Automatic Labelling of Tumourous Frames in Free-Hand Laparoscopic Ultrasound Video

Jeremy Kawahara[1], Jean-Marc Peyrat[2], Julien Abinahed[2], Osama Al-Alao[3], Abdulla Al-Ansari[2,3], Rafeef Abugharbieh[4], and Ghassan Hamarneh[1]

[1]Medical Image Analysis Lab, Simon Fraser University, Burnaby, Canada
[2]Qatar Robotic Surgery Centre, Qatar Science & Technology Park, Doha, Qatar
[3]Urology Department, Hamad General Hospital, Hamad Medical Corporation, Qatar
[4]BiSICL, University of British Columbia, Vancouver, Canada

Abstract. Laparoscopic ultrasound (US) is often used during partial nephrectomy surgeries to identify tumour boundaries within the kidney. However, visual identification is challenging as tumour appearance varies across patients and US images exhibit significant noise levels. To address these challenges, we present the first fully automatic method for detecting the presence of kidney tumour in free-hand laparoscopic ultrasound sequences in near real-time. Our novel approach predicts the probability that a frame contains tumourous tissue using random forests and encodes this probability combined with a regularization term within a graph. Using Dijkstra's algorithm we find a globally optimal labelling (tumour vs. non-tumour) of each frame. We validate our method on a challenging clinical dataset composed of five patients, with a total of 2025 2D ultrasound frames, and demonstrate the ability to detect the presence of kidney tumour with a sensitivity and specificity of 0.774 and 0.916, respectively.

1 Introduction

Among patients with early-stage kidney cancer, undergoing a partial rather than a radical nephrectomy has been shown to improve patient survival [9]. During a partial nephrectomy, the surgeon's goal is to remove the entire tumour from the kidney while maximally sparing the kidney's healthy tissue. Minimally invasive surgery (MIS) can be used to perform a partial nephrectomy and has been shown to reduce operation time, lessen blood loss during surgery, reduce infection rates, and allow for shorter hospital stays and recovery times [5]. During MIS, an endoscope and often a laparoscopic ultrasound (US) probe are inserted into the insufflated patient's body. The endoscope is used to visualize the surfaces of the organs while the US probe provides a way to visualize the internal structures, allowing surgeons to accurately localize the tumour when choosing the correct areas to resect (i.e., the tumour-free margins). Currently, the US probe is manually controlled by one of the surgeon's assistants who sweeps and rotates the US probe over the tumour and the healthy kidney while the surgeon carefully observes both the real-time US and endoscopic video sequences.

The US scanning process produces a sequence of US frames showing the transition between tumour and healthy tissue. However, distinguishing tumour from healthy tissue in US is challenging due to several reasons. The US signal-to-noise-ratio is quite low, making the visual appearance of the tumour similar to that of healthy tissue. Moreover, within the scan of the same patient, the tumour often looks quite different depending on the position and orientation of the probe. This is complicated and challenging to model as the tumour appearance can change across patients who in some cases may also suffer from other abnormalities such as cysts. Such problems are prevalent in US, which is known to be operator-skill dependent and challenging to read [4]. We note that while an ultrasound radiologist is more experienced in interpreting US, such specialists are usually not present in the operating room, leaving junior surgeons with less US interpretation training to navigate these challenges.

A system capable of automatically detecting tumours in US sequences is valuable as it could help identify those frames containing tumour, which in turn may increase the speed and confidence at which the surgeon determines the correct tumour-free margins. Such an automated system also reduces intra/inter-operator variability and may serve as a second expert opinion mimicking the role of a trained ultrasound radiologist in determining the correct tumour boundaries.

Detecting tumour using US images has wide clinical applications and has previously been studied not only for the kidney [1] but also for other organs such as the breast [2, 6, 7]. Hao *et al.* [6] segmented lesions from static 2D US breast images by combining superpixels, detection windows, learned weights and boosted features within a conditional random field formulation. While they showed promising results in segmenting lesions in 2D, they did not examine temporal US sequences which is our focus. Ahmad *et al.* [1] segmented a *phantom* kidney tumour from freehand ultrasound sequences by requiring the user to trace the correct boundary on the first US frame. This contour then acted as a seed for the next image followed by further refinement of the boundary. Jiang *et al.* [7] detected breast tumours in 2D US images by first training AdaBoost on Haar-like features to detect candidate tumour bounding boxes. They then used a support vector machine to discriminate between tumour and non-tumour regions. Finally, the centre points of the classified tumour boxes were used to seed a random walker segmentation.

Exploiting the temporal nature of US sequences provides valuable regularization context. However, very few works dealt with detecting tumours from sequences of 2D US (videos). Bocchi *et al.* [2] incorporated information from short video clips of 2D US frames to classify benign and malignant tumours. However, their method required manual input to obtain a segmentation that was then used in the classification.

These state-of-the-art methods detected tumour in static (single) US images [6, 7], used an external US probe [6, 7], required user input [1, 2], or were only tested on phantom data [1] instead of clinical videos. In contrast, our work uses *in vivo* free-hand laparoscopic ultrasound sequences and is validated on

challenging clinical patient video data composed of significant kidney and tumour variability.

We propose an automatic *tumour detection* strategy in endoscopic US video. In our first step, we extract features computed over the 2D scan and train a random forest classifier to assign probabilities of a tumour and tumour-free label for each frame. In our second step, we use these extracted probabilities as the data term in an energy function regularized by a binary term penalizing neighbouring temporal frames with different labels. We encode our energy formulation within a trellis that allows for the derivation of a globally optimal labelling over the entire US sequence. We validate this approach on real clinical data and show that our temporal regularization provides marked improvements over independently labelling the frames.

2 Method

In this section we describe our energy function which includes the learned data and regularization terms. We also detail how we encode our problem as a minimal path optimization, which results in a fast globally optimal labelling of the frames. An overview of our method can be seen in Fig. 1.

2.1 Energy Function for US Frame Labelling

Given a video sequence X of n 2D US frames $\boldsymbol{x}_i; i = 1 \dots n$, we want to find a set of labels Y where each label $y_i; i = 1 \dots n$ is assigned either a tumour or non-tumour designation, $y_i \in \{0 : \text{non-tumour}; 1: \text{tumour}\}$. The optimal labelling of frames Y^* minimizes the following energy function,

$$Y^* = \underset{Y}{\operatorname{argmin}}\, E(Y, X) = \underset{Y}{\operatorname{argmin}} \sum_{i=1}^{n} D(y_i, \boldsymbol{x}_i) + \sum_{i=1}^{n-1} \lambda R(y_i, y_{i+1}). \tag{1}$$

The energy function $E(Y, X)$ is composed of a data term D and a regularization term R. Given a single US frame $\boldsymbol{x}_i$ and a candidate label for the frame y_i, the data term D is based on the probability of the single frame containing tumour and is defined as,

$$D(y_i, \boldsymbol{x}_i) = \begin{cases} -\log p(y_i = 1 | \Phi(\boldsymbol{x}_i)), & \text{if } y_i = 1 \\ -\log p(y_i = 0 | \Phi(\boldsymbol{x}_i)), & \text{if } y_i = 0 \end{cases} \tag{2}$$

where the probability of the tumour given the image $p\left(y_i | \Phi(\boldsymbol{x}_i)\right)$ is learned using random forests trained on features $\Phi(\boldsymbol{x}_i)$ (described in the next section) extracted from the US image. The regularization term R is defined using a 1D Potts model,

$$R(y_i, y_j) = 1 - \delta(y_i - y_j) \tag{3}$$

where $\delta(y_i - y_j)$ returns 1 if the two neighbouring labels, y_i, y_j, are equal or 0 otherwise. This term penalizes neighbouring labels that are different from each

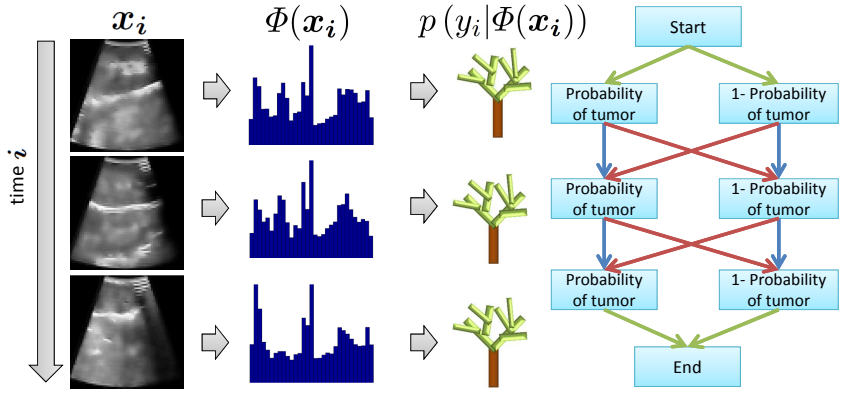

Fig. 1. An overview of our method showing three 2D US frames over time (*left*). We extract features $\Phi(\boldsymbol{x}_i)$ (*middle histogram*) from these US frames and pass them to trained random forests. The random forests outputs the probability of the frame containing a tumour. These probabilities are used as the data term and encoded in the trellis (*right*). A start and an end node are added with uniform weights (*green arrows*). The nodes on the *left* represent frames that are labelled non-tumour, while the nodes on the *right* represent a labelling of the frame as tumour. Crossing edges (*red arrows*) represent a change in the labels and are penalized higher than those that maintain the same label (*blue arrows*).

other. To balance between these two terms, we weigh the regularization term by a constant λ which we empirically set to 12. We observed that in practice our method is fairly insensitive to the choice of λ.

2.2 Random Forest Data Term and Extracted Features

For our data term in Eq. 2, we use random forests to learn the probability $p(y_i|\Phi(\boldsymbol{x}_i))$ of a US frame containing a tumour independent of the other frames. We chose random forests due to the probabilistic output they produce (i.e., uncertainty-encoding) and their overall ability to generalize well to unseen data with little parameter tuning when compared to other machine learning approaches [3].

As input to the random forests, we first extract features $\Phi(\boldsymbol{x}_i)$ from the US frame capable of discriminating between tumourous and healthy tissue despite their heterogeneous appearance across patients. This is especially challenging for our particular dataset as the appearance of the tumour widely differed depending on the angle and position of the free-hand laparoscopic probe.

Prior to extracting features, we down-sample the image by a factor of eight to reduce run times. We then apply a Weiner filter to each 2D US scan to adaptively smooth the image based on local image variance. We extract a histogram of gradient directions from the entire image as our feature vector $\Phi(\boldsymbol{x}_i)$, where the angles are grouped into bins of 12 degrees. This results in a total

of 30 features (360/12) extracted from each image. We empirically found these low-dimensional features to discriminate well between tumour and healthy US scans across patients in our dataset and chose simplicity and speed to support real-time processing and clinical translation. We note the recent success of using Histograms of Oriented Gradients to improve the specificity of discriminating between masses and healthy tissues in mammography images [8]. While it is non-trivial to analyze the use of a histogram of gradient directions combined with random forests (Fig. 1), our experience to date suggests that these features, computed over the entire image, capture variations in textural patterns caused by differences in acoustic impedance, attenuation, reflection, and scattering properties between healthy tissue and tumour. Furthermore, the histogram encourages a pose-invariant representation which is useful for the random forest classifier.

2.3 Minimal Path Optimization

We formulate the optimization of Eq. 1 as a minimal path problem and apply a standard search algorithm (Dijkstra's/Viterbi algorithm) to obtain a globally optimal frame labelling Y^*. In particular, we represent our problem as a graph in the form of a trellis (Fig. 1 - right), such that each node in the graph represents a potential label of a US frame in our sequence over time. At each node (representing the label for a US frame) we apply a directed edge to the two nodes representing the 'tumour' and 'non-tumour' labels of the next US frame. Each connected node is then weighted by the energy as specified in Eq. 1. A start node is added to the beginning with equal weights to either label of the nodes that represent the first US frame, and an end node with equal weight transition probabilities is added after the nodes that represent the final frame.

3 Results

Abiding by ethical review board requirements, we collected data from five kidney tumour patients undergoing a robot-assisted partial nephrectomy using the da Vinci Si surgical system (Intuitive Surgical, Inc.). Brightness mode (B-mode) ultrasound sequences were acquired using an Ultrasonix SonixTablet ultrasound system (Ultrasonix, Analogic Ultrasound). Each US video comprised a sequence of 2D B-mode frames recorded at a rate of 10 frames per second, resulting in 405 US frames per patient. The frames captured the underlying tissue as the human operated free-hand laparoscopic probe swept over the kidney and the tumour.

Obtaining labelled (ground-truth) data is quite difficult as it is not always possible to visually distinguish between tumourous and healthy kidney tissue from US information alone. We therefore manually linked the US sequences temporally with the endoscopic video data such that both videos were played at the same time. This was accomplished using the Kinovea software[1] which

[1] http://www.kinovea.org/

allowed us to view two videos simultaneously, change the playback speed of each independently, annotate the videos, and crop them across time. With the US and endoscopic video synchronized, we were able to simultaneously view the location of the probe relative to the exophytic tumour (visible in the endoscopic video) and see the approximate corresponding US frame. This endoscopic video gave us cues to help distinguish between tumour and healthy kidney when creating the manual ground-truth. Manual contours were hand drawn in consultation with our clinical collaborators, where we outlined the tumours on specific frames based on the information in both the US image and the endoscopic video. A label for each frame was obtained by assigning the label of 'tumour' to all frames with a drawn tumour in them. Thus the ground-truth labelling of 'tumour' or 'non-tumour' was assigned to each frame guided *collectively* by the endoscopic video, the information contained in the US frame, and the appearance of the tumour across time. All final labels were confirmed with the clinical collaborator. The classes were fairly balanced with 42.5% of the frames being labeled as 'tumour'.

We performed leave-one-patient-out cross validation where we omited one patient entirely from the training dataset, trained our method on the remaining four patients, and then tested on the omitted patient. For all experiments we used 80 trees in our random forests as this was sufficient for the out-of-bag error to stabilize. Our method assigned each of the frames the label of 'tumour' or 'non-tumour', which we compared against the ground-truth labelling. To quantitatively evaluate our method, we computed the *sensitivity*, $TP/(TP + FN)$, as a measure of how well the method detected tumour in a frame when there actually was a tumour, and *specificity*, $TN/(FP + TN)$, as a measure of how well the method reported no tumour when there was no tumour in a frame. We also calculated the overall *accuracy*, $(TP + TN)/(TP + FP + FN + TN)$, and *precision*, $TP/(TP + FP)$, of our method.

We validated our method over the aforementioned dataset and report our results in Table 1. We compared our method first by training the classifier to detect the tumour within an image based on the data term only, i.e., without temporal regularization. This achieved the least accurate results (Table 1 - row 1) as each label is assigned without considering its neighbours. By adding the regularization term, we obtained significant improvements to the results. Our improvements are most noticeable in specificity, which is critical to prevent overestimating tumour boundaries and the subsequent removal of healthy nephrons.

Exemplar qualitative results using our method can be seen in Fig. 2, highlighting where our method was successful and unsuccessful in discriminating between tumour vs. non-tumour frames. We note that many areas where our method incorrectly classified the frames occur in areas where the difference between tumour and healthy kidney is highly ambiguous within the B-mode. Errors were also observed when the patient's tumour had very different appearance information when compared to the other training data.

Training the random forests over a training set of 1620 frames requires approximately 8 seconds on a 2.33 GHz quad-core machine. Once trained, given a novel patient our method labels the remaining 405 US frames in under 2 seconds

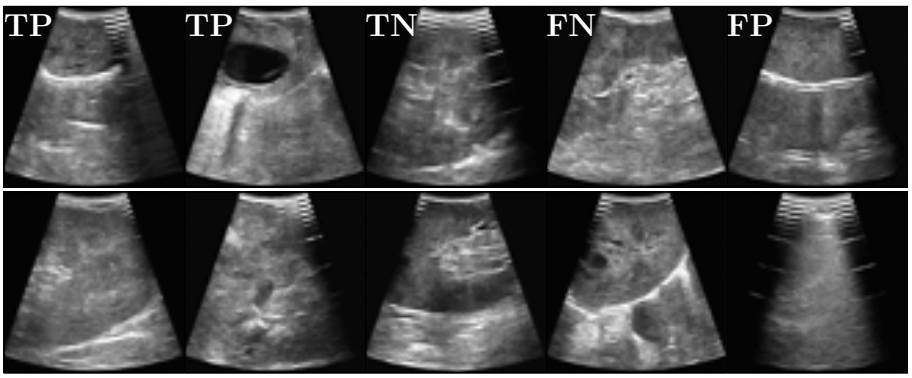

Fig. 2. Typical 2D slices from various patients illustrating the challenges of detecting tumour in US. The first two columns (TP) show examples of where our method was able to correctly identify frames as containing tumour. The third column (TN) shows healthy kidney our method correctly labelled as non-tumour. The fourth column (FN) shows two examples containing tumour our method incorrectly labelled as healthy. The fifth column (FP) shows healthy slices our method incorrectly labelled as tumour.

using un-optimized and non-parallelized MATLAB code. As any training can be done offline prior to the surgery, our approach appears to be a promising practical solution for real-time intra-operative clinical use.

4 Conclusions

To the best of our knowledge, we presented the first work that automatically labels tumourous frames in free-hand laparoscopic ultrasound video with the capacity for real-time processing. Our method was validated on a relatively large clinical dataset of minimally invasive image guided partial nephrectomies. Our method is a first step towards providing an automated "US radiologist" support system in the operating room. We encode a learned random forest based probabilistic data term regularized temporally as a minimal path problem, allowing us to optimally label frames using standard minimal path solvers. With continuous acquisition of additional datasets by our surgical team, we expect to significantly expand our training sets allowing our model to better learn the high variability between patients and improve our performance even further. Our fast offline version, in addition to being a crucial component for tasks such as automatically marking tumour in endoscopic video, also serves as a first step for a future online version. Future work will be focused on incorporating data from other imaging modalities to improve the detection rate in areas where the US image information on its own is insufficient.

Acknowledgements. We thank Alborz Amir-Khalili, Ivan Figueroa-Garcia, and Masoud S. Nosrati for their assistance with data acquisition and helpful

Table 1. Results from applying our method to 2025 2D US scans from five separate patients. The first two rows represent the results computed across all patients. The first row acts as our baseline test by thresholding the probability of a tumourous frame using a trained random forests, *RF*. In the second row we add the temporal regularization term. The remaining five rows separate each patient to show individual results across 405 2D sequential US frames. We note the low sensitivity in patient 4 is largely due to a tumour with appearance that is not seen in the other patients (Fig. 2 row 1, col 4), hence not accurately learned in our supervised method.

Method	Patient	Sensitivity	Specificity	Accuracy	Precision
RF	all	0.765	0.848	0.813	0.788
proposed	all	**0.774**	**0.916**	**0.855**	**0.872**
proposed	1	1.000	0.705	0.802	0.626
	2	0.943	1.000	0.978	1.000
	3	0.810	1.000	0.877	1.000
	4	0.423	1.000	0.696	1.000
	5	0.862	0.942	0.923	0.818

feedback. This publication was made possible by NPRP Grant #4-161-2-056 from the Qatar National Research Fund (a member of the Qatar Foundation). The statements made herein are solely the responsibility of the authors.

References

1. Ahmad, A., Cool, D., Chew, B.H., Pautler, S.E., Peters, T.M.: 3D Segmentation of Kidney Tumors from Freehand 2D Ultrasound. In: SPIE. vol. 6141. Medical Imaging 2006: Visualization, Image-Guided Procedures, and Display (2006)
2. Bocchi, L., Gritti, F., Manfredi, C., Giannotti, E., Nori, J.: Semiautomated Breast Cancer Classification from Ultrasound Video. In: ISBI IEEE. pp. 1112–1115 (2012)
3. Caruana, R., Niculescu-Mizil, A.: An Empirical Comparison of Supervised Learning Algorithms. In: ICML. pp. 161–168. ACM (2006)
4. Finberg, H.J.: Whither (Wither?) the Ultrasound Specialist? Journal of Ultrasound in Medicine 23(12), 1543–1547 (2004)
5. Gill, I.S., Kavoussi, L.R., Lane, B.R., et al.: Comparison of 1,800 Laparoscopic and Open Partial Nephrectomies for Single Renal Tumors. Urology 178(1), 41–46 (2007)
6. Hao, Z., Wang, Q., Wang, X., Kim, J.B., Hwang, Y., Cho, B.H., Guo, P., Lee, W.K.: Learning a Structured Graphical Model with Boosted Top-Down Features for Ultrasound Image Segmentation. In: Mori, K., Sakuma, I., Sato, Y., Barillot, C., Navab, N. (eds.) MICCAI 2013. pp. 227–234. Springer Berlin Heidelberg (2013)
7. Jiang, P., Peng, J., Zhang, G., Cheng, E., Megalooikonomou, V., Ling, H.: Learning-Based Automatic Breast Tumor Detection and Segmentation in Ultrasound Images. In: IEEE ISBI. pp. 1587–1590 (2012)
8. Pomponiu, V., Hariharan, H., Zheng, B., Gur, D.: Improving Breast Mass Detection using Histogram of Oriented Gradients. In: Aylward, S., Hadjiiski, L.M. (eds.) SPIE Medical Imaging. vol. 9035 (2014)
9. Tan, H.J., Norton, E.C., Ye, Z., Hafez, K.S., Gore, J.L., Miller, D.C.: Long-Term Survival Following Partial vs Radical Nephrectomy Among Older Patients with Early-Stage Kidney Cancer. JAMA 307(15), 1629–1635 (2012)