Force-Ultrasound Fusion: Bringing Spine Robotic-US to the Next "Level"

Maria Tirindelli, Maria Victorova, Javier Esteban, Seong Tae Kim, David Navarro-Alarcon, Yong Ping Zheng, and Nassir Navab

***Abstract*—Spine injections are commonly performed in several clinical procedures. The localization of the target vertebral level (i.e. the position of a vertebra in a spine) is typically done by back palpation or under X-ray guidance, yielding either higher chances of procedure failure or exposure to ionizing radiation. Preliminary studies have been conducted in the literature, suggesting that ultrasound imaging may be a precise and safe alternative to X-ray for spine level detection. However, ultrasound data are noisy and complicated to interpret. In this study, a robotic-ultrasound approach for automatic vertebral level detection is introduced. The method relies on the fusion of ultrasound and force data, thus providing both "tactile" and visual feedback during the procedure, which results in higher performances in presence of data corruption. A robotic arm automatically scans the volunteer's back along the spine by using force-ultrasound data to locate vertebral levels. The occurrences of vertebral levels are visible on the force trace as peaks, which are enhanced by properly controlling the force applied by the robot on the patient back. Ultrasound data are processed with a Deep Learning method to extract a 1D signal modelling the probabilities of having a vertebra at each location along the spine. Processed force and ultrasound data are fused using both a non deep learning method and a Temporal Convolutional Network to compute the locations of the vertebral levels. The benefits of fusing force and image signals for the identification of vertebrae locations are showcased through extensive evaluation.**

***Index Terms*—Medical robots and systems, computer vision for medical robotics.**

I. Introduction

LUMBAR spinal injections are commonly performed in different clinical procedures as facet joint or epidural injections [1], [2]. Such procedures typically require the correct localization of the target vertebra to effectively release pharmaceuticals. In clinical practice, vertebral level detection is achieved either through palpation or X-ray guidance. Although X-ray guidance can improve the overall precision of the procedure, the use of ionizing radiation is considered a hazard for the patient and especially for the clinicians and assistants. On the other hand, the accuracy of the palpation technique is lower, especially for less experienced clinicians. Furthermore, the incorrect chosen level of injection can lead to avoidable complications, such as headaches, nerve damage, and paralysis [3].

Ultrasound (US) has proven to be an alternative to X-ray, providing precise guidance and preventing patients and clinicians from unnecessary radiation [4]. Despite being real-time and non-invasive, ultrasound guidance is particularly challenging in spine procedures due to artifacts and noise caused by the curvature of the spinal bones and the layer of soft tissue covering the spine. To address these issues, various authors have proposed to use image processing techniques to support the clinician in the detection of vertebral levels.

In [5] a method is proposed to automatically classify images acquired during manual ultrasound-guided epidural injections. In this work, a Convolutional Neural Network (CNN) is used to classify the acquired images as either "vertebra" or "intervertebral gap" and State Machine is implemented to refine the results. In [6] and [7] panorama image stitching is used to obtain a 2-Dimentional (2D) representation of vertebral laminas along the spine in the paramedian-sagittal plane. In [6] a set of filters are applied to the panorama image to enhance bony structures. Local minimums in the resulting pattern are extracted and labelled as vertebrae. In [7] the identification of vertebrae is performed on the panorama image using a template matching approach.

The aforementioned methods provide support tools for the interpretation of ultrasound data during manual injection procedures. However, they still rely on the operator's skills to manually find correspondence between ultrasound images and patient anatomy. Few studies have been conducted to evaluate the potential of robots integration in the clinical environment for injection procedures. In [8], a robotic-ultrasound system for precise needle placement is described in an initial clinical study. In this study, a robotic system with a calibrated ultrasound probe is used to scan the patient back. The acquired US volume is then used by the operator to select the needle insertion path. The manipulator, equipped with a calibrated needle holder, moves to the desired insertion point to offer visual guidance during the insertion. Although showing promising results, these systems still rely on the operator in the interpretation of ultrasound

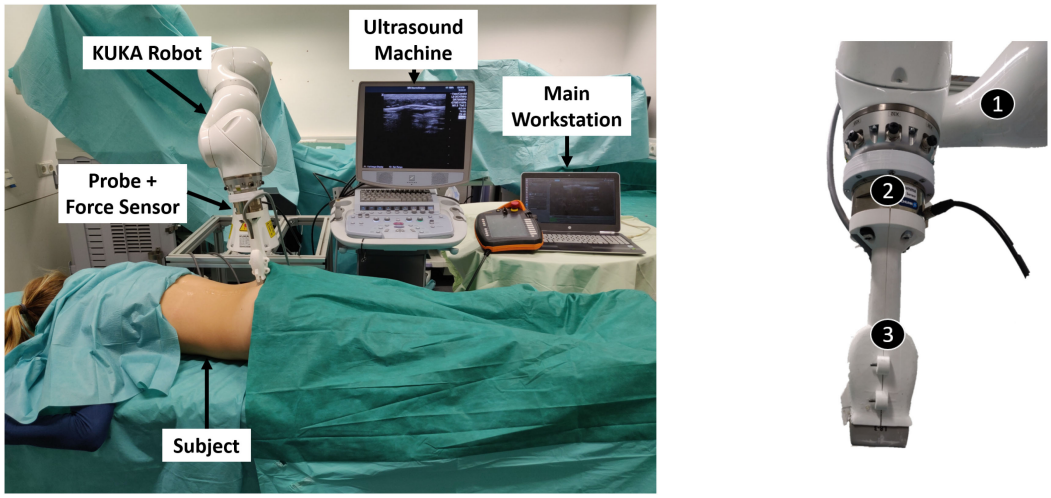
Manuscript received February 23, 2020; accepted June 4, 2020. Date of publication July 14, 2020; date of current version July 28, 2020. This letter was recommended for publication by Associate Editor E. De Momi and Editor P. Valdastri upon evaluation of the Reviewers' comments. This work was supported by the Bayerische Forschungsstiftung, under Grant DOK-180-19. *(Maria Tirindelli and Maria Victorova contributed equally to this work.) (Corresponding author: Maria Tirindelli.)*

Maria Tirindelli, Javier Esteban, and Seong Tae Kim are with Computer Aided Medical Procedures, Technische Universität München, Munich 80333, Germany (e-mail: maria.tirindelli@mail.polimi.it; javier.esteban@tum.de; seongtae.kim@tum.de).

Maria Victorova, David Navarro-Alarcon, and Yong Ping Zheng are with The Hong Kong Polytechnic University, Hung Hom, Hong Kong (e-mail: maryviktory@gmail.com; davidnavarroalarcon@gmail.com; yongping.zheng@polyu.edu.hk).

Nassir Navab is with the Computer Aided Medical Procedures, Technische Universität München, Munich 80333, Germany and also with the, Computer Aided Medical Procedures Johns Hopkins University, Baltimore, MD 21218 USA (e-mail: navab@cs.tum.edu).

Digital Object Identifier 10.1109/LRA.2020.3009069



(a) (b)

Fig. 1. (a) The Robotic Ultrasound System for Vertebral Level Classification Setup. (b) The Robot end-effector configuration: 1 - Robotic arm, 2 - External force sensor, 3 - Ultrasound linear probe + 3D printed probe holder.

images. Furthermore, they do not provide any tactile feedback, which, for the standard procedure, is given by palpation.

The contribution of this work is a robotic-ultrasound approach combining force and ultrasound data for automatic lumbar vertebral level classification in the spine. The target spinal region is the lumbar region (i.e. vertebrae levels from L5 to L1), where spinal injections commonly take place. Force feedback reproduces the tactile information the operator can get through palpation while ultrasound images provide continuous visual feedback during the procedure. Compared to the previously presented methods for vertebrae level classification, the proposed approach combines the benefits of both robotics and standard procedures. Furthermore, it does not only rely on visual feedback, but it exploits multiple sensors information. It is demonstrated that fusing ultrasound and force data ensures higher performances in the presence of data corruption and single-sensor misclassifications. The potential of the proposed approach is explored for an example application, i.e. automatic target plane detection for facet injection procedures.

II. Methods

A. Materials and Experimental Setup

The robotic ultrasound system setup is presented in Fig. 1(a). The system consists of a main workstation (Intel i7, GeForce GTX 1050 Mobile), a robotic arm certified for human interaction (KUKA LBR iiwa 7 R800) combined with a Six-Axis Force/Torque Sensor System FTD-GAMMA (SCHUNK GmbH & Co. KG) and a Zonare z.one ultra sp Convertible Ultrasound System with an L8-3 linear probe, with purely linear and steered trapezoidal imaging (Fig. 1(b)). The ultrasound system is connected to the main workstation through an Epiphan DVI2USB 3.0 frame-grabber (Epiphan Systems Inc. Palo Alto, California, USA), with an 800x600 resolution and a sampling frequency of 30 fps.

Deep Learning models were trained on an NVIDIA Titan V 12 GB HBM2, using Pythorch 1.1.0 as Deep Learning framework for both training and inference. ImFusion Suite Version 2.9.4 (ImFusion GmbH, Munich, Germany) is used for basic image processing and visualization.

Three different datasets were used for training of Deep Learning models and testing. The datasets were acquired for different subjects with different ultrasound (constant Gain = 92%, Frequency = 14 Hz), robot force and speed settings. The acquisition was performed in the lumbar region, from L5 to L1. The Body Mass Index (BMI) of the scanned subject is in the range 20–30 for all the 3 databases. The dataset size and acquisition parameters are reported for the three datasets in Table I. Two ultrasound experts manually labeled ultrasound data independently. All the ultrasound sweeps where the number of labeled vertebral levels did not coincide between the annotators were discarded.

In Fig. 2(a) flowchart of the method is shown. In Section II-(b), II-(c), II-D and II-E a detailed description of each pipeline step is provided.

B. Scanning Procedure

Before starting the procedure, the robotic arm is manually placed at the base of the sacrum with a transverse probe orientation. After probe placement, the robot starts moving in the caudo-cranial direction towards the subject's head, while force and ultrasound data are simultaneously collected (Fig. 3(a), black arrow). The subjects are asked to hold their breath for the whole duration of the scan (around 10 sec.), which is comparable to the breath-hold time of standard imaging procedures, as abdominal MRI or PET/CT [9], [10]. Once the scan is completed, the collected data are processed, to provide the location of the vertebral level at which the injection must be performed (Fig. 3(a), red cross).

Depending on the clinical application, further data can be acquired of the target vertebral level, to identify specific anatomical features. For the explored example application (i.e. automatic facet plane detection for facet joint injections), a further scan of the target vertebral level is performed in the

TABLE I
DATASET TABLE WITH CORRESPONDENT SIZE, DATA AND SENSOR SETTINGS

Dataset	**N. Subjects**	**Acquired Data**	**Probe Orientation**	**US Parameters**	**Applied Force** [N]	**Robot Speed** [mm/s]
Dataset 1	19	B-Mode Linear US	Transverse	Depth = 4cm	2	20
Dataset 2	14	B-Mode Linear US Force Data	Transverse	Depth = 4cm	[2, 10, 15]	[12, 20, 40]
Dataset 3	19	B-Mode Convex US	Paramedian-Sagittal	Depth = 7cm	2	5

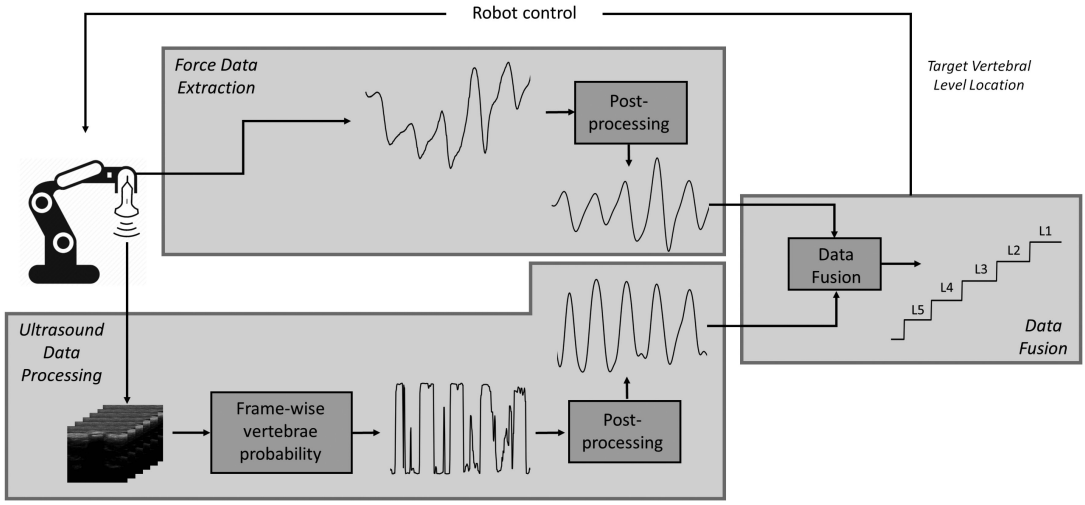


Fig. 2. The method pipeline, including force data extraction, ultrasound data processing and the fusion method.

latero-lateral direction with the probe in paramedian-sagittal orientation (Fig. 3(a)). The facet plane is identified within the scan so that the robot can then move to the plane position.

C. Force Data Extraction

In Fig. 4(a), a model of the vertebra-robot interaction is provided. In absence of vertebrae, the robot moves on a surface (the patient back) which can be considered flat. The reaction force is directed along the z-axis and its modulus balances the force applied by the robot, which is constant and set prior to the acquisition (Point A). In correspondence to a vertebra, the local direction of the subject back changes yielding to the generation of a non-null y-axis component of the reaction force (point B).

Once the vertebral peak has been reached (point C), the inclination of the plane changes again (point D) leading to the generation of non-null y-component of the reaction force, with an opposite sign with respect to point B. When the original surface direction is recovered, the y-component of the reaction force vanishes and the initial force value is recovered. The variations in the force y-component due to reaction forces are recorded by the force sensor and result in a very characteristic pattern in the force trace (Fig. 4(b)). This pattern can be used to count the vertebral levels while the patient back is scanned. In Fig. 4(b), a plot of the y-component of the force signal is provided, in relation to the points A, B and C.

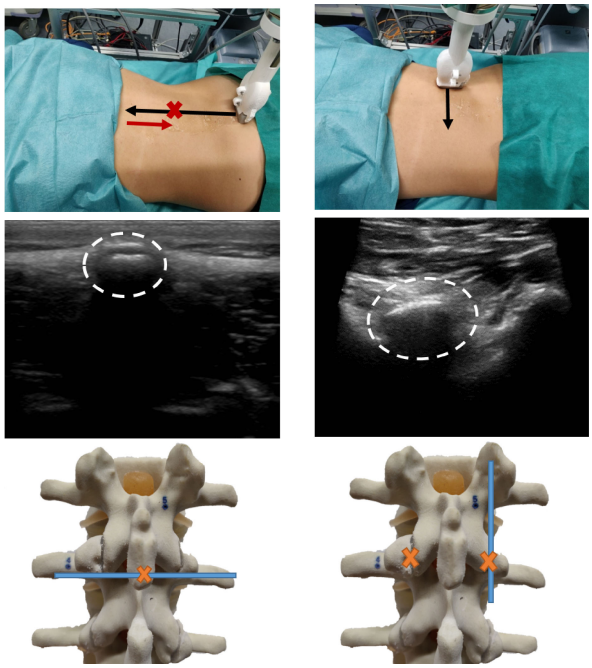
The recorded force in the y-direction (F_y) is pre-processed to remove the low-frequency drift, appearing due to the robot initial and final acceleration/deceleration. Drift removal is done by subtracting from the original signal its filtered version obtained applying a second-order Butterworth filter with cutoff frequency at 0.05 Hz. The "un-drifted" signal is then low pass-filtered with a second-order Butterworth filter with cutoff frequency at 0.3 Hz, normalized between 0 and 1 and re-sampled in equally spaced space-grid.

As mentioned above, the force applied by the robot along the z-direction ($F_{Ro,z}$) is constant and manually set before the acquisition takes place. The robot complies to the Force Control Scheme as described in [11].

Control over z-axis is also designed to compensate residual breathing motions, which main component is along the z-axis.

The value of the force z-component has a notable impact on the quality of the force signal recorded along the y-axis (F_y) and on the visibility of vertebral patterns. In particular, higher values of $F_{Ro,z}$ lead to more visible and defined vertebral spikes. However, high values of $F_{Ro,z}$ also result in less comfort for the subjects, especially for those with a thin muscle/fat layer. In this study, the quality of the force signal recorded along the spine direction is evaluated for three different values of $F_{Ro,z}$ on a group of 14 subjects with BMI ranging from 20 to 30 (Dataset 2). The selected force values are comparable to those which are used in clinical experimentation [8]. Each subject was asked to report the comfort level of the procedure on a scale ranging from 1 to 4, designed in the following way: 1 - very uncomfortable, 2 - uncomfortable, 3 - slightly uncomfortable, 4 - comfortable.

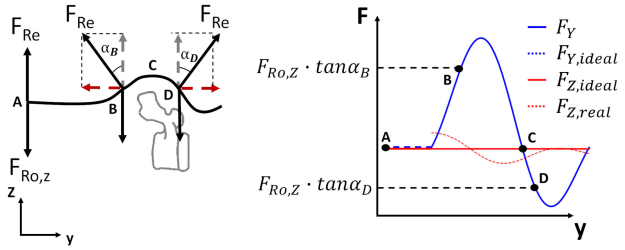
For none of the subject, the procedure resulted to be "very uncomfortable" or "uncomfortable". However, subjects with lower BMI tended to rate the procedure performed with $F_z = 15\ N$ as slightly uncomfortable.



(a) Data acquisition with a probe in transverse orientation and the respective ultrasound image of the spinous process.

(b) Data acquisition with the probe in paramedian sagittal orientation and the respective ultrasound image of the facet joint.

Fig. 3. Robot Trajectory during the procedure (arrows), target anatomies (dash line) and corresponding ultrasound images of acquired anatomies with planes of scanning (blue line).



(a) (b)

Fig. 4. (a) The modelled interaction between robot and patient back during the robotic scanning procedure. (b) Z component (red) and Y component (blue) of the force signal recorded over a single vertebra.

For this reason, the force applied by the robot along the z-axis is set to 10 N for subjects with lower BMI ($BMI < 23$) and to 15 N for subjects with higher BMI ($BMI > 23$). In Fig. 5, the force signals are reported for 3 different values of F_z (i.e. 2 N, 10 N, 15 N) for two subjects with different BMI. For both subjects, the amplitude of the spikes in the force trace increases with increasing force. However, for the subject with lower BMI, the spikes are still clearly recognizable in the signals obtained with lower pressures along the z-direction.

D. Ultrasound Data Processing

The informative component of the force signal (along y-axis F_y is a 1D signal providing spatial information about the spine anatomy along the spine direction. However, ultrasound data

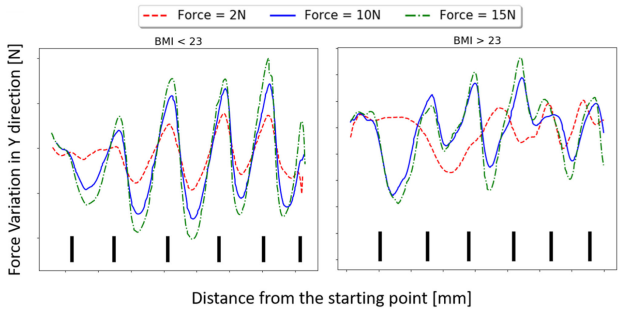


Fig. 5. The force signal recorded in the y-axis with 3 different values (2, 10 and 15 N) of the z-force applied by the robot for subjects with $BMI < 23$ (left) and for subjects with $BMI > 23$ (right).

are 3D data, where each position along the spine corresponds to a 2D (B-mode) ultrasound frame. To be able to effectively compare the information from the two sensors, ultrasound data are reduced to a 1D vector, defined along the spine direction. The dimension reduction is achieved by analyzing each ultrasound frame in the acquired sweeps and defining the probability for each of them to contain a vertebra.

Therefore the problem is designed as a binary classification problem in which the network learns to classify each frame along the spine as either "vertebra" or "intervertebral gap". The concatenation of the resulting values along the spine direction is a 1D signal where high probability peaks ideally coincide with vertebrae and therefore corresponds to peaks in the force signal.

The vertebra probability value is extracted from each frame using a Convolutional Neural Network trained for the task of classification. In order to ensure the best classification results, three state of the art classification networks were tested and compared: ResNet18 [12], DenseNet121 [13], VGG11 with batch normalization [14]. The training and validation performances were evaluated for all the architectures in the following cases: a) Training the network with randomly initialized weights; b) Using ImageNet [15] weights as initialization (pre-trained network) and fine-tuning all layers; c) Using ImageNet weights as initialization and fine-tuning the last layer only. Each model was trained using Adam optimizer, Cross-entropy loss function, learning rate of 0.0005 and a learning rate decay of 0.1 every 5 epochs for 30 epochs. The data for CNN training and testing were sampled from the Dataset 1.

Labels are represented as boolean values, where 1 corresponds to "vertebra" and 0 to "intervertebral gap". The training dataset consisted of 15 subjects (12 for training and 3 for validation), for a total of 1986 images for each class to ensure class balance. The test set consisted of 4 subjects, for a total of 696 images for each class. A 5-fold cross-validation study was performed over the training and validation datasets to exclude false-positive results. The obtained 1D signal is smoothed using a second-order Butterworth filter with cutoff frequency at 0.3 Hz and re-sampled in equally spaced space-grid.

E. Force - Ultrasound Data Fusion

The extracted and pre-processed force and ultrasound 1D signals represent variations of the inner/outer spine anatomy along the spine direction. In optimal conditions, both signals

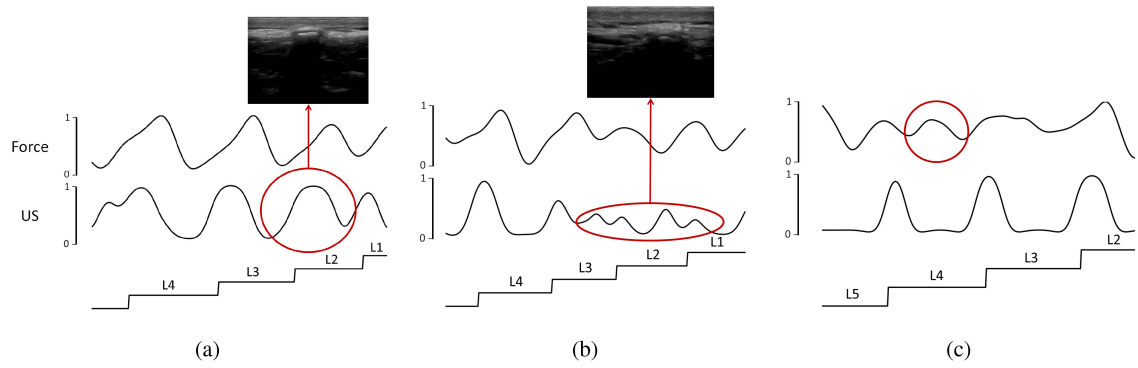


Fig. 6. (a) Force signal, Ultrasound signal and labels in the presence of non-corrupted force and ultrasound data. (b) Force signal, Ultrasound signal and labels in the presence of noisy ultrasound data. (c) Force signal, Ultrasound signal and labels in presence of noisy force signal.

present well visible peaks in correspondence with vertebral levels (Fig. 6(a)). However, in some cases one (or both) signals may be corrupted by noise, making it challenging to identify the real position of the vertebral levels. Noise in the signal extracted from the ultrasound data typically arises from the scarce visibility of the spinous process in the ultrasound sweep (Fig. 6(b)). This can be related to several factors as device-specific noise, non-optimal couplings between the probe and the patient skin or subject-specific anatomy and tissue distribution. Noise in the force signal may arise from sudden movements of the subject during the acquisition, or from subject-specific anatomical features (e.g. vertebral peaks may be less evident in particularly muscular subjects) (Fig. 6(c)). Labels for ground truth images were generated by manually labeling each image as one of the classes (L1 to L5). To make the method more robust against single-sensor misclassifications, a force-ultrasound fusion method was implemented. In particular, a Temporal Convolutional Network (TCN) was trained to classify vertebral levels from the input signals. The vertebral level counting problem is modelled as a classification problem, where the network is trained to classify each vertebral level in the lumbar region.

A multi-stage temporal convolutional network is devised based on [16], where the overall architecture consists of three stages and each stage is trained to classify the input data. Each stage refines the results from previous stages, yielding smoother and more accurate classification results. Each stage consists of an initial 1x1 convolution layer which re-sizes the input into a 32 x N sequence, where N is the original signal length (number of samples along the spine direction). The initial layer is followed by 9 1xD dilated convolution layers with kernel size 3 and increasing dilation size (Fig. 7). Dilated convolution is defined as:

$$(F *_l k)_t = \sum_{s+lt=p} F(S)k(t) \tag{1}$$

where F is the input signal, k is the filter kernel and l is the dilation factor. It can be seen from the formula that, compared to standard convolution, the result at each point of the convoluted signal is obtained considering a larger spatial field in the input signal, therefore allowing the network to exploit a broader spatial context for the input's classification. A softmax layer is added after the last convolution layer, to retrieve class probabilities

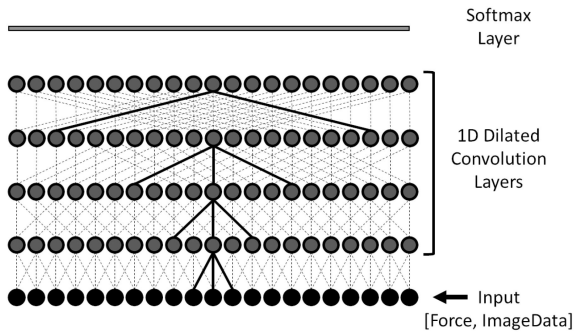


Fig. 7. The architecture of the single stage of the 1D convolutional network for data fusion.

(Fig. 7). The cross-entropy and an additional smoothing factor are used as the loss function for network training, as described in [16]. The convolutional network for force and ultrasound fusion was trained using Adam optimizer, learning rate of 0.0005 and batch size 1 for 110 epochs. The data for network training and testing were sampled from Dataset 2. The training dataset consisted of multiple sweeps acquired over 9 subjects (7 for training and 2 for validation) sampled from Dataset 2, for a total of 27 sequences for training and 7 for validation. The test set consists of 4 unseen subjects, acquired with the optimal robot parameters (force equal to 10 N or 15 N depending on subject BMI and robot speed equals to 20 mm/s). The test subjects were excluded from network training and validation, to avoid positive results related to network overfitting on the training dataset. No data acquired on any of the test subjects were used for model training at any step of the pipeline, ensuring the test set was unseen at test time as well as test subjects anatomy.

III. Results and Discussion

1) Ultrasound Data Processing: In Table II the test accuracy is reported for each CNN architecture (ResNet18, DenseNet121, VGG11) for the 3 training cases (a - training entire network with randomly initialized weights; b - using a pre-trained network with ImageNet weights as initialization and fine-tune all layers; c - training only the last layer of the network). The best accuracy on the test set is obtained by fine-tuning all the layers

TABLE II
RESULTS OF 5-FOLDS CROSS-VALIDATION STUDY FOR VARIOUS MODELS WITH DIFFERENT TRAINING MODES

	ResNet18	DenseNet121	VGG11
Case a	0.817 ± 0.118	0.878 ± 0.047	0.635 ± 0.15
Case b	$\mathbf{0.929 \pm 0.006}$	0.89 ± 0.014	0.878 ± 0.055
Case c	0.6 ± 0.02	0.577 ± 0.006	0.63 ± 0.03

TABLE III
CONFUSION MATRIX FOR THE BEST MODEL PERFORMANCE EVALUATED ON THE TEST SET OF 4 SUBJECTS

		Predicted	
	n = 1392	Vertebra	Intervertebral Gap
Actual	Vetrebra	True Positive 0.459 (n = 640)	False Negative 0.04 (n = 56)
	Intervertebral Gap	False Positive 0.02 (n = 30)	True Negative 0.478 (n = 666)

of ResNet18 from the pre-trained model, providing an average accuracy of 0.929 ± 0.006.

The ResNet18 model with the best performance was tested on a testing database of 4 subjects, yielding an overall accuracy of 0.938. The confusion matrix computed on the test data is displayed in Table III. The values are normalized by the total number of frames, the number of images n = 1392, the correspondent number of frames is shown in the parenthesis.

2) Force-Ultrasound Data Fusion: The performances of the force-ultrasound data fusion method were evaluated in terms of its capability to correctly label each vertebral level. The test group consists of 5 (unseen) subjects, for a total of 25 vertebral levels.

The TCN results were evaluated in the following conditions: using the pure force, the pure image and both force and image signals as input. The performance of the TCN for vertebral level counting was compared with a conventional peak detector (CPD). For the peak detector, the parameters were empirically chosen on the training set and were constant across all of the experiments (amplitude threshold 0.5, the minimum distance between peaks - 10 samples). The fusion input signal for the peak detector was obtained as the sum of force and ultrasound input signals.

Table IV reports the counting results in terms of detection accuracy and distance from the ground truth for all methods. A vertebral level classification is here considered successful if an overlap higher than 0.5 exists between labels and predictions, similarly as in [5]. Given the small training data set, the results of the 5-fold cross-validation study for all TCN methods are reported in Table IV. The test subjects were divided into three groups according to their height: *i)* below average ($<$ 163 cm), *ii)* average (163 cm $< - <$ 183 cm), *iii)* above average ($>$ 183 cm) The threshold values were chosen according to the training set height distribution (173 ± 10 cm). It can be noticed that the TCN methods outperform the peak detector for all of the input signals for subjects with average height, and outperforms the peak detector method when using pure force and pure image as input signals, for all of the height categories. This superiority could be attributed to the capability of the TCN

TABLE IV
THE CLASSIFICATION PERFORMANCES AND DISTANCE FROM THE GROUND TRUTH VERTEBRAE POSITION FOR ALL TESTED METHODS. FOR THE TCN METHODS THE RESULTS ARE REPORTED AS MEAN (STD) FOR THE 5-FOLD CROSS VALIDATION

		Correctly Classified Levels [num/total]				Distance from Ground Truth Label [mm]			
		Below	Average	Above	Overall	Below	Average	Above	Overall
CPD	Image	0.4	0.73	1.0	0.72	27.359 (26.9)	9.85 (14.0)	3.079 (1.96)	9.74 (15.95)
	Force	0.2	0	0	0.04	20.01 (9.779)	37.58 (7.10)	30.7 (3.70)	32.09 (10.3)
	Fusion	**1.0**	0.933	**1.0**	**0.96**	**2.495 (3.2)**	**2.357 (1.8)**	**2.386 (2.196)**	**2.39 (2.23)**
TCN	Image	0.48 (0.09)	**1.0** (0.0)	0.68 (0.097)	0.832 (0.03)	10.93 (0.90)	3.7 (1.17)	8.224 (1.20)	6.05 (0.735)
	Force	0.439 (0.079)	0.92 (0.06)	0.72 (0.16)	0.784 (0.04)	14.74 (2.8)	6.18 (1.79)	8.88 (1.46)	8.43 (1.02)
	Fusion	0.439 (0.149)	**1.0** (0.0)	0.6 (0.0)	0.808 (0.03)	12.64 (1.6)	3.76 (0.99)	8.72 (0.70)	6.52 (0.5)

to learn an anatomical prior based on the training data and hence compensate for missing or corrupted peaks in the input signals. This characteristic of TCN is relevant when using the pure force signal as input, where the first vertebral peak is often non well visible, due to both L5 anatomy (on average smaller and less prominent than the other lumbar vertebrae) and to the noise introduced by the acceleration of the robot at movement initiation. This leads to a shift in the vertebral level classification when using a peak detector on the force signal (Fig. 8).

When using a TCN, the pure image and fusion methods perform similarly. The slight difference between these methods could be attributed to the labeling process, which is performed on ultrasound data and not on the force signal. During labeling all the cases where the spinous processes were not visible were discarded because they did not pass the double-blind labeling process. This approach introduces a bias toward the pure-image based TCN, which can, therefore, be considered in this case the upper bound in terms of accuracy for TCN-based vertebral level classification.

The peak detector applied to the fusion signal outperforms the TCN fusion method for subjects above and below the average height. This can be explained by the limited size and variability of the training set (9 subjects). Nevertheless, the capacity of the TCN can be further improved with the acquisition of a more extensive data set with higher variability in terms of spine anatomy. Furthermore, compared to the peak detector method, the TCN does not rely on empirical parameter setting. In terms of the distance between ground truth and detected vertebral levels, the pure image and fusion methods perform similarly and marginally better than the pure force method. Compared to the peak detector, the distance errors of the TCN are more consistent, according to the reported standard deviation.

In Fig. 8 the results for all the analyzed methods are reported for one of the test subjects (Height: 186 cm). It can be seen that the peak detector on the pure image signal is only able to detect L5 and L4. As for most of the subjects, the peak detector on the pure force signal cannot correctly detect L5, which leads to an offset in the classification of all the remaining vertebrae. This offset is corrected by the TCN, which can correctly classify the first vertebral level. The peak detector on the fusion signal also successfully integrate the information from the two signals and correctly classifies all the vertebral levels.

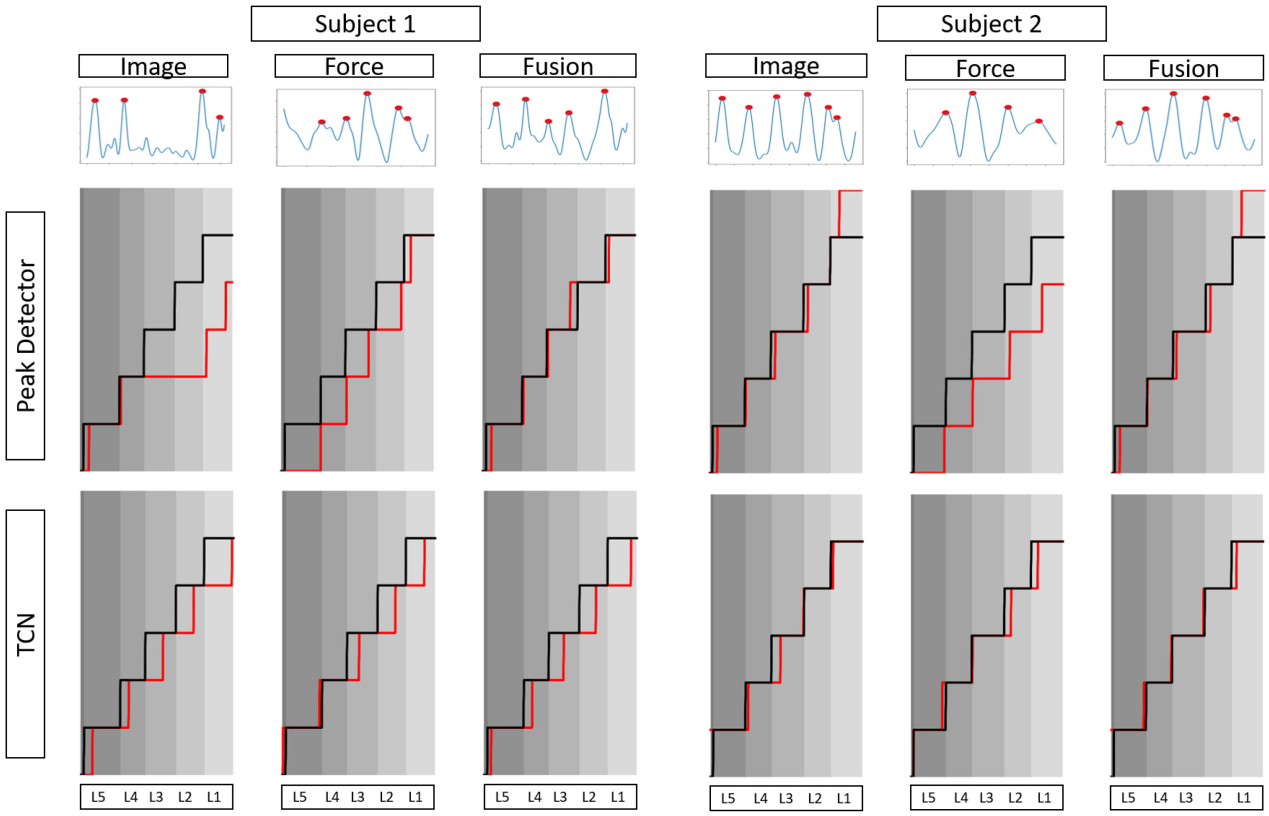


Fig. 8. The predicted (red line) and ground-truth (black line) vertebral levels for pure force-based, pure ultrasound-based and force-ultrasound fusion both when using a peak detector and a tcn. Subject 1 with anatomical characteristics non well-represented in the tcn training set (Subject Gender: Male, BMI: 30, Height: 186 cm). Subject 2 with anatomical characteristics well-represented in the tcn training set (Subject Gender: Female, BMI: 22, Height: 172 cm).

In Fig. 8 the results for Subject 2 (Height: 172 cm) for the three methods are shown in the presence of noisy ultrasound data. In this example it can be noticed that the TCN is able to compensate both for the first missing peak (L5) in the force signal and for the spurious peak in the image signal (L1).

3) The Potential Application: The performances of the presented method were tested for an example application, i.e. automatic target plane selection for facet injection procedures. The facet injection procedure is performed to deliver anaesthetics at the level of facet joints, i.e. the anatomical structures connecting consecutive vertebrae (Fig. 3(b)).

Using the proposed vertebral level classification method, the correct vertebral level can be selected, and a sweep can be taken at the correct level with the probe in a paramedian sagittal orientation, to identify the target injection plane.

The method for facet plane identification is similar to the one presented in [17]. Each frame in the sweep is classified as either "facet" or "non-facet" plane and the two frames with the highest probability in the sweep are labelled as right and left facet planes. The labels are represented as boolean values where 1 corresponds to "facet" and 0 to "non-facet" plane. The plane classification task is performed using ResNet18, given its high performances in the ultrasound classification task (Section. III-1). The model was pre-trained on ImageNet and fine-tuned on a training set sampled from Dataset 3. The spatial errors between identified facet joint planes and labelled planes were calculated on 4 test subjects sampled from Dataset 3, which consisted of 20 vertebrae sweeps (5 vertebrae for each subject), each containing two facet joints, resulting in 40 facet joints in total. For 37 facet joints out of 40, the mean distance error between the detected and manually labelled facet planes is 2.08 ± 2.63 mm. According to [18] an error below 5 mm leads to an effective anaesthetic result for the facet joint injections. For the rest 3 facet joints out of 40, the error is 8.43 ± 8.98 mm since the CNN output resulted to be less precise, due to the poor image quality.

IV. Conclusion

Currently, clinical routine spine injections procedures completely rely on the expertise of the surgeon, both to ensure the accuracy of the procedure and to limit the exposure time to the ionizing radiation. In this study, a robotic-ultrasound method for vertebral level detection and counting was developed for spine injection procedures. To the best of our knowledge, it is the first robotic system integrating visual and force feedback for vertebra level classification.

The current work shows that the fusion of force and ultrasound data is effective for vertebral level counting compared to only using ultrasound or force data separately. The use of a TCN network was explored and compared with a peak detector method. The TCN showed improved performance in both force and image methods, which highlights the ability of the TCN to learn not only the peaks in the input signal but also an anatomical prior on the positions of vertebral levels. This is highly beneficial in

the case of missing or corrupted peaks in the input signal, where the network can use the anatomical prior to compensating for the missing information.

Nevertheless, this may cause misclassification for subjects where the spine anatomy is significantly different from those in the training dataset. However, the best detection performances across all methods were obtained with the fused input, proving that the combination of force and image is beneficial for vertebral level counting.

The method was tested on a group of healthy volunteers, chosen to maximize the inter-subject variability in terms of gender and BMI. The detection accuracy is reported for a group of healthy volunteers. However, the appearance of the spine in the US images of pathological patients might not be as clear as in healthy subjects. Therefore, further clinical studies should be conducted to evaluate the method accuracy on real patients, in more challenging scenarios. The robotic procedure can be used for most of the patients undergoing facet joint injections. However, for patients with pathologies as, spondylolisthesis or tumour the usage of the robotic system is not advisable. In principle, there is no contraindication for hernia patients, although lying in a prone position might not be comfortable enough. It is anyways advisable to test both the position and the force on the patient's back prior to the procedure, to ensure the patient's comfort. Despite the promising results of the presented system, further steps are to be taken for the deployment of the system in a clinical environment. The initial position of the probe should be automatized to avoid any misplacement, e.g. by integrating the proposed system with methods for automatic sacrum localization as in [19]. Furthermore, to account for the curvature of the spine in medical conditions as scoliosis, the spinous process tracking during the procedure might be enabled. In fact, in these conditions, the spine might fall out of the field of view of the probe. Tracking could be improved by using, for example, landmark localization algorithms to center the robot on the spine. Enabling tracking of the spine potentially extends the application of the system to another clinical scenario as scoliosis assessment. Curvature reconstruction using robotic ultrasound systems was already explored in the literature [20] and might benefit from the proposed vertebrae level counting and additional tracking. The final step to be explored toward fully automatic injection procedures is the automation of the target point localization as well as of the mechanical injection itself. Despite further steps are to be taken for the deployment of the system in a clinical environment, this work opens the path for future exploration toward fully automatic injection procedures.

REFERENCES

[1] C. E. Alexander and M. Varacallo, "Lumbosacral facet syndrome," in *Proc. StatPearls [Internet]*. StatPearls Publishing, 2019.

[2] I. M. Skaribas, J. L. Erian, D. Reynolds, and E. E. Skaribas, "Lumbar interlaminar epidural injection," in *Proc. Deer's Treatment Pain*. Springer, 2019, pp. 405–412.

[3] J. Boon, P. Abrahams, J. Meiring, and T. Welch, "Lumbar puncture: Anatomical review of a clinical skill," *Clinical Anatomy (New York, N.Y.)*, vol. 17, pp. 544–53, 10 2004.

[4] I. Evansa, I. Logina, I. Vanags, and A. Borgeat, "Ultrasound versus fluoroscopic-guided epidural steroid injections in patients with degenerative spinal diseases: A randomised study," *Eur. J. Anaesthesiology*, vol. 32, no. 4, pp. 262–268, May 2015.

[5] J. Hetherington, V. Lessoway, V. Gunka, P. Abolmaesumi, and R. Rohling, "SLIDE: Automatic spine level identification system using a deep convolutional neural network," *Int. J. Comput. Assist. Radiology Surg.*, vol. 12, no. 7, pp. 1189–1198, Jul. 2017.

[6] B. Kerby, R. Rohling, V. Nair, and P. Abolmaesumi, "Automatic identification of lumbar level with ultrasound," in *Proc. 30th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc., EMBS'08 - "Personalized Healthcare through Technol."*, pp. 2980–2983, 2008.

[7] S. Yu, K. K. Tan, B. L. Sng, S. Li, and A. T. H. Sia, "Real-time automatic spinal level identification with ultrasound image processing," in *Proc. IEEE 12th Int. Symp. Biomed. Imag.*, Apr. 2015, pp. 243–246.

[8] J. Esteban *et al.*, "Robotic ultrasound-guided facet joint insertion," *Int. J. Comput. Assisted Radiology Surg.*, vol. 13, no. 6, pp. 895–904, 2018.

[9] R. B. van Heeswijk, G. Bonanno, S. Coppo, A. Coristine, T. Kober, and M. Stuber, "Motion compensation strategies in magnetic resonance imaging," *Critical Reviews. Biomed. Eng.*, vol. 40, no. 2, 2012.

[10] A. Pepin, J. Daouk, P. Bailly, S. Hapdey, and M.-E. Meyer, "Management of respiratory motion in pet/computed tomography: The state of the art," *Nucl. Med. Commun.*, vol. 35, no. 2, pp. 113–122, 2014.

[11] O. Zettinig *et al.*, "3D ultrasound registration-based visual servoing for neurosurgical navigation," vol. 12, no. 9, pp. 1607–1619, 2017.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 770–778.

[13] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 4700–4708.

[14] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Representations*, 2015.

[15] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. CVPR09*, 2009, pp. 248–255.

[16] Y. A. Farha and J. Gall, "MS-TCN: Multi-stage temporal convolutional network for action segmentation," in *Proc. IEEE Conf. Comput. Vision Pattern Recognition*, 2019, pp. 3575–3584.

[17] M. Pesteie *et al.*, "Real-time ultrasound image classification for spine anesthesia using local directional hadamard features," *Int. J. Comput. Assisted Radiology Surg.*, vol. 10, no. 6, 2015.

[18] M. Greher*et al.*, "Ultrasound-guided lumbar facet nerve block: A sonoanatomic study of a new methodologic approach," *Anesthesiology: The J. Amer. Soc. Anesthesiologists*, vol. 100, no. 5, pp. 1242–1248, 2004.

[19] H. Hase *et al.*, "Ultrasound-guided robotic navigation with deep reinforcement learning," 2020, *arXiv:2003.13321*.

[20] M. Victorova, D. Navarro-Alarcon, and Y. Zheng, "3d ultrasound imaging of scoliosis with force-sensitive robotic scanning," in *Proc. Third IEEE Int. Conf. Robotic Comput.*, Feb. 2019, pp. 262–265.