operatorname{argmin}_a \left(x - x_t\right) d_x^{(a)} + \left(y - y_t\right) d_y^{(a)} \tag{2}$$

By replacing $d_x^{(a)}$ and $d_y^{(a)}$ with their actual values, the selection of $\hat{a}$ falls into four regions (one for each action type), where the regions are partitioned by two lines with slopes of ±1 crossing the landmark (see the top panel in Fig. 2), i.e., $y = x + \left(y_t - x_t\right)$ and $y = -x + \left(x_t + y_t\right)$. This generates an action map representing the pixel-wise optimal action step moving toward the target landmark location. For example, suppose one starts searching the landmark at a random location, say in the red region as show in Fig. 2, the optimal actions will keep moving up until hitting the line and then following the line to reach the target landmark. Using this action map representation, the landmark detection is essentially converted into an image partitioning problem.

2.2 Deep Image-to-Image Network Learning for Action Map Estimation

To estimate the action map for a given image, we employ a fully convolutional neural network given its efficient sampling scheme and large receptive field for comprehensive feature learning. Since both input (raw image) and output (action map) are images with the same size, we also call it a deep image-to-image network (DI2IN). Specifically, we follow the symmetric network architecture of SegNet [7]. The network is constructed with an encoder using the same structure as the fully convolutional part of VGG-16 network [4], and a decoder that replaces the pooling layers with upsampling layers and then essentially reverses the encoder structure. Batch normalization is used for each convolutional layer, and the max-pooling indices are kept during pooling and restored during upsampling. A softmax layer is used to provide categorical outputs, while cross-entropy loss is calculated and weighted by pre-computed class frequencies.

2.3 Action Map Aggregation for Landmark Detection

The landmark location needs to be derived from the estimated action map. However, the action map estimated by DI2IN may not always be in perfect shape as how it is constructed. There can be uncertainties around the partition lines between action types. It is also possible that there are islands of different action types, which are false predictions, inside a particular action partition. This undermines the robustness of lots of possible approaches for landmark derivation. For example, starting from a random point and moving along with the estimated action steps like DRL may not guarantee the convergence at the target landmark. Similarly, linear regression of the two partition lines may be disrupted even though the slopes are known, while dynamic programming based on the action flows can encounter dead locks. Here we propose an aggregative approach. With the output action map A from DI2IN, the estimated landmark location coordinates (x', y') are determined by maximizing an objective function $C(\cdot)$ summed up with that of each action type $C_a(\cdot)$.

$$x', y' = \operatorname{argmax}_{x,y} C(x, y) = \operatorname{argmax}_{x,y} \sum_a C_a(x, y) \tag{3}$$

where the action-wise objective function at pixel (x, y) is aggregated by the pixels with that specific action on the same row or column, specifically

$$C_a(x, y) = \begin{cases} d_x^{(a)} \left(\sum_{i=x}^{\infty} \delta(A(i, y) = a) - \sum_{i=-\infty}^{x} \delta(A(i, y) = a) \right) & if \left| d_x^{(a)} \right| = 1 \\ d_y^{(a)} \left(\sum_{j=y}^{\infty} \delta(A(x, j) = a) - \sum_{j=-\infty}^{y} \delta(A(x, j) = a) \right) & if \left| d_y^{(a)} \right| = 1 \end{cases} \tag{4}$$

Note that the objective function increments with pixels pointing towards (x, y), while decrements with pixels pointing away from (x, y) (Fig. 3). Such aggregation enables robust location coordinate derivation even with suboptimal action map from the DI2IN output.

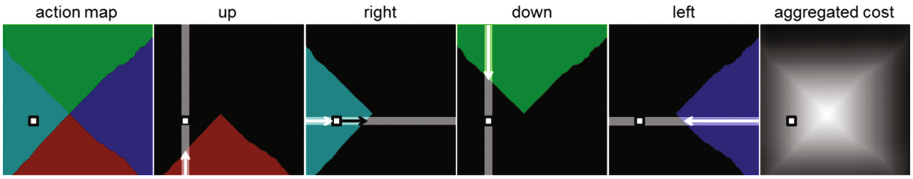

Fig. 3. Illustration of action map aggregation on a single pixel. White arrow indicates value increase of the objective function, while black arrow indicates value decrease.

3 Methods and Results

3.1 Data

Two ultrasound datasets are used in this study including a cardiac and an obstetric dataset with 1353 and 1642 patients, respectively. Both datasets are collected and anonymized in the process of clinical routine. Landmarks of interest are annotated by clinical experts on each scan. We collect 8892 frames in the cardiac dataset in total across the entire heart cycles rather than collect the images just around end-systole and end-diastole as in [6]. Therefore, the cardiac dataset in our experiment presents larger contextual variations and greater challenges for landmark detection. On a cardiac scan, landmarks are defined as the two annulus points, which are the roots of mitral valve in apical 2 chamber (A2C) view and apical 4 chamber (A4C) view. In the obstetric dataset, each patient has only one scanned image. On an obstetric scan, the first landmark is annotated at the anterior end of the fetal head, while the second is at the posterior end. Note that the orientations of fetal head can essentially cover 360° across the ultrasound scans. Therefore, detecting these two landmarks on an obstetric scan is not an easy task even for humans. Careful identification of the internal brain structure is necessary for consistent manual annotation. For each dataset, 80% patients are randomly selected as training set, and the remaining 20% are used for testing. All images are normalized into 480 × 480 before further processing.

3.2 Experimental Setup

We apply the proposed approach to the cardiac and the obstetric ultrasound datasets individually. For each landmark, we train a DI2IN to learn its associated action map. The DI2IN are trained using the Caffe framework on a Linux workstation equipped with an Intel 3.50 GHz CPU and a 12GB NVidia Titan X GPU. The encoder part of DI2IN of is initialized with the weights of VGG-16 trained from ImageNet. During training, the mini batch size is set to 2, standard stochastic gradient descent is used for updates with learning rate as 1e–3 and momentum as 0.9 through 80,000 iterations. We compare the proposed SAC with other learning-based approaches on the same dataset including PBT, DRL, and a state-of-the-art regression-based approach using DI2IN [8] (we refer to it as I2I). Note that I2I and SAC uses similar network structure, while representing the landmark differently. For each method, we try our best to tune the configuration to provide reasonably good results. Distance error of landmark position in pixels is used for comparison since all images are in normalized space.

3.3 Qualitative and Quantitative Results

The action maps estimated from SAC (Fig. 4) are clean (very few islands of false predictions) and smooth (sharp separations between regions of different action types). It turns out to be beneficial to keep the pooling indices in DI2IN, which enforcing the smoothness of the estimated action map. Overall, the action maps look very reasonable even though they are not exactly the same as the ground truth (the partitioning lines are not straight). The derived landmark locations from the estimated action maps are also close to those of the manual annotations.

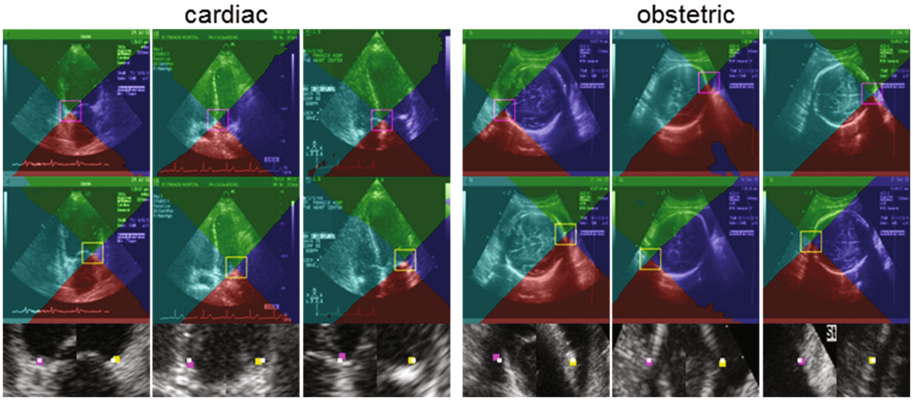

Fig. 4. Example landmark detection results. The first two rows present the action maps estimated from DI2IN for the first and second landmark, respectively. The last row demonstrates the two predicted landmark locations (magenta and yellow dots) along with the manual annotation (white dot, smaller than the magenta and yellow dots) by zooming in the local patches in the first two rows, where the left patch corresponds to the first row, and the right patch corresponds to the second row.

For cardiac scans, it is not too hard to identify a rough location of the target landmarks in the middle of left ventricle and left atrium; however, it is challenging to have precise localization given that we include cardiac phases throughout heart cycles, where the relative locations vary a lot between the annulus points of mitral valves and the surrounding structures. Overall, compared to PBT and DRL, our proposed method provides consistently better accuracy and robustness (Table 1). Compared to I2I, SAC presents slightly better overall performances.

Table 1. Distance errors of landmark detection in pixels.

		PBT		DRL		I2I		SAC	
		lmk1	lmk2	lmk1	lmk2	lmk1	lmk2	lmk1	lmk2
CA	Mean	10.45	13.85	7.69	10.02	6.73	9.02	**6.31**	**8.01**
	50%	5.74	8.11	5.43	7.63	5.00	6.40	**4.35**	**5.88**
	80%	11.11	16.18	9.33	13.73	8.54	11.40	**7.54**	**10.83**
OB	Mean	59.23	130.66	29.99	32.45	30.07	21.97	**14.94**	**16.76**
	50%	35.31	139.49	11.69	13.17	5.39	6.08	**4.85**	**5.91**
	80%	109.84	193.64	43.98	45.76	13.34	15.54	**11.76**	**13.67**

Note that the best performance for each landmark is highlighted in bold. CA indicates cardiac scans, while OB indicates obstetric scans. 50% and 80 % indicate median and 80 percentile, respectively. Across all tests, our method presents significant improvements over other methods statistically ($p < 0.05$, t-test).

For the obstetric scans, it is very hard to identify the landmark location correctly without capturing the context in a large receptive field given lots of ambiguities around the almost radially symmetric structure. It is very likely to make major failures, especially if only local context are used for feature modeling (PBT and DRL), while confusion of head orientation can be substantially prevented using DI2IN (I2I and SAC). SAC demonstrates the best performance among all tested methods.

4 Discussion

In this paper, we introduce a new perspective to address landmark detection; we propose a novel approach inspired from DRL by converting the landmark detection problem into a supervised image partition task in the form of action maps. This landmark-to-image conversion enables the classifier to not only sample data in a more balanced manner (compared to PBT), but also capture more comprehensive image context across the entire image for the guidance of landmark detection (compared to DRL). Based on this conversion, we formulate a complete workflow by leveraging a deep DI2IN for action map estimation, and designing an action map aggregation for landmark estimation. Using this workflow, we present competitive performances against other state-of-the-art approaches on cardiac and obstetric ultrasound datasets. Further investigation on its clinical value will be performed as our next step, where more training data will be used for better performance, more engineering efforts will be spent for faster and smoother detection across frames, and more evaluations will be focused on the measurements derived from landmarks against human errors.

Our SAC approach is generic, and it has great synergy with DI2IN as observed in our experiments on ultrasound datasets. We observe big opportunities to improve the performances by integrating new technologies of training DI2IN, e.g., deep supervision [9] and skip connection [10]. Meanwhile, given the promising results in 2-D ultrasound for single landmark detection, it is worthwhile to explore the extension of SAC in (1) 3-D, (2) other image modalities, and (3) multi-landmark detection, where the action map generation and aggregation need to be adapted.

References

1. Tu, Z.: Probabilistic boosting-tree: learning discriminative models for classification, recognition, and clustering. In: Tenth IEEE International Conference on Computer Vision (ICCV 2005), vol. 2. IEEE (2005)
2. Viola, P., Jones, M.: Fast and robust classification using asymmetric adaboost and a detector cascade. In: Advances in Neural Information Processing System, vol. 14 (2001)
3. Zhou, S.K., Comaniciu, D.: Shape regression machine. In: Karssemeijer, N., Lelieveldt, B. (eds.) IPMI 2007. LNCS, vol. 4584, pp. 13–25. Springer, Heidelberg (2007). doi: 10.1007/978-3-540-73273-0_2
4. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
5. Long, J., et al.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2015)

6. Ghesu, F.C., Georgescu, B., Mansi, T., Neumann, D., Hornegger, J., Comaniciu, D.: An artificial agent for anatomical landmark detection in medical images. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9902, pp. 229–237. Springer, Cham (2016). doi:10.1007/978-3-319-46726-9_27
7. Badrinarayanan, V., et al.: SegNet: a deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling. arXiv preprint arXiv:1505.07293 (2015)
8. Yang, D., et al.: Automatic vertebra labeling in large-scale 3D CT using deep image-to-image network with message passing and sparsity regularization. In: Niethammer, M., Styner, M., Aylward, S., Zhu, H., Oguz, I., Yap, P.-T., Shen, D. (eds.) IPMI 2017. LNCS, vol. 10265, pp. 633–644. Springer, Cham (2017). doi:10.1007/978-3-319-59050-9_50
9. Xie, S., Tu, Z.: Holistically-nested edge detection. In: Proceedings of the IEEE International Conference on Computer Vision (2015)
10. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). doi:10.1007/978-3-319-24574-4_28