Quality Assessment of Echocardiographic Cine Using Recurrent Neural Networks: Feasibility on Five Standard View Planes

Amir H. Abdi[1(✉)], Christina Luong[2], Teresa Tsang[2], John Jue[2], Ken Gin[2], Darwin Yeung[2], Dale Hawley[2], Robert Rohling[1], and Purang Abolmaesumi[1]

[1] Electrical and Computer Engineering Department, University of British Columbia, Vancouver, Canada
amirabdi@ece.ubc.ca
[2] Cardiology Lab, Vancouver General Hospital, Vancouver, Canada

Abstract. Echocardiography (echo) is a clinical imaging technique which is highly dependent on operator experience. We aim to reduce operator variability in data acquisition by automatically computing an echo quality score for real-time feedback. We achieve this with a deep neural network model, with convolutional layers to extract hierarchical features from the input echo cine and recurrent layers to leverage the sequential information in the echo cine loop. Using data from 509 separate patient studies, containing 2,450 echo cines across five standard echo imaging planes, we achieved a mean quality score accuracy of 85% compared to the gold-standard score assigned by experienced echosonographers. The proposed approach calculates the quality of a given 20 frame echo sequence within 10 ms, sufficient for real-time deployment.

Keywords: Convolutional · Recurrent Neural Network · LSTM · Deep learning · Quality assessment · Echocardiography · Echo cine loop

1 Introduction

Despite advances in medicine and technology, cardiovascular disease remains the leading cause of mortality worldwide. Cardiac ultrasound, better known as echocardiography (echo), is the standard method for screening, detection, and monitoring of cardiovascular disease. This noninvasive imaging modality is widely available, cost-effective, and is used for evaluation of cardiac structure and function. Standard echo studies include assessment of chamber size and function as well as valvular stenosis and competence. However, the clinician's ability to interpret an echo study highly depends on image quality, which is

C. Luong—Co-first author.

T. Tsang is the Director of the Vancouver General Hospital and University of British Columbia Echocardiography Laboratories, supervisor of the Cardiology Team, and Co-Principal Investigator of the CIHR-NSERC grant supporting this work.

M. Descoteaux et al. (Eds.): MICCAI 2017, Part III, LNCS 10435, pp. 302–310, 2017.
DOI: 10.1007/978-3-319-66179-7_35

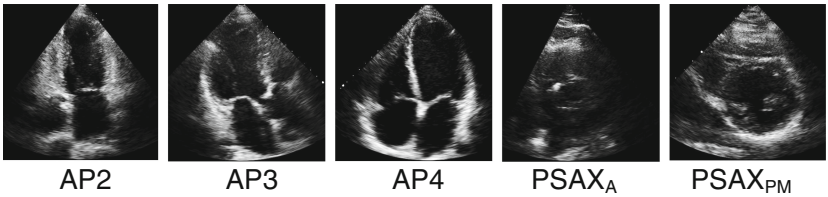


Fig. 1. The five standard echo view planes targeted in this study.

closely tied to the sonographer's skill and patient characteristics. Suboptimal images compromise interpretation and can adversely alter patient care.

Comprehensive evaluation with echo requires the acquisition of standardized views for 2D measurements and precise ultrasound transducer alignment. As ultrasound becomes increasingly available, less experienced clinicians are using this tool with potential hazards due to inconsistent image quality and limited expertise. Unlike other imaging modalities, ultrasound systems do not have automated image acquisition. The images obtained rely on scanner knowledge of the cardiac structures.

To improve the consistency of ultrasound image acquisition, efforts have been invested in detecting shadows and aperture obstructions [8,10] and optimizing the image acquisition parameters [5]. However, those methods are generic and not specific to echo acquisition. Quality of echo data is also dependent on optimizing the imaging plane to obtain sharp edges of desired anatomic structures for each standard view. View-specific quality assessment has been investigated through searching for a binary atlas using generalized Hough transform [11] or defining a goodness-of-fit to a parametric template model [13]. However, those techniques mainly rely on presence of sharp edges in the image, hence are likely to fail in low contrast settings, which is very common in clinical echo data.

In our previous work [1], we proposed an echo quality assessment approach using convolutional neural networks which focused only on the apical four-chamber view. However, the proposed method did not take advantage of the information available in sequential echo images (cine echo), and the assessment was limited to end-systolic frames.

In this work, we propose a deep learning model for quality assessment of echo cine loops across five standard imaging planes, based on analyzing the entire cine echo. The five standard view planes we analyze in this work are apical 2-chamber (AP2), apical 3-chamber (AP3), apical 4-chamber (AP4), parasternal short axis at the aortic valve level (PSAX$_A$), and parasternal short axis at the papillary muscle level (PSAX$_{PM}$) (Fig. 1). We designed a deep neural network with convolutional and recurrent layers and, a shared architecture to leverage transfer learning. This model automatically extracts hierarchical features from different echo view planes and relates them to a quality score determined by expert echocardiographers. In this research, we use data from 509 separate patients

studies, with a total of 2,450 echo cines. Using GPU-computing, the network is able to assess an echo cine loop and assign a quality score in real time.

2 Materials and Method

2.1 Dataset and Gold Standard Quality Assessment

To train the deep learning model, we accessed an echo database on the Picture Archiving and Communication System at Vancouver General Hospital. Different ultrasound machines from Philips and GE, and different image acquisition parameters contributed to the dataset. The majority of studies were performed by certified sonographers with a small proportion scanned by cardiology and sonography trainees. The dataset was randomly selected from the database and is therefore expected to contain a uniform distribution among easy and difficult patient cases. For each patient, 2D cine loops were available from standard views. In this paper, we focused on five standard 2D views, i.e. AP2, AP3, AP4, PLAX_A, and PLAX_{PM} (Fig. 1). These views provide comprehensive evaluation of chamber size, systolic function, and gross valvular function.

We used 2,450 cine loops from 509 echo studies with ethics approval of the Clinical Medical Research Ethics Board and consultation with the Information Privacy Office. The dataset was evaluated for image quality by one of two physicians trained in echocardiography. A semi-quantitative scoring system was defined for each view, modeled after a system proposed by Gaudet et al. [6], which is summarized in Table 1. The scores were obtained by semi-quantitative evaluation of component structures. Each component was assigned a quality score of up to 2 points that were summed to produce an overall view-specific image score, based on the following observations: 0 point) the structure was not imaged or was inadequate for assessment; 1 point) the structure was adequately viewed; 2 points) the view was optimized for the structure. Other components of the score included appropriate centering (1 point), correct depth setting (0.5 points), proper gain (0.5 points), and correct axis (1 point). Since the maximum possible score value was different for each view, the quality scores for all views were normalized to one. We refer to the normalized ground-truth values assigned by the trained echocardiographer as the Clinical Echo Score (CES).

Table 1. Summary of dataset and criteria for quality assessment. Note that each echo cine can contain multiple sequences of 20 consecutive frames.

View plane	#Cines	#Seqs	Criteria for clinical quality assessment	Score range
AP2	478	1131	Centering, depth, gain, LV, LA, MV	0–8
AP3	455	1081	Centering, depth, gain, AV, MV, LA, LV, septum	0–7
AP4	575	1270	Centering, depth, gain, LV, RV, LA, RA, MV, TV	0–10
PLAX_A	480	1148	Centering, depth, gain, AV and leaflets	0–4
PLAX_{PM}	462	1189	Centering, depth, gain, papillary muscles, axis	0–5
Total	**2450**	**5819**		

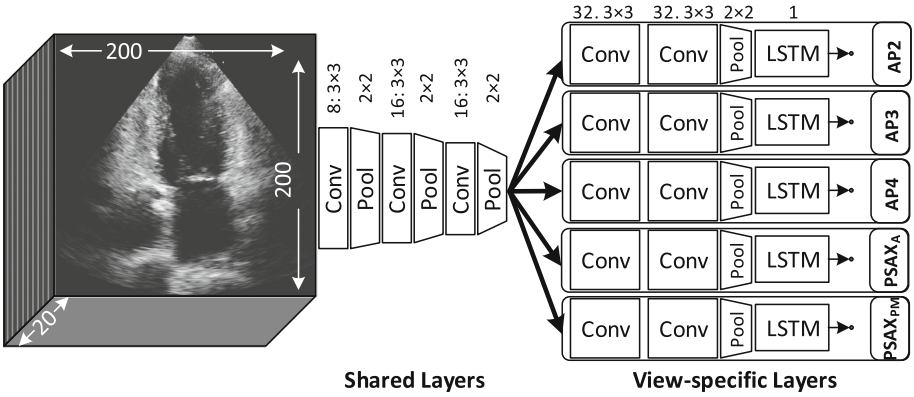


Fig. 2. The proposed multi-stream network architecture. Number of kernels in each layer and their corresponding sizes are presented above each layer.

2.2 Network Architecture

The proposed deep neural network is a regression model, consisting of convolutional (`conv`), pooling (`pool`), and Long Short Term Memory (`LSTM`) layers [4], and is simultaneously trained on the five echo view planes. The quality score estimate by the neural network is referred to as the Network Echo Score (NES).

The architecture, depicted in Fig. 2, represents a multi-stream network, i.e., five regression models that share weights across the first few layers. Each stream of the network has its own view-specific layers and is trained based on the *mean absolute error* loss function (ℓ_1 norm), via a stochastic gradient-based optimization algorithm.

All `conv` layers have kernels with the size of 3×3 following the VGG architecture [12], with the number of kernels doubling for deeper `conv` layers, i.e., from 8 to 32 kernels. The `conv` layers extract hierarchical features in the image, with the first three shared layers modeling high level spatial correlations, and the next two `conv` layers focusing on view-specific quality features. Activation function of the `conv` layers are Rectified Linear Units (ReLUs). In this design, all the `pool` layers are 2×2 max-pooling with a stride of 2 to select only superior invariant features and divide the input feature-map size to half in both dimensions to reduce feature variance and train more generalized models. The `conv` and `pool` layers are applied to each frame of the echo cine, independently. To prevent co-adaptation of features and over-fitting on the training data, a dropout layer with the dropout probability of 0.5 was used after the third pooling layer.

The feature map of the final `pool` layer is flattened and sent to an `LSTM` unit, a special flavor of Recurrent Neural Networks (RNN) that uses a gated technique to selectively add or remove information from the cell state [7]. A single `LSTM` cell analyzes 20 feature-sets corresponding to the 20 consecutive input frames, and only the last output of the sequence is used. The `LSTM` layer uses hard sigmoid functions for inner and output activations.

2.3 Training

We partitioned the data into two mutually exclusive sets of training-validation (80%) and test (20%). Network hyper-parameters were optimized by cross-validation on the training-validation set to ensure that the network can sufficiently learn the distribution of all echo views without over-fitting to the training data. After finalizing the network architecture, the network was trained on the entire training-validation set and its performance was reported on the test set.

Sequence Generation: Due to the variability in heart rate and frame acquisition rates the number of frames per cardiac cycle varied from study to study. We used a static sequence size of 20 frames, which encompasses nearly half the average cardiac cycle in our dataset. This duration was sufficient to capture the quality distribution of the echo imaging planes without adversely affecting the run-time of the model. As a result, frames in each echo sequence sample are not synced with the cardiac cycle, neither in the training-validation nor in the test data set. This design decision ensured that the estimated quality score for a given input sequence was independent of the starting phase of the cardiac data.

After partitioning studies into training-validation and test sets, each echo cine loop was split into as many consecutive sequences of 20 frames as possible, all of which were assigned the same quality-score as the original echo cine. As a result, the average number of training-validation sequences per echo view was 935 (4,675 in total), and the average number of test sequences per echo view was 228 (1,144 in total), each with equal length of 20 frames (Table 1).

Batch Selection: The five regression models were trained simultaneously and each batch consisted of eight sequences from each view. Each sequence was a set of 20 consecutive gray-scale frames, which were downsized to 200×200 pixels; no preprocessing was applied on the frames.

Since distribution of samples for each view was not uniform and the dataset held more mid to high quality images, a stratified batch selection strategy was implemented to prevent biases towards the quality-levels with the majority of samples [14]. For each view plane of each mini-batch, eight quality-levels were randomly selected and a sample, corresponding to the quality-level, was randomly fetched.

The above strategy benefited the training in two ways: (1) training samples did not follow a predefined order; (2) it guaranteed that, from the network's perspective, the training samples have a uniform distribution among quality-levels for all the five echo views.

Data Augmentation: Data augmentation was applied to achieve a more generalized model and to reduce the probability of over-fitting. To promote rotational invariance, each sequence of each batch was rotated, on-the-fly, with a random value uniformly drawn from the range $[-7, +7]$. A cardiologist confirmed that

this amount of rotation does not degrade the clinical quality of the cine. Translational invariance was achieved by shifting each cine of each batch in the horizontal and vertical directions with a random value uniformly drawn from the range $[-D/15, +D/15]$, where D is the width or height of the frame.

Training: The deep learning model was trained using the adam optimizer with the same hyper-parameters as suggested in the original research [9]. The weight of the `conv` layers were initialized randomly from a zero-mean Gaussian distribution. To prevent the deep network from over-fitting on the training data, ℓ_2 norm regularization was added to the weights of the `conv` kernels. Keras deep learning library with TensorFlow backend, was used to train and test the models [3].

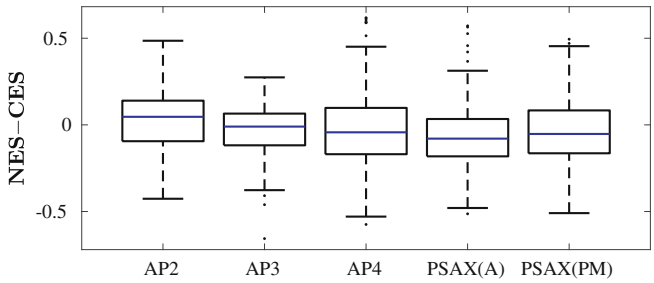
3 Experiments and Results

The error distribution for each echo view, calculated as $NES - CES$, is depicted in Fig. 3a. Figure 3b shows the average accuracy percentage calculated as

$$Acc_{view} = (1 - \sum_{i}^{T_{view}} |NES - CES|) \times 100, \tag{1}$$

where T_{view} is the total number of test sequences for the echo view. Performance of the model on the test data shows an average accuracy of 85% ± 12 against the expert scores. The accuracy of the trained models for the views are in the same order ranging from 83%–89%. Example test results are shown in Fig. 4.

By leveraging the GPU-based implementation of neural networks by TensorFlow, the trained model was able to estimate quality of an input echo cine with 20 frames of 200 × 200 pixels in 10 ms, suitable for real-time deployment.



(a)

View	Acc (%)
AP2	86 ± 9
AP3	89 ± 9
AP4	83 ± 14
$PSAX_A$	84 ± 12
$PSAX_{PM}$	83 ± 13
Total	85 ± 12

(b)

Fig. 3. (a) Distribution of error in each echo view. (b) Performance of the trained models for each view calculated via Eq. (1).

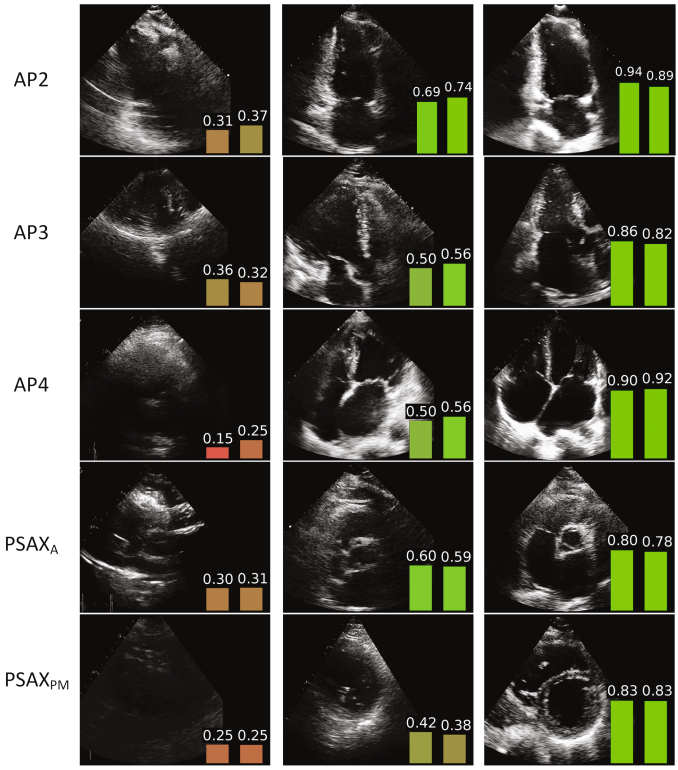


Fig. 4. Sample test results for the five standard echo imaging planes. The left bar in each sub-figure shows the gold-standard score by an expert echocardiographer (CES), and the right bar shows the estimated score by our approach (NES).

4 Discussion and Conclusion

Studies suggest that real-time feedback to sonographers can help optimize the image quality [13]. Here, we propose a deep learning framework to estimate the quality of a given echo cine and to provide feedback to the user in real time.

The results show that the trained model works with an acceptable 85% average accuracy across all the five targeted echo view planes (Fig. 3), which is superior to the performance of 82% which was achieved in our previous study on single end-systolic frames of the AP4 view [1]. More importantly, as demonstrated in Fig. 3, performance of the model is the same across all the views with the error distributed evenly around zero. As a result of the stratified batch-selection technique (Sect. 2.3), the model observes a uniformly distributed training-set, eliminating potential biases towards a quality-level with the majority of samples.

The five echo imaging planes were chosen based on their importance in echo studies. We did not provide any a priori information to the model regarding the visual perception of these views and the proposed method does not use view-specific templates [11,13]; hence, we expect that this approach can be easily extended towards other echo imaging planes. More importantly, this is the first study to leverage from the sequential information in echo cine to estimate the quality of the cine loop. Moreover, by designing a cross-domain architecture (Fig. 2), we leverage transfer learning to share the training sequences of each view with other views [2]. As a result, the proposed approach requires fewer training samples per echo view to achieve the same accuracy.

As the method does not rely on any pre-processing steps and takes advantage of GPU computing, the model can compute the quality of a 20 frame cine in real time. This is comparable to the speed achieved in our previous study [1] and faster than the Hough transform method suggested by Pavani *et al.* [11].

References

1. Abdi, A.H., et al.: Automatic quality assessment of apical four-chamber echocardiograms using deep convolutional neural networks. In: Proceedings of SPIE, vol. 10133, pp. 101330S–101330S-7 (2017)
2. Chen, H., Zheng, Y., Park, J.-H., Heng, P.-A., Zhou, S.K.: Iterative Multi-domain Regularized Deep Learning for Anatomical Structure Detection and Segmentation from Ultrasound Images. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9901, pp. 487–495. Springer, Cham (2016). doi:10.1007/978-3-319-46723-8_56
3. Chollet, F.: Keras (2015). https://github.com/fchollet/keras
4. Donahue, J., et al.: Long-term recurrent convolutional networks for visual recognition and description. IEEE Trans. Pattern Anal. Mach. Intell. **39**(4), 677–691 (2017)
5. El-Zehiry, N., Yan, M., Good, S., Fang, T., Zhou, S.K., Grady, L.: Learning the manifold of quality ultrasound acquisition. In: Mori, K., Sakuma, I., Sato, Y., Barillot, C., Navab, N. (eds.) MICCAI 2013. LNCS, vol. 8149, pp. 122–130. Springer, Heidelberg (2013). doi:10.1007/978-3-642-40811-3_16
6. Gaudet, J., et al.: Focused critical care echocardiography: development and evaluation of an image acquisition assessment tool. Crit. Care Med. **44**(6), e329–e335 (2016)
7. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
8. Huang, S.W., et al.: Detection and display of acoustic window for guiding and training cardiac ultrasound users. In: Progress in Biomedical Optics and Imaging - Proceedings of SPIE, vol. 9040, p. 904014 (2014)
9. Kingma, D.P., Ba, J.L.: Adam: a Method for Stochastic Optimization. In: International Conference on Learning Representations 2015, pp. 1–15 (2015)
10. Løvstakken, L., et al.: Real-time indication of acoustic window for phased-array transducers in ultrasound imaging. In: Proceedings of IEEE Ultrasonics Symposium, pp. 1549–1552 (2007)
11. Pavani, S.K., et al.: Quality metric for parasternal long axis B-mode echocardiograms. MICCAI **2015**, 478–485 (2012)

12. Simonyan, K., Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition. In: ICRL 2015, pp. 1–14 (2015)
13. Snare, S.R., et al.: Real-time scan assistant for echocardiography. IEEE Trans. Ultrason. Ferroelectr. Freq. Control **59**(3), 583–589 (2012)
14. Zhao, P., Zhang, T.: Accelerating minibatch stochastic gradient descent using stratified sampling. arXiv preprint arXiv:1405.3080, pp. 1–13 (2014)