Multimodal Image Registration with Deep Context Reinforcement Learning

Kai Ma[1(✉)], Jiangping Wang[1], Vivek Singh[1], Birgi Tamersoy[2], Yao-Jen Chang[1], Andreas Wimmer[2], and Terrence Chen[1]

[1] Medical Imaging Technologies, Siemens Medical Solutions USA, Inc., Princeton, NJ 08540, USA
kai.ma@siemens.com

[2] Siemens Healthcare GmbH, Forchheim, Germany

Abstract. Automatic and robust registration between real-time patient imaging and pre-operative data (e.g. CT and MRI) is crucial for computer-aided interventions and AR-based navigation guidance. In this paper, we present a novel approach to automatically align range image of the patient with pre-operative CT images. Unlike existing approaches based on the surface similarity optimization process, our algorithm leverages the contextual information of medical images to resolve data ambiguities and improve robustness. The proposed algorithm is derived from deep reinforcement learning algorithm that automatically learns to extract optimal feature representation to reduce the appearance discrepancy between these two modalities. Quantitative evaluations on 1788 pairs of CT and depth images from real clinical setting demonstrate that the proposed method achieves the state-of-the-art performance.

1 Introduction

Depth sensing technologies using structured light or time-of-flight become popular in recent years. Their applications have also been widely studied in the healthcare domain, such as patient monitoring [1], patient positioning [16] and computer-aided interventions [19]. In general, depth imaging provides real-time and non-intrusive 3D perception of patients that could be used for markerless registration, to replace conventional RGB cameras, and potentially to achieve higher robustness against illumination and other data variability.

To enable such clinical applications, one of the fundamental steps is to align the pre-operative image such as CT or MRI, with the real-time patient image from the depth sensor. This requires an efficient and accurate registration or ego-positioning algorithm. As depth sensors capture the 3D geometric surface of the patient while skin surface can be readily extracted from CT scans, surface-based registration methods [2,14,19] have been intuitively proposed. However, those methods usually fail to perform robustly due to several challenges: (1) the surface

Electronic supplementary material The online version of this chapter (doi:10.1007/978-3-319-66182-7_28) contains supplementary material, which is available to authorized users.

M. Descoteaux et al. (Eds.): MICCAI 2017, Part I, LNCS 10433, pp. 240–248, 2017.
DOI: 10.1007/978-3-319-66182-7_28

data obtained from the depth sensor is noisy and suffers from occlusions; (2) the surface similarity is tampered due to the patients' clothing or protective covers; (3) the two modalities may have a different field of view. CT data, for example, often only covers a part of the patient's body; (4) the patient's pose/shape may vary between the two imaging processes. To overcome these challenges, most of the existing solutions still rely on marker-based approaches [5].

Another way to formulate the depth-CT registration problem is to utilize the internal body information that the CT scan naturally captures. Unfortunately, the physical principles used in the depth sensing and CT imaging are so different that the information from the two modalities has little in common. To measure the similarity between different modalities, learning-based algorithms have been actively explored [4,15]. Most recently, there has been a significant progress in feature representation learning using deep convolutional neural networks, which can extract hierarchical features directly from raw visual input. The high level features encode large contextual information which are robust against noise and other data variations. Moreover, by combining deep convolutional neural network with reinforcement learning, the deep reinforcement learning (DRL) has demonstrated superhuman performance in different applications [10,13].

In this paper, we propose a deep reinforcement learning based multimodal registration method that handles the aforementioned challenges. An overview of the system algorithm workflow is shown in Fig. 1. Our major contributions are summarized as follows: (1) We propose a learning-based system derived from deep Q-learning [13] that automatically extracts compact feature representations to reduce the appearance discrepancy between depth and CT data. It is the first time a state-of-the-art DRL method is used to solve the multimodal registration problem in an end-to-end fashion. (2) We also propose to use the contextual information for the depth-CT registration. Compared to conventional methods that compute surface similarities, our algorithm learns to exploit the relevant contextual information for optimal registration.

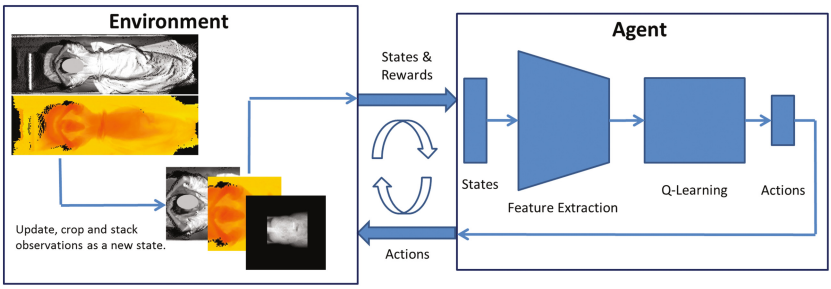

Fig. 1. Run-time workflow of the proposed DRL registration framework. The iterative observe-action process gradually aligns the multimodal data until termination.

2 Related Work

Registration of multimodal data recently attracts increasing attention on medical use cases. Different information is extracted and fused from different modality scans to provide pieces of an overall picture of pathologies. In general, most of the multimodal registration (MMR) approaches can be categorized as one of the two types. The first category algorithms attempt to locate invariant image features [2,17], while the second category approaches apply statistical analysis such as regression to find a metric that measures dependency between two modalities [4,7]. Different from those approaches, our method learns both feature representations and alignment metric implicitly in an end-to-end fashion with DRL.

DRL is a powerful algorithm that trains an agent which interacts with an environment, with image observations and rewards as the input, to output a sequence of actions. The working mechanism makes it suitable to solve the sequential decision making problems, for example the landmark detection in medical images with trajectory learning [6]. To the best of our knowledge, the most relevant registration work is proposed in [11]. They solve the 3D CT volume registration problem with a standard deep Q-learning framework. To speedup the training process with the 6 degree-of-freedom transformation, they replace the agent's greedy exploration process with a supervised learning scheme. In our scenario, due to the appearance discrepancies as well as ambiguities due to missing observations, we instead encourage the agent to explore the search space freely rather than exploiting the shortest path. Furthermore, we utilize the history of actions to help agent escape from local loops caused by the incorrect initialization, which differentiates our work from theirs.

3 Method

We propose a novel MMR algorithm that aligns the depth data to the medical scan. Our work is inspired by the process of how human experts perform the manual image alignment, which can be described as an iterative observe-action process. Similarly, the DRL algorithm trains an agent with observations from environment to learn a control policy, which is reflected by the capability of making sequential alignment actions with given observations. The rest of the section will reveal more details of the proposed registration method.

3.1 Environment Setup

In deep reinforcement learning, the environment E is organized as a stochastic finite state machine. It takes agent's action as the input and outputs states and rewards. The agent is designed to have zero knowledge about the internal model of the environment, besides the observed states and rewards.

States: In our setup, the state is represented by a 3D tensor consisting of cropped images from both data modalities. At the beginning of each training episode, the environment is initialized either randomly or roughly to align the two data

Fig. 2. The derived dueling network architecture used in the proposed method.

sources. A fixed size window is applied to crop the depth image with current transformation, where the cropped image is stacked with the projected CT data (Sect. 3.3) as an output state. In the following iterations, a new action output from the agent is used to update the transformation accordingly.

Rewards: Given a state s_t, a reward r_t is generated to reflect the value of current action a_t given by the agent. A small reward value is given to the agent during the regular exploration steps, while the terminal state triggers a much larger reward. The sign of the reward is determined by the current distance to the ground truth compared to the previous step.

3.2 Training the Agent

Let I_d represent the depth image and I_t represent the projected CT image. The goal here is to estimate the rigid transformation T that aligns the moving image I_t to the fixed image I_d with a minimal error. A common method to find the optimal parameters of T is by maximizing a similarity function $S(I_d, I_t)$ with a metric. Instead of applying a manually defined metric, we adopt the reinforcement learning algorithm to implicitly learn the metric. The optimization process is recast as a Markov Decision Process following the Bellman equation [3]. More precisely, we train an agent to approximate the optimal action-value function by maximizing the cumulative future reward [13]. Different from the deep-Q network, the proposed method is derived from the Dueling Network [18] with some modifications (Fig. 2):

- We add more convolution and pooling layers to make the network deep enough to extract high-level contextual features.
- We add batch normalization layer after the input data layer to minimize the effect of intensity distribution discrepancy across different modalities.
- We concatenate the feature vector extracted from the last convolution layer with an action history vector that records the actions of the past few frames. In our experiment, the concatenation of the action history vector alleviates the action oscillation problem around certain image positions.

The insight behind the dueling network is that certain states include more critical information than others to help the agent make the right decision. For example, during the chest region registration, getting the head region rather than the arms within the observation will significantly help the agent move toward the right direction. Compared to the deep Q-network, the dueling network has the capability of providing separate estimates of the value and advantage functions, which allow for a better approximation of the state values. In our setup, the final Q value function is formulated as:

$$Q(s,h,a;\theta,\alpha,\beta) = V(s,h;\theta,\beta) + (A(s,h,a;\theta,\alpha) - max_{a'}A(s,h,a';\theta,\alpha)) \quad (1)$$

where h is the history action vector, θ is the convolution layers' parameters, α and β are the parameters of the two streams of fully-connected layers. To further stabilize the training process, double DQN [8] is also adopted to update the network weights.

3.3 Data Projection

The two data modalities in our scenario are the 2.5D depth image and 3D CT volume data. One way to align the two modalities is to reconstruct the depth image to a 3D surface, and then apply the registration algorithm in the 3D space. However, feature learning with the 3D convolution requires tremendous computation. Meanwhile, the DRL algorithm with a greedy exploration policy has to explore millions of observations to properly train an agent. To reduce the computation complexity and speedup the training process, we reformulate the 2.5D-3D registration problem to a 2D image registration problem. We simplify the 3D volume data to a 2D image through a projection process. Note that the simplification is only for speedup purpose and the proposed workflow can be extended to the 2.5-3D registration with minor modifications.

To best utilize the internal information that CT data naturally captures, we project the CT volume to a 2D image using the following equation.

$$I_t(x,y) = \frac{1}{h} \times \sum_{z=0}^{h} CT(x,y,z) \quad (2)$$

where h is the size of the CT volume along the anterior axis. The intensity of each pixel on the projected image is the summation of the voxel readings of the volume along the projection path. We apply an orthographic projection for both depth data and volume data, and Fig. 3 shows an example of the projected images. The projected image of volume data is visually similar to a topogram image. Since the medical scans often only have a partial view of the patient, it is challenging even for a human expert to align the two modalities from the surface, especially over the flat regions such as the chest and the abdomen. On the contrary, the topogram-like image reveals more contextual information of the internal structures of the patient to better handle the data ambiguity problem, compared to the surface representation.

Fig. 3. Orthographically projected CT and depth images. Left image shows a CT abdomen scan in a larger scale. Middle image shows a depth image rendered in color. Right image displays the overlay of the two modalities with the ground truth.

Although the depth-CT registration involves a six degree-of-freedom transformation, we simplify the search space into two translations T_R (along the Right axis in the RAS coordinate system), T_S (along the Superior axis) and one rotation R_A (along the Anterior axis). The rest of the transformation can be determined/inferred through the sensor calibration process together with the depth sensor readings. For example, the relative translation offset along the Anterior axis can be calculated by deducting the actual distance between table and camera from the distance recorded during the calibration time.

4 Experiments and Results

We installed Microsoft Kinect2 cameras to the ceilings of clinical CT-scan rooms. Depth images were collected when the patient lay down on the table and adjusted the pose for the scan. We took several snapshots during the positioning process. We reconstruct the depth image to a 3D point cloud and orthographically re-project the point cloud to a 2D image. We also reconstruct the patient's CT data with full FOV to avoid cropping artifacts. The two imaging systems, Kinect2 and CT scanner, can be pre-calibrated through a standard extrinsic calibration process [12]. As long as the patient remains stationary during the two imaging processes, the ground truth alignment of the two data modalities can be determined from the table movement offsets and the extrinsic parameters.

We collect two datasets that consist of thorax and abdomen/pelvis scans, which ends up with 1788 depth-CT pairs across several clinical sites. We randomly split the training and testing set for each experiment and guarantee each training set have 800 data. The rest of them is used as the testing data. We also add random perturbations to the training data to avoid overfitting.

The network configuration is shown in Fig. 2. The input images are cropped with the same size (200×200) at a resolution of 5 mm. The network output is a 6D vector (4 translations and 2 rotations). The action history vector has a length of 24 (6 actions $\times$ 4 histories). We use RMSprop optimizer without the momentum to update network weights. The learning rate is initially set to 0.00002 with a decay of 0.95 every 10,000 iterations. The mini-batch size is 32. γ equals to 0.9. We randomly initialize the transformation with a translation offset ± 500 mm and a rotation offset $\pm 30°$ from the ground truth location, to start training the agent. The non-terminal rewards are ± 0.1 and the terminal rewards

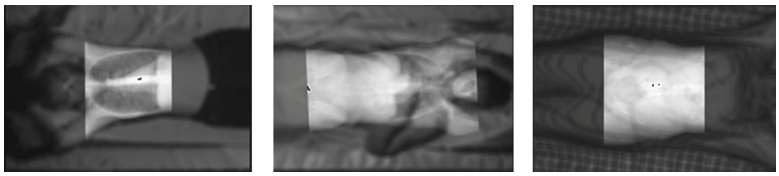

Fig. 4. Qualitative impression with the proposed algorithm. Left image is a perfect thorax alignment. Middle one is a good thorax alignment though the patient's poses at the two imaging time were different. Right image shows a perfect abdomen alignment.

are ±10. For each dataset, we train an agent with a single TitanX Pascal GPU for 1.2M iterations and each of the training lasts about 4 days.

System performance is reported as the average Euclidean distance between the network estimation and the ground truth. We compare the performance with several baseline approaches as well as different DRL networks. The landmark baseline [6] trains detectors to detect surface landmarks, such as shoulders and pelvises, to align with the CT anatomy landmarks. The Haustorff baseline minimizes the surface distance between CT and depth in 3D with the Haustorff metric. The ICP baseline aligns the two surfaces with the standard ICP algorithm. The DQN baseline is configured with the original setup [13]. The Dueling Network [18] is similar to our proposed method but configured with the original setup. We also test the proposed network without history information and batch normalization [9] separately. The quantitative accuracy comparison among all methods is shown in Table 1 as well as the computation time. A qualitative analysis of the results generated by the proposed method is shown in Fig. 4.

Table 1. Results comparison of thorax and abdomen (ABD) dataset.

Methods	Region	T_S (mm)	T_R (mm)	R_A (◦)		T_S (mm)	T_R (mm)	R_A (◦)	Time (s)
Landmark [6]	Thorax	36.1 ± 19.7	7.3 ± 2.1	-	ABD	47.8 ± 25.6	7.2 ± 3.7	-	**0.06**
Hausdorff	Thorax	14.6 ± 7.1	5.9 ± 4.4	-	ABD	20.2 ± 16.8	9.3 ± 6.1	-	11.8
ICP	Thorax	18.3 ± 9.5	5.1 ± 2.2	4.2 ± 2.2	ABD	25.9 ± 18.2	11.2 ± 2.7	5.1 ± 3.4	2.35
DQN [13]	Thorax	27.3 ± 5.9	4.9 ± 1.5	7.2 ± 1.8	ABD	33.2 ± 9.3	9.1 ± 2.2	4.8 ± 2.2	1.37
Dueling [18]	Thorax	19.7 ± 6.2	5.2 ± 1.3	6.9 ± 1.4	ABD	22.4 ± 10.5	7.6 ± 3.3	6.3 ± 2.4	1.40
Proposed	Thorax		**2.7 ± 1.5**	**2.5 ± 0.4**	ABD	**15.2 ± 5.8**	**4.6 ± 1.9**	**2.9 ± 0.8**	1.42
– w/o history	Thorax	**9.1 ± 3.7**	4.2 ± 1.5	3.1 ± 1.4	ABD	19.2 ± 8.3	7.3 ± 2.9	2.9 ± 0.9	1.42
– w/o BN [9]	Thorax	11.5 ± 6.4	4.7 ± 1.4	6.2 ± 1.7	ABD	22.8 ± 8.2	7.1 ± 2.9	3.4 ± 1.5	1.41
		17.7 ± 6.9							

5 Conclusion and Future Work

A novel depth-CT registration method based on deep reinforcement learning is proposed. Our approach investigates the correlations between surface readings from depth sensors and internal body structures captured by the CT imaging. The experimental results demonstrate that our approach reaches the best accuracy with the least deviation. The better performance compared to two original

DRL methods suggests that our modifications improve the network learning for the multimodal registration. Higher errors in the abdomen cases, compared to the chest cases, may be caused by the larger appearance variations. The proposed approach also has no limitations to be applied to register images from other modalities. Future research direction includes combining the surface metric together with the contextual information to further improve performance. Extra efforts are also required to improve the training and testing efficiency.

References

1. Achilles, F., Ichim, A.-E., Coskun, H., Tombari, F., Noachtar, S., Navab, N.: Patient MoCap: human pose estimation under blanket occlusion for hospital monitoring applications. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9900, pp. 491–499. Springer, Cham (2016). doi:10.1007/978-3-319-46720-7_57
2. Bauer, S., Wasza, J., Haase, S., Marosi, N., Hornegger, J.: Multi-modal surface registration for markerless initial patient setup in radiation therapy using Microsoft's Kinect sensor. In: ICCV Workshops (2011)
3. Bellman, R.: A Markovian decision process. Indiana Univ. Math. J. **6**, 679–684 (1957)
4. Cao, X., Gao, Y., Yang, J., Wu, G., Shen, D.: Learning-based multimodal image registration for prostate cancer radiation therapy. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9902, pp. 1–9. Springer, Cham (2016). doi:10.1007/978-3-319-46726-9_1
5. Elmi-Terander, A., Skulason, H., Söderman, M., et al.: Surgical navigation technology based on augmented reality and integrated 3D intraoperative imaging: a spine cadaveric feasibility and accuracy study. Spine **41**, 303–311 (2016)
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. Gutiérrez-Becker, B., Mateus, D., Peter, L., Navab, N.: Learning optimization updates for multimodal registration. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9902, pp. 19–27. Springer, Cham (2016). doi:10.1007/978-3-319-46726-9_3
8. Hasselt, H., Guez, A., Silver, D.: Deep reinforcement learning with double Q-learning. In: AAAI (2016)
9. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift. In: arXiv (2015)
10. Levine, S., Pastor, P., Krizhevsky, A., Quillen, D.: Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. In: ISER (2016)
11. Liao, R., Miao, S., de Tournemire, P., Grbic, S., Kamen, A., Mansi, T., Comaniciu, D.: An artificial agent for robust image registration. In: AAAI (2017)
12. Ma, K., Chang, Y.J., Singh, V.K., O'donnell, T., Wels, M., Betz, T., Wimmer, A., Chen, T.: Calibrating RGB-D sensors to medical image scanners. US Patent 9,633,435

13. Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., Hassabis, D.: Human-level control through deep reinforcement learning. Nature **518**, 529–533 (2015)
14. Nutti, B., Kronander, S., Nilsing, M., Maad, K., Svensson, C., Li, H.: Depth sensor-based realtime tumor tracking for accurate radiation therapy. In: Eurographics (2014)
15. Simonovsky, M., Gutiérrez-Becker, B., Mateus, D., Navab, N., Komodakis, N.: A deep metric for multimodal registration. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9902, pp. 10–18. Springer, Cham (2016). doi:10.1007/978-3-319-46726-9_2
16. Singh, V., Chang, Y., Ma, K., Wels, M., Soza, G., Chen, T.: Estimating a patient surface model for optimizing the medical scanning workflow. In: Golland, P., Hata, N., Barillot, C., Hornegger, J., Howe, R. (eds.) MICCAI 2014. LNCS, vol. 8673, pp. 472–479. Springer, Cham (2014). doi:10.1007/978-3-319-10404-1_59
17. Toews, M., Zöllei, L., Wells, W.M.: Feature-based alignment of volumetric multi-modal images. Inf. Process. Med. Imaging **23**, 25–36 (2013)
18. Wang, Z., de Freitas, N., Lanctot, M.: Dueling network architectures for deep reinforcement learning. In: ICML (2016)
19. Xiao, D., Luo, H., Jia, F., Zhang, Y., Li, Y., Guo, X., Cai, W., Fang, C., Fan, Y., Zheng, H., Hu, Q.: A Kinect camera based navigation system for percutaneous abdominal puncture. Phys. Med. Biol. **61**, 5687–5705 (2016)