Automatic Multi-View Pose Estimation in Focused Cardiac Ultrasound

João Freitas[1,2]
jpav.freitas@gmail.com
Jaime C. Fonseca[3]
jaime@dei.uminho.pt
Sandro Queirós[1,2]
sandroqueiros@med.uminho.pt

[1] Life and Health Sciences Research Institute (ICVS) School of Medicine, University of Minho
[2] ICVS/3B's - PT Government Associate Laboratory, Braga/Guimarães, Portugal
[3] Center Algoritmi, School of Engineering, University of Minho

Abstract

Focused cardiac ultrasound (FoCUS) emerges as a valuable on-the-spot technique for assessing cardiovascular structures and performance. Nevertheless, its applicability is curbed by equipment constraints and the proficiency of the operator, which leads to mostly qualitative evaluations. This study presents a novel framework that aims to automatically estimate the 3D spatial relationship between standard FoCUS views. The proposed framework uses a multi-view U-Net-like convolutional neural network to regress line-based heatmaps representing the most likely areas of intersection between input images. The lines that best fit the regressed heatmaps are then extracted, and a system of nonlinear equations is created to determine the relative 3D pose between all input views. The feasibility and accuracy of the proposed pipeline were validated using a novel realistic *in silico* FoCUS dataset, revealing auspicious outcomes that suggest its potential value within clinical contexts. This framework, by estimating the 3D pose, could help enabling comprehensive 3D quantitative assessments of FoCUS evaluations, enhancing diagnostic proficiencies, especially within urgent and high-care scenarios where swift and precise evaluations are paramount.

1 Introduction

Focused cardiac ultrasound (FoCUS) stands as a point-of-care imaging methodology leveraging ultrasonography for assessing both cardiac structure and function. In contrast to conventional echocardiography, which demands skilled sonographers for a comprehensive evaluation, FoCUS is often conducted by less seasoned clinicians with the intent of addressing specific queries directly at the patient's bedside. To do so, it requires the acquisition of fewer cardiac views, in addition to smaller ultrasound devices, offering enhanced portability at the cost of somewhat diminished image quality [1].

While recent research has proposed avenues to extract quantitative information from FoCUS assessments, the emphasis has predominantly revolved through two-dimensional (2D) images [2], a process profoundly reliant on the operator. While volumetric quantification could supply more precise and accurate indices [3], its implementation often requires three-dimensional (3D) imaging, an impracticality in routine FoCUS evaluations due to equipment and operator constraints.

In light of this, we tackle this challenge head-on by introducing an innovative deep learning (DL)-driven framework for the automatic estimation of the relative 3D pose across all acquired views. This strategy has the potential to surmount the primary hurdle limiting the integration of 3D cardiac image analysis methodologies into the FoCUS approach, thereby augmenting its diagnostic capabilities.

2 Methodology

The proposed pose estimation framework is built on the observation that the relative 3D pose of a set of input images can be determined if the intersection between them is known, provided that a sufficient number of images, from distinct views, are available. Hence, the framework is divided into three modules (Fig. 1): heatmap regression, line extraction and 3D view positioning.

2.1 Heatmap regression

The heatmap regression module aims to regress line-based heatmaps representing the most probable area of intersection between pairs of input images (as projected on both of them). Given the potential for shared knowledge among models trained to detect the intersection between distinct pairs of FoCUS views, we propose to employ a multi-encoder multi-decoder U-Net-like architecture that simultaneously estimates the heatmaps of all relevant views.

The proposed network receives five input images, each representing a typical view obtained in a FoCUS examination (apical 4- and 2-chambers, parasternal long- and short axis, and subxiphoid). It presents a multi-encoder architecture, allowing each input to have its dedicated contraction path. This enables the encoders to learn view-specific convolutional filters and gain a better understanding of the details of each view. Similarly, a multiple-decoder strategy is implemented, where the outputs are separated to allow each decoder to focus on a specific intersection between a pair of views. In total, the network produces 20 heatmaps, obtained from pairwise combinations of the intersections between the five input views. Each heatmap represents the network's confidence in the location of the intersection line between the respective pair of images. The proposed network was implemented using TensorFlow.

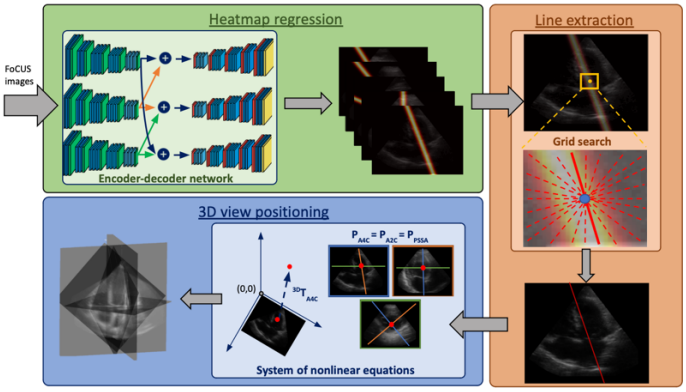

Fig 1 - Overview of the proposed multi-view pose estimation framework.

2.2 Line extraction

This module aims to extract the lines that best fit the predicted heatmaps. To achieve this, a coarse-to-fine grid search algorithm (based on [4]) was employed. The algorithm explores all possible position and orientation values to find the line with the highest score (computed by summing the heatmap values of pixels that fall within a perpendicular distance of ≤1 pixel from the evaluated line). To optimize the search and extraction process, five grid resolutions are used. Initially, a step of 15 is employed to determine the optimal line in the entire image. Subsequently, a refined search is performed in a window of size 30 (15 to each side) using a step of 5. This process is repeated for the last three resolutions (of step 1, 0.5 and 0.1), progressively increasing the accuracy of the search procedure.

2.3 3D view positioning

The main goal of this module is to estimate the relative 3D pose between all input views based on the extracted lines. Since the left ventricle is present in all considered views and no duplicated ones are included in the model, it is reasonable to assume that all images intersect with each other. It is also unlikely, although not impossible, for three images to not intersect at a single point, as collinearity only occurs between specific pairs (apical 4- and 2-chambers, or parasternal long- and -short axis) and no more than two views are acquired from a given cardiac window. These intersections among view triplets form a system of nonlinear equations that enables the estimation of the images' unknown poses.

Let us now consider a specific view triplet V consisting of images A, B and C. Through the previous modules, one has estimated two lines representing the intersection of any image i in V with the other two images. For any of the three images, the intersection between these two lines represent the point where the three images intersect, and therefore corresponds to the same point in 3D space. Hence, it is possible to express the equality between these points based on ${}^{3D}T_i$ as follows:

$$
\begin{aligned}
{}^{3D}T_A \cdot P_A &= {}^{3D}T_B \cdot P_B \\
{}^{3D}T_A \cdot P_A &= {}^{3D}T_C \cdot P_C \\
{}^{3D}T_B \cdot P_B &= {}^{3D}T_C \cdot P_C
\end{aligned}
$$

P_i represents the 2D point estimated for image i. However, assuming all unknown variables to be zero would result in a set of true equations that do not correspond to a valid solution. To address this, one of the input images must be set as reference by assigning its transformation matrix to the identity matrix. The pose of the remaining images will then be given with respect to it. In our experiments, the A4C view was chosen as the reference image. Additionally, the three triplets with the poorest performance on the estimation of the intersection lines were excluded to prevent a detrimental influence on the pose estimation accuracy.

3 Results and discussion

3.1 Dataset

A total of 7800 realistic synthetic FoCUS images [5] were used to implement and evaluate the proposed framework. The dataset was divided into two groups in a patient-disjoint manner: (1) a train/validation set (~80%) used for architecture design and hyperparameter tuning (in a 5-fold cross validation); and (2) a test set (~20%). It is worth nothing that after obtaining the fine-tuned architecture and optimal hyperparameters, the final model was trained using the complete train/validation set, allowing the network to be exposed to a wider range of diverse anatomical structures during the training phase.

3.2 Evaluation metrics

To assess the images' final positioning, we evaluate: the displacement error (d_{3D}), which was determined by calculating the 3D Euclidean distance between the image's centre point when positioned according to the predicted or ground truth pose; the rotational error (θ_{3D}), which quantifies the magnitude of rotation required to transform the predicted pose to the ground truth one through a single rotation.

Considering the absence of comparable studies in the literature, we established a baseline (Baseline-mid) to evaluate the results of the proposed framework. It consists in the expected error when using the 'average' intersection line and any other intersection line within pairs of random images from a given view pair. This 'average' line is obtained by intersecting the first image with the reference transducer's pose (which effectively represents the average pose of all simulated images for said view pair).

3.3 3D positioning

Table 1 provides an analysis of the accuracy of the predicted relative poses by comparing the median θ_{3D} and d_{3D} errors of the proposed approach with those obtained using the baseline. The results demonstrate that the proposed model outperforms the baseline, particularly in terms of the rotational component (approximately 75% and 67% reduction in θ_{3D} and d_{3D}, respectively).

Table 1 – Global performance of the proposed pose estimation framework.

		A2C	PSLA	PSSA	SX	Average
θ_{3D} (°)	Baseline-mid	50.60	58.04	60.30	74.32	60.02
	Proposed	**14.94**	**14.45**	**19.10**	**10.75**	**14.81**
d_{3D} (mm)	Baseline-mid	28.11	34.75	19.76	44.29	31.73
	Proposed	**9.84**	**12.47**	**10.10**	**9.52**	**10.48**

The best result per metric, both per view and averaged, is highlighted in bold. No results are provided for the A4C view since it is used as reference.

The results shown in Table 1 demonstrate the ability of the proposed pose estimation framework to assemble all views relative to one another, serving as a fundamental aspect of our approach. It not only allows the development and utilization of 3D image analysis algorithms in FoCUS, facilitating the extraction of the relevant chambers (*e.g.*, the left ventricle) and computation of the necessary indices in a three-dimensional manner, but also potentially addresses the impact of foreshortening by enabling the use of any acquired views not affected by it. These aspects provide a potential advantage over existing approaches that rely on a single view or on geometrical assumptions, and may result in the extraction of more accurate, reliable, and reproducible clinical indices.

Despite the framework's merits, this study presents some limitations. Firstly, while the use of synthetic images is a common practice in computer vision research, it may not fully capture the intricacies and variations present in real data. To address this concern, we conducted an exploratory study using real FoCUS images of a patient from our in-house dataset (depicted in Fig.2). Although ground truth data was unavailable to confirm the accuracy of the 3D view positioning, our framework demonstrated a remarkable ability to seamlessly position all views. This finding holds immense significance, as it shows the potential applicability of our framework in a clinical setting. However, these findings need to be confirmed using a large real FoCUS dataset that includes pose information (namely by using electromagnetic tracking technology to determine the transducer's pose). Secondly, while our proposed method yielded promising results, it is important to note that these findings were compared against theoretical baselines due to the absence of existing works addressing this specific topic. We acknowledge the limitations of these baselines and emphasize the importance of conducting further research to validate our approach against other benchmarks or by leveraging alternative datasets.

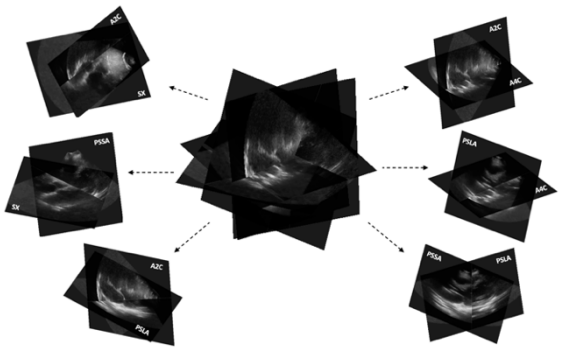

Fig 2 – Estimated relative 3D pose for a set of images from our in-house dataset, and highlight the intersection between six pairs of vires

4 Conclusion

In summary, we have introduced and validated a novel automatic framework for multi-view pose estimation in FoCUS. The experiments have demonstrated the accuracy and effectiveness of the proposed framework in estimating the relative pose of all input views. Ultimately, the proposed framework opens up new avenues for 3D analysis in FoCUS, showcasing its potential to enhance the diagnostic capabilities of this imaging modality in clinical practice.

Acknowledgements

This work was supported by National funds, through the Foundation for Science and Technology (FCT, Portugal), through projects UIDB/50026/2020, UIDP/50026/2020 and PTDC/EMD-EMD/1140/2020, and grant CEECIND/03064/2018 (S.Q.).

References

[1] Soliman-Aboumarie, H. *et al.*, 2021. How-to: Focus cardiac ultrasound in acute settings. European Heart Journal-Cardiovascular Imaging .

[2] Jafari, M.H. *et al.*, 2019. Automatic biplane left ventricular ejection fraction estimation with mobile point-of-care ultrasound using multi-task learning and adversarial training. International Journal of Computer Assisted Radiology and Surgery 14, 1027–1037.

[3] Lang, R.M *et al.*, 2015. Recommendations for cardiac chamber quantification by echocardiography in adults: an update from the American Society of Echocardiography and the European Association of Cardiovascular Imaging. European Heart Journal-Cardiovascular Imaging 16, 233–271.

[4] Wei, D., Ma, K., Zheng, Y., 2021. Training automatic view planner for cardiac MR imaging via self-supervision by spatial relationship between views, in: International Conference on Medical Image Computing and Computer- Assisted Intervention, Springer. pp. 526–536.

[5] Freitas, J. *et al.*, 2023. Automatic Generation of Multi-View Synthetic Echocardiographic Images in 18th International Symposium on Computer Methods in Biomechanics and Biomedical Engineering: Abstract book Oral presentations.