Table Motion Detection in Interventional Coronary Angiography

Junaid R. Rajput[1,2], Karthik Shetty[1,2], Andreas Maier[1], Martin Berger[2]

[1]Pattern Recognition Lab, FAU Erlangen-Nürnberg
[2]Siemens Healthcare GmbH, Forchheim, Germany
junaid.rajput@fau.de

Abstract. The most common method for detecting coronary artery stenosis is interventional coronary angiography (ICA). However, 2-D angiography has limitations because it displays complex 3-D structures of arteries as 2-D X-ray projections. To overcome these limitations, 3-D models or tomographic images of the arterial tree can be reconstructed from 2-D projections. The 3-D modeling process of the arterial tree requires accurate acquisition geometry since in many ICA acquisitions the patient table is translated to cover the entire area of interest, the original calibrated geometry is no longer valid for the 3-D reconstruction process. This study presents methods for identifying the frames acquired during table translation in an angiographic scene. Spatio-temporal methods based on deep learning were used to identify translated frames. Three different architectures − 3-D convolutional neural network (CNN), bi-directional convolutional long short term memory (CONVLSTM), and fusion of bi-directional CONVLSTM and 3-D CNN − were trained and tested. The combination of CONVLSTM and 3-D CNN surpasses the other two methods and achieves a macro f1-score (mean f1-scores of two classes) of 93%.

1 Introduction

X-ray based interventional coronary angiography (ICA) is the most widely used technique to detect coronary diseases, which is still considered the gold standard for diagnosis and treatment. The 2-D X-ray image of the 3-D vessel structure lacks depth information, resulting in projection artifacts such as vessel overlap and foreshortening. These inherent limitations of ICA can be overcome by reconstructing a 3-D representation of the arterial tree that allows 3-D quantitative coronary analysis (QCA) [1]. This can be done either by creating a symbolic 3-D model of the coronary tree or by reconstructing a tomographic volume [2].

The symbolic 3-D image of the coronary tree can be reconstructed from at least two projection images of the arterial tree obtained at least 30° apart for an identical cardiac phase with minimum vessel overlap [3]. This requires prior information about the acquisition geometry, e.g., the distance between source and image, the pixel spacing of the detector, and the distance between source

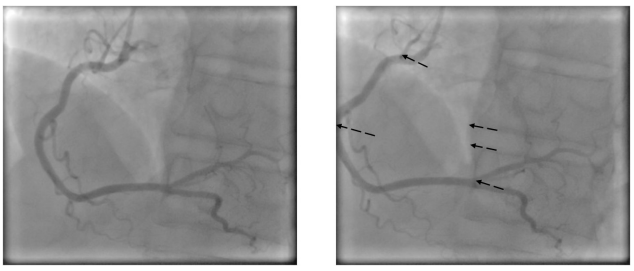
and isocenter. The translation of the patient table which regularly occurs in ICA acquisitions to cover the entire arterial tree in projection images, disrupts the original imaging geometry. Hence, that initial calibrated geometry does not remain valid for reconstruction. An example of a right coronary artery (RCA) acquisition with typical table translation is illustrated in Fig. 1. In prior work, an additional translation vector based on the known distance of table displacement was integrated into the optimization of correspondence matching in projection images [4]. In another study, the translation vector was integrated into physical space instead of in projections [5]. However, these methods require prior information about the table displacement and manual selection of centerline correspondences of two projection images.

Spatio-temporal neural networks can detect and distinguish movements in a 3-D scene, and they have outperformed traditional machine learning algorithms [6]. This work attempts to use these spatio-temporal networks to identify the frames in ICA acquisitions that are recorded during the table translation. This information allows to accurately determine when exactly in an acquisition the table has been moved, i.e., which frames are deviating from the known acquisition geometry. Based on that, a subset of frames from different view angles, which still correspond to the given acquisition geometries, can be selected automatically.

2 Materials and methods

2.1 Datasets and annotation

The datasets used for all experiments consisted of angiographic scenes from three different clinical sites. All scenes were acquired at 15 frames per second (FPS), with varying scene length, image size, and pixel spacing using Siemens Artis imaging devices (Siemens Healthineers, Erlangen, Germany). All required information, e.g., the corresponding electrocardiogram (ECG), the orientation of the scene, and the image data were stored as DICOM files. A total of 158 scenes from 28 patients were manually annotated, frames were labeled as translated frames (moving) if the rib or spine changed position compared to the previous frame, otherwise frames were termed static. Tab. 1 provides a detailed overview of the datasets.



(a) Before translation. (b) After translation.

Fig. 1. Impact of table translation on angiographic scene.

Table 1. Detailed overview of the angiographic scenes.

Site	Number of scenes	Table movement	Left coronary artery (LCA)	Right coronary artery (RCA)	Frames per second	Pixel spacing
Training and validation sets						
Site 1	104	94	74	30	15	0.18 mm
Site 2	2	2	1	1	15	0.16 mm
Site 3	2	2	0	2	15	0.4 mm
Test set						
Site 1	24	23	10	14	15	0.18 mm
Site 2	25	17	19	6	15	0.16 mm
Site 3	1	1	0	1	15	0.4 mm

2.2 Preprocessing

The angiographic scenes were of varying lengths, with a number of 12 to 182 frames. Fig. 2 depicts the overview of the preprocessing pipeline. The scenes were divided into temporal blocks so that each block of the scene has 32 frames − for the training, blocks of 32 frames were extracted from the angiographic scenes without any overlap. If the scene length was not a multiple of 32, temporally symmetrically mirrored frames were added to fill the blocks, e.g. the second block of the scene with 40 frames consists of [33, 34, . . . , 40, 39, 38, . . . , 17, 16] frames − then each frame of these temporal blocks was resized to 512×512, afterwards contrast was enhanced using adaptive histogram equalization, and in the final step the frames were downsampled to 128×128 using minimum pooling. The application of minimum pooling was to highlight high-density structures (ribs and spine) in the angiography scene. During the training, the first frame of each block was marked as static. During the evaluation, the scene was split up so that the following block starts with the last frame of the previous block, i.e. if the scene has 63 frames, then block 1 and 2 have frames 1 - 32 and 32 - 63 respectively, the prediction of frame number 32 comes from block 1.

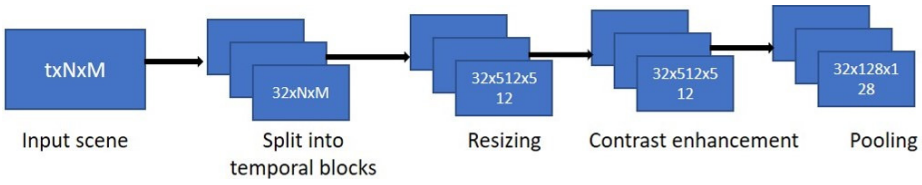


Fig. 2. Overview of the preprocessing pipeline for angiographic input scenes.

2.3 Network architectures

Three different spatio-temporal networks were designed and trained. Fig. 2 shows the overview of network architectures. Network 1 consisted of a 3-D convolutional neural network (CNN), 3-D CNN captures the spatial and temporal information

among adjacent frames of the scene [7]. Network 2 consisted of bi-directional convolutional long short term memory (CONVLSTM) layers. This layer contains a convolution operation within the LSTM cell, allowing it to learn spatial information and long-term dependencies between frames of the scene [8]. Network 3 (3-D CNN-CONVLSTM) contained both 3-D CNN and the CONVLSTM layers. The bi-directional CONVLSTM layer stacked on top of the 3-D CNN layer in the fusion network enables the learning of longer temporal information with much less complexity compared to the CONVLSTM network.

3 Results

All three networks were trained with the binary cross-entropy loss and the Adam optimizer. The class imbalance was addressed by adding multiple copies of temporal blocks that consisted only of moving frames from the respective training and validation sets. Tab. 2 gives an overview of the performance of network architectures for the test set. The 3-D CNN-CONVLSTM network outperformed the other two networks and was able to successfully predict the moving frames based on the rigid structures of the angiography. It achieved a macro f1-score of 93% on the complete test set, the macro f1-score was 93%, 92%, 95% for Site 1, Site 2, and Site 3. Macro f1-score was 94% and 92% for RCA and LCA scenes, respectively. The results for Site 2 and RCA show that the trained model generalized well, as they were severely underrepresented in the training data. Increasing the depth of CONVLSTM and 3-D CNN architectures leads to overfitting and further degrades their performance.

Fig. 4a shows the binary classification and the ground truth labels for an example scene. The angiographic scene was resliced along a vertical line in temporal direction (cf. right Fig. 4b), highlighting a part that focuses on a selected vertebra. The superimposition of a selected vertebra and the classification results in Fig. 4a confirm that the 3-D CNN-CONVLSTM architecture can accurately distinguish between static and moving frames. Furthermore, the overlaid ECG signal shows that there is a sufficient number of cardiac cycles before table translation, hence, these frames still correspond to a correctly known acquisition geometry and could be used directly for further processing, e.g., 3-D QCA.

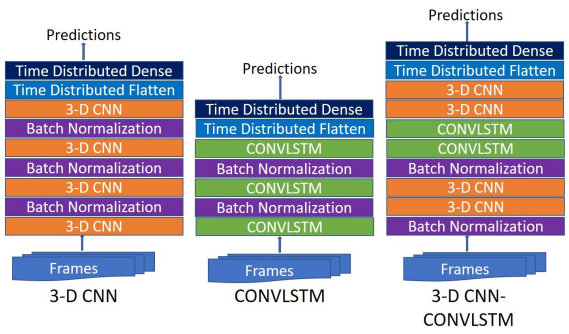
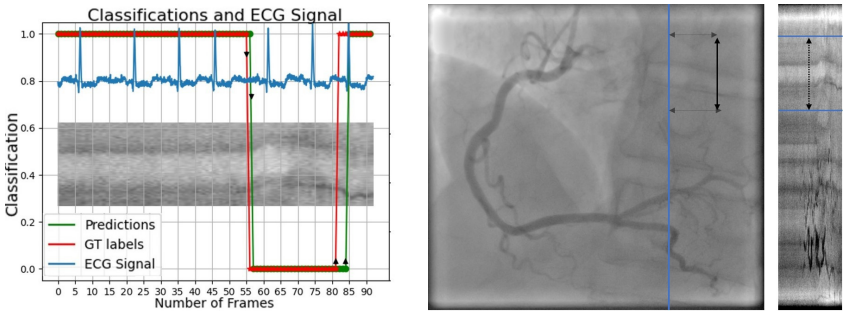


Fig. 3. Overview of network architectures.

Table 2. Overview of the performance of network architectures.

Frames	Precision	Recall	F1-score	Number of frames	Macro f1-score
3-D CNN					
Moving	0.74	0.64	0.69	969	0.8
Static	0.89	0.92	0.91	2968	
CONVLSTM					
Moving	0.85	0.71	0.78	969	0.86
Static	0.91	0.96	0.93	2968	
3-D CNN-CONVLSTM					
Moving	0.91	0.87	0.89	969	0.93
Static	0.96	0.97	0.97	2968	

Misclassifications were mainly due to the difference between the leading edge (static to moving) and the trailing edge (moving to static) of the predictions and the ground truth, shown as black arrows pointing up and down in Fig. 4a. This provides a quantitative measure for temporal accuracy. Fig. 5 shows the distributions of these temporal errors. Both the leading and trailing edges have almost similar temporal accuracy with a mean deviation of 1.37 and 1.61 frames, respectively. An Overlay of all data points on the box plots shows that the maximum error was 5 frames for both the leading and trailing edge.



(a) Classification result superimposed on the temporal movement of a disc.

(b) Temporal movement along a vertical line within an angiographic scene.

Fig. 4. Output of the 3-D CNN-CONVLSTM network for an example scene.

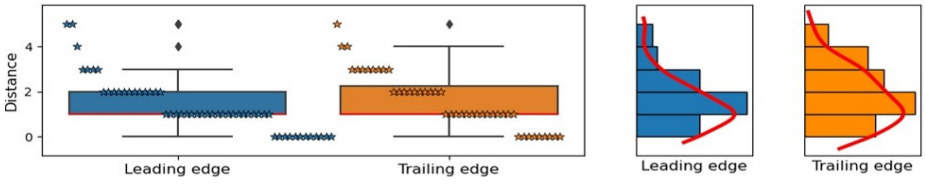


Fig. 5. Distance between the leading and trailing edges of ground truth and predictions.

4 Discussion

The 3-D CNN-CONVLSTM network successfully distinguishes table motion in the angiographic scene from heart and respiratory movements as it encodes longer temporal information. The CONVLSTM network also works reasonably well, but it usually incorrectly classified static frames as moving when there was a sudden change in successive frames, usually at the beginning of the contrast flow. The 3-D CNN network only worked well when there was no significant diaphragm motion and the effects of heart movement and heart shadow were minimal. The results show that the trained model can robustly classify table motion frames since all three clinical sites and RCA and LCA scenes yield similar f1-scores. The temporal error of the leading and trailing edge was less than 2 frames on average. This uncertainty can easily be taken into account when selecting suitable frames by adding an appropriate safety margin to the classification results. Currently, this study is limited to angiographic acquisitions with a framerate of 15 FPS. In future work, angiographic scenes will be up- or downsampled to a common framerate using motion interpolation techniques [9]. Furthermore, we aim to incorporate the information of the table motion predictor from this study to an extended network trained to estimate the optical flow [10]. This information could then be used to further reduce the effects of table motion when modeling the arterial tree.

References

1. Pantos I, Efstathopoulos EP, Katritsis DG. Two and three-dimensional quantitative coronary angiography. Cardiol Clin. 2009;27(3):491–502.
2. Çimen S, Gooya A, Grass M, et al. Reconstruction of coronary arteries from X-ray angiography. Med Image Anal. 2016;32:46–68.
3. Chen SJ, Hoffmann KR, Carroll JD. Three-dimensional reconstruction of coronary arterial tree based on biplane angiograms. Proc SPIE Med Imaging. 1996;2710:103–114.
4. Chen SJ, Carroll JD. 3-D reconstruction of coronary arterial tree to optimize angiographic visualization. IEEE Trans Med Imaging. 2000;19:308–336.
5. Yang J, Wang Y, Liu Y, et al. Novel approach for 3-D reconstruction of coronary arteries from two uncalibrated angiographic images. IEEE Trans Med Imaging. 2009;18:1563–1572.
6. Xiao X, Xu D, Wan W. Video recognition from handcrafted method to deep learning method. Proc ICALIP. 2016; p. 646–651.
7. Tran D, Bourdev L, Fergus R, et al. Learning Spatiotemporal Features with 3D Convolutional Networks. Proc IEEE ICCV. 2015; p. 4489–4497.
8. Shi X, Chen Z, Wang H, et al. A machine learning approach for precipitation nowcasting. Proc NIPS. 2015; p. 802–810.
9. Ce Liu. Beyond pixels: Exploring new representations and applications for motion analysis. USA: Massachusetts Institute of Technology; 2009.
10. Fischer P, Dosovitskiy A, Ilg E, et al. FlowNet: Learning optical flow with convolutional networks. Proc IEEE ICCV. 2015; p. 2758–2766.