International Journal of Computer Assisted Radiology and Surgery (2020) 15:1177–1186
https://doi.org/10.1007/s11548-020-02151-w

ORIGINAL ARTICLE

Detecting the occluding contours of the uterus to automatise augmented laparoscopy: score, loss, dataset, evaluation and user study

Tom François[1,2] · Lilian Calvet[1] · Sabrina Madad Zadeh[1] · Damien Saboul[2] · Simone Gasparini[1] · Prasad Samarakoon[1] · Nicolas Bourdel[1] · Adrien Bartoli[1]

Received: 21 November 2019 / Accepted: 31 March 2020 / Published online: 5 May 2020

Abstract

Purpose The registration of a preoperative 3D model, reconstructed, for example, from MRI, to intraoperative laparoscopy 2D images, is the main challenge to achieve augmented reality in laparoscopy. The current systems have a major limitation: they require that the surgeon manually marks the occluding contours during surgery. This requires the surgeon to fully comprehend the non-trivial concept of occluding contours and surgeon time, directly impacting acceptance and usability. To overcome this limitation, we propose a complete framework for object-class occluding contour detection (OC2D), with application to uterus surgery.

Methods Our first contribution is a new distance-based evaluation score complying with all the relevant performance criteria. Our second contribution is a loss function combining cross-entropy and two new penalties designed to boost 1-pixel thickness responses. This allows us to train a U-Net end to end, outperforming all competing methods, which tends to produce thick responses. Our third contribution is a dataset of 3818 carefully labelled laparoscopy images of the uterus, which was used to train and evaluate our detector.

Results Evaluation shows that the proposed detector has a similar false false-negative rate to existing methods but substantially reduces both false-positive rate and response thickness. Finally, we ran a user study to evaluate the impact of OC2D against manually marked occluding contours in augmented laparoscopy. We used 10 recorded gynecologic laparoscopies and involved 5 surgeons. Using OC2D led to a reduction of 3 min and 53 s in surgeon time without sacrificing registration accuracy.

Conclusions We provide a new set of criteria and a distance-based measure to evaluate an OC2D method. We propose an OC2D method which outperforms the state-of-the-art methods. The results obtained from the user study indicate that fully automatic augmented laparoscopy is feasible.

Keywords Edge detection · Distance-based score · Edge detector evaluation · Convolutional neural network · Deep learning · Laparoscopy · Augmented reality

Introduction

Augmented monocular laparoscopy requires the registration of a preoperative 3D model to laparoscopy images. As shown in Fig. 1, the state-of-the-art registration systems [2,4,10] rely on visual cues extracted from laparoscopy images, especially the organ's occluding contours. For a given imaged object, an *occluding contour* refers to any boundary fragment where the object is an occluder, and is thus part of the object's *silhouette*. The occluding contours are essential to constrain the registration of a deformable biomechanical model, as shown for the uterus [4] and the liver [2,10]. These systems are well advanced in terms of registration computation. However, they require the surgeon to mark the occluding contours manually on laparoscopy images during surgery. This significantly reduces the acceptance and usability of augmented laparoscopy because the concept of occluding contour is non-trivial and marking them requires surgeon time. We propose to detect the organ's occluding contours

✉ Tom François
tom.francois@etu.uca.fr

[1] Université Clermont Auvergne, CHU Clermont-Ferrand, CNRS, SIGMA, Institut Pascal, Clermont-Ferrand, France

[2] Be-Studys, A Brand of Be-Ys Group, 123 Route de Meyrin, 1219 Châtelaine, Suisse, Vernier, Switzerland

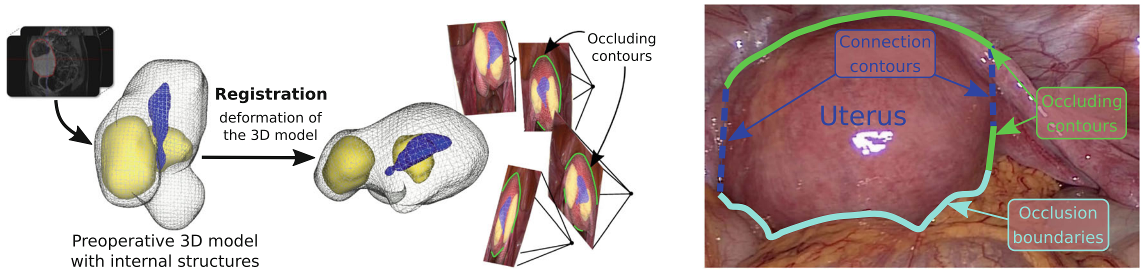


Fig. 1 (Left) In augmented laparoscopy, the preoperative 3D model is registered by fitting the occluding contours of the organ in laparoscopy images. The current systems require the surgeon to mark these contours manually during surgery. (Right) An occluding contour arises at an organ boundary where the organ occludes another structure, as opposed to an occlusion boundary where the organ is occluded by another structure. The set of occluding contours is the silhouette. OC2D is the task of detecting the occluding contours for a specific object, here the uterus. It forms a task of semantic detection far more challenging than organ segmentation

automatically in order to build the critically missing component needed to automatise the existing systems. We tackle the general problem, which we refer to as *object-class occluding contour detection* (OC2D) and specialise our detector to the uterus.

OC2D is an open problem, closely related to semantic edge detection and occlusion boundary detection. Semantic edge detection finds the boundary of objects and is somehow dual to semantic segmentation. The early methods relied on edges, located on abrupt brightness changes [3,8,15]. However, object boundaries do not always lie on edges, especially when the object and background colours are similar. Recent CNN-based approaches thus combine higher-level features with learnt shape and appearance object priors [12,24,26]. Occlusion boundary detection finds the boundary of all objects and classifies them according to their occlusion relationship. This classification makes the task more difficult than semantic edge detection. CNN-based approaches have shown to perform well over a large number of object classes in natural images. OC2D combines the difficulty of a specific object class and of the occlusion relationship. Its application to the uterus in laparoscopy images increases the difficulty as the colours are clearly not discriminative. The literature lacks a specific solution method for OC2D, as well as several critical parts which we discuss in the next paragraphs.

The first missing part for OC2D is an evaluation score complying with all the relevant performance criteria. Three performance criteria were defined by Canny in 1986 for edge detection in his seminal work [3]: C1, true contours should not be missed and responses not spurious; C2, responses should be close to true contours, and C3, each true contour should only produce a single response. As discussed in [13,14], the evaluation scores used in the literature fail one or several of Canny's criteria. Most of them are derived from classification frameworks and rely on precision-recall measures at the pixel level. They fail C2 as they equally penalise mislocalised responses irrespective of their distance to true contours. The use of a tolerance region allowing one to consider slightly mislocalised responses as true responses is used in [8,15]. Yet, their score fails C2 as the response-to-true contour distance is not considered. They also fail C3 as several responses can match a true contour within the tolerance region. In contrast, we propose an evaluation score complying with all of Canny's criteria and with two other proposed criteria. These, named C4 and C5, ensure that the score is left invariant by changing object deformation, camera intrinsics and pose. They are important because we want the occluding contours to equally constrain registration over the set of images. Specifically, for a given object and amount of occlusion, we have that the score should be invariant to: C4, image resolution and C5, the amount of true contours. We compare the proposed score to existing ones [13–15] on synthetic contours.

The second missing part for OC2D is the detector itself, specifically the loss to train a CNN end to end. Using a CNN is a natural approach, as in related tasks [1,5,16,22,23,25,26]. These methods do not address OC2D specifically but reveal the important potential problem of thick responses [1,5,25]. These approaches require complex learning pipelines and a large body of training data. In contrast, we propose an end-to-end OC2D method which encourages 1-pixel thickness responses. We use a U-Net, and our contribution lies in a loss combining cross-entropy with two new penalties we call BiP and TiP for binarising the outputs and thinning the contours. We propose training strategies with these penalties.

The third missing part for OC2D is a dataset, specifically in laparoscopy. Existing datasets [11,18,20,21] do not comprise labels for anatomical structures and the type of occlusion. We propose a dataset of 3818 carefully labelled laparoscopy images of the uterus meant to address gynecologic surgery. The labels are as in Fig. 1, the occluding contours, the occlusion boundaries and the connection contours of the uterus.

We evaluated our detector on randomly chosen test images from the proposed dataset. We used U-Net trained with cross-entropy as baseline. We also compared with CASENet [26], which we specialised to OC2D for the uterus. All three methods show similar FN (false-negative) rates, but ours is substantially better in terms of FP (false-positive) rate and thickness of response fragments.

Lastly, we conducted a user study to evaluate the gain of using OC2D in augmented laparoscopy in an existing system [4], against manual marking by the surgeon. The user study was performed on 9 recorded laparoscopy videos and involved 5 surgeons. Intraoperative surgeon time was substantially reduced, registration accuracy was preserved, and the system became usable by any surgeon, without the need to understand the concept of occluding contour. This confirmed the crucial importance of automation in augmented laparoscopy.

Related work

Evaluation score The principle of the score from [8,15] is widely used in semantic edge detection [1,25,26]. The score is based on precision-recall obtained by matching responses and true edges. It also uses a tolerance region to deal with spurious responses. Unfortunately, the matching requires to solve the minimum flow over a bipartite graph, which is in practice only solvable approximately. Also recall that the score fails C2. An exhaustive list of scores for edge detection is given in [13], following three categories: local, statistical and distance-based. Strong arguments in favour of distance-based scores are given in [14], which gives a thorough comparison and proposes a distance-based score integrating the number of FP and FN. These are, however, unequally weighted, causing the score to be overly sensitive to spurious responses, failing C3–C5. In contrast, the score we propose shares the same desirable features but gracefully copes with spurious responses.

Detection methods and loss OC2D has not been specifically addressed in the literature, but semantic edge detection and occlusion boundary detection are closely related tasks. For both, the best results are currently obtained with CNNs. In semantic edge detection, the task is to detect the boundary of multiple specific objects [1,5,12,25,26]. Weighted cross-entropy is commonly used to compensate the imbalanced distribution between the edge and non-edge classes over the image. This weighting, however, has the negative effect to favour response fragments thicker than the true edges. However, [1,25] suggest that these may be due to the imperfect labelled contours and adjust them during training to address this problem, while [1,5] propose a specific loss based on the reciprocal Dice coefficient.

In occlusion boundary detection, the task is to detect all occlusion boundaries in the image. Existing methods use a two-stage approach, where the object boundaries are first detected and then ordered depthwise. Some methods use a shared encoder and multiple decoders. SharpNet [16] uses a U-Net with three decoders to predict depthmaps, occluding contours and normals. Other methods [7,22,23] combine two parallel streams estimating boundary location and occlusion orientation. In [22], a specific loss is proposed to boost detection nearby class-agnostic object boundaries once the cross-entropy loss stalls.

Our proposed detector designed for OC2D takes inspiration from these related tasks. We use a U-Net and weighted cross-entropy as most methods. Similarly to [22], we boost the detection once mere cross-entropy stalls by adding penalties. The penalties we propose are, however, radically new. Our binarising penalty favours binary responses of the network to encourage sharp contour maps, and our thinning penalty favours well-localised responses to encourage thin contours.

Datasets There exist datasets of labelled laparoscopy images for supervised learning-based detection of surgical actions [11], surgical phases [18,21] and anatomical structures [11]. These datasets are procedure specific, namely cholecystectomy [11,18,21] and fibroid resection [11,20]. There exist datasets for semantic segmentation of robotic surgical instruments, stereo correspondence and reconstruction in endoscopy [9]. However, there do not exist public datasets of laparoscopy images labelled for semantic segmentation of the anatomical structures. The proposed dataset is thus the first of its kind. It includes advanced organ boundary information, namely the occlusion boundary, occluding contour and connection contour, carefully labelled on 3818 images extracted from various procedures.

Evaluation score

We propose a contour evaluation score complying with the five performance criteria C1–C5. We then compare it with existing scores.

Formulation and compliance with the five performance criteria

Formulation Let $\mathcal{I}$ be the set of all image pixels coordinates, $C \subset \mathcal{I}$ the true contours and $R \subset \mathcal{I}$ the responses of a contour detector. We use a tolerance distance $d_{\max}$ such that a *missed contour* is defined as a true contour with no response located at a distance lower than $d_{\max}$ from it and a *spurious response* is a response located at a distance greater than $d_{\max}$ from true contours. In practice, $d_{\max}$ is chosen as 2% of the image diagonal [8]. A missed contour and a spurious response are considered as FN and FP, respec-

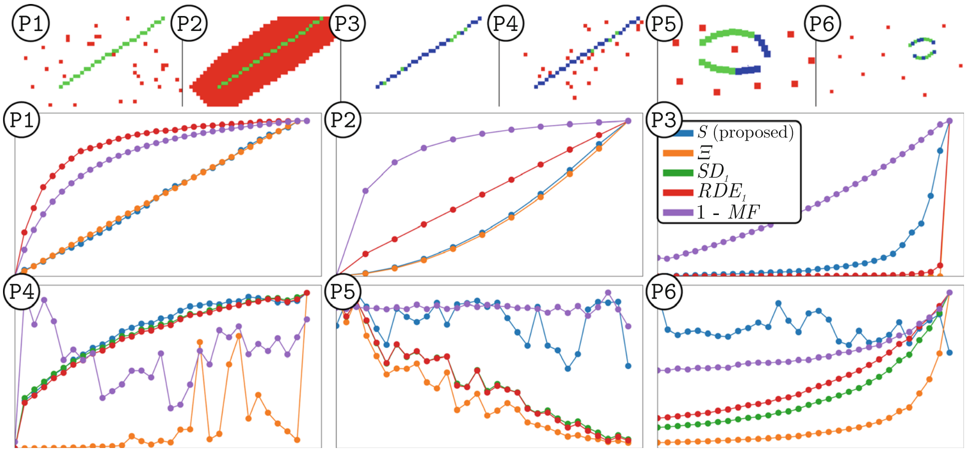


Fig. 2 (top row) The six types of contour perturbation P1–P6 with TP in green, FP in red, FN in blue and TN in white. (Bottom row) The evaluation scores rescaled to fit the graphs

tively, in the sequel. The responses in the tolerance region $\mathcal{T} = \{r \in \mathcal{I} \mid d(r, C) < d_{\max}\}$ are then TP (true positives) and the responses outside $\mathcal{T}$ are FP.

The proposed score $S(R, C)$ combines $d_{\max}$ with the distance between true contours and responses for the first time. It combines the following three terms:

$$S_{\text{TP}} = \frac{1}{2}\left(\frac{1}{|C|}\sum_{r\in R\cap\mathcal{T}} d(r, C\backslash\text{FN}) + \frac{1}{|C|}\sum_{c\in C\backslash\text{FN}} d(c, R\cap\mathcal{T})\right),$$
$$S_{\text{FP}} = \frac{d_{\max}}{|I| - 2|C|d_{\max}}|\text{FP}| \;\; \text{and}$$
$$S_{\text{FN}} = \frac{d_{\max}}{|C|}|\text{FN}|.$$

Specifically, $S(R, C)$ sums the three terms and normalises by $d_{\max}$:

$$\begin{aligned} S(R, C) &= \frac{1}{d_{\max}}\left(S_{\text{TP}} + S_{\text{FP}} + S_{\text{FN}}\right) \\ &= \frac{1}{2|C|d_{\max}}\left(\sum_{r\in R\cap\mathcal{T}} d(r, C\backslash\text{FN}) + \sum_{c\in C\backslash\text{FN}} d(c, R\cap\mathcal{T})\right) \\ &\quad + \frac{|\text{FP}|}{|I| - 2|C|d_{\max}} + \frac{|\text{FN}|}{|C|}. \end{aligned}$$

Compliance with C1, C2 S_{TP} is a symmetric distance between the true contours and responses.

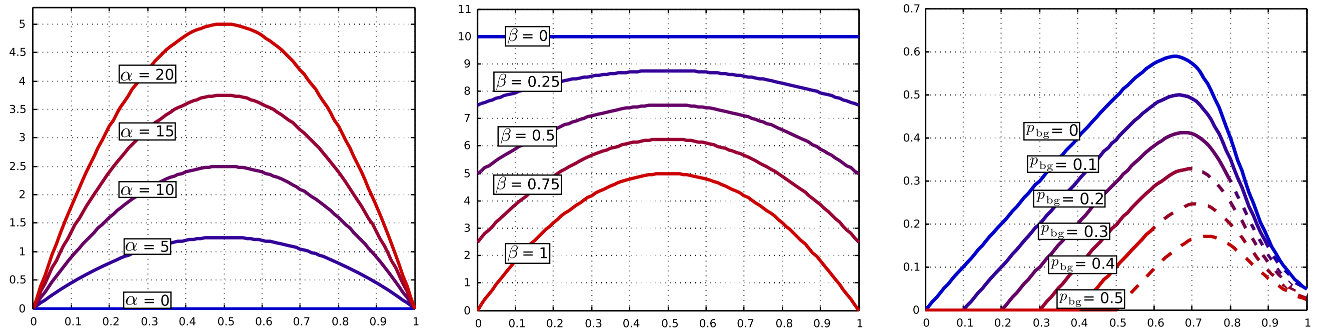
It thus encourages C2, namely responses close to true contours. S_{FP} and S_{FN} are the normalised FP and FN, respectively, each counting for $d_{\max}$. They thus encourage C1, namely no spurious responses and no missed contours, respectively, while equally penalising spurious responses irrespective of their distance to true contours.

Compliance with C3 The difficulty in complying with C3 arises from the distance in S_{TP} which uses the nearest true contour to each response, which possibly associates the same true contour to multiple responses. We handle this by penalising deviation between the number of true contours and responses within the tolerance region, using normalisation by $|C|$, whilst previous work use $|R \cap \mathcal{T}|$ [6].

Compliance with C4, C5 A high FN rate tends to have lower impact than a high FP rate and requires proper weighting [14]. We assume that the probability of having a spurious response is (1) uniform within the tolerance region and (2) similar to the probability of missing a true contour. In order to equally penalise FP and FN inside and outside the tolerance region, our weighting is to normalise S_{FP} and S_{FN} by their spatial extent, specifically $||\mathcal{I}|-2|C|d_{\max}|$ pixels, considered a good approximation of the number of pixels outside the tolerance region, for S_{FP}, and $|C|$ pixels for S_{FN}. In summary, all three terms are normalised according to the number of true contours $|C|$ while the second term also integrates the image resolution to satisfy C4 and C5.

Evaluation

As shown in Fig. 2, we simulated six types of perturbation, P1–P6, between true contours and responses, some borrowed from [13], to test C1–C5. *P1: adding FP, 1* (C1 and C3): an increasing number of random false responses are added. *P2: adding FP, 2* (C3): an increasing number of false responses are added by dilating true contours to simulate thick responses within $d_{\max}$. *P3: adding FN* (C1): an increasing amount of random true responses is deleted. *P4: locations* (C2): the location of true responses are independently randomly perturbed with an increasing magnitude. *P5: downsampling* (C4): the image is increasingly downscaled with constant FN rate. *P6: downscaling* (C5): the contours are downscaled with constant FN rate. Importantly, a score verifying C1–C5 is expected to increase for P1–P4 and to



(a) $p_*, * \in \{\mathrm{oc}, \mathrm{ob}, \mathrm{bg}\}$ (b) $p_*, * \in \{\mathrm{oc}, \mathrm{ob}, \mathrm{bg}\}$ (c) $p_\#, \# \in \{\mathrm{oc}, \mathrm{ob}\}$

Fig. 3 **a** Binarising penalty in the amplitude strategy, BiPα, with $\alpha \in [0, 20]$ and $\beta = 1$. **b** Binarising penalty in the frequency strategy, BiPβ, with $\alpha = 20$ and $\beta \in [0, 1]$. **c** Thinning penalty, TiP, for $p_{\mathrm{bg}} \in [0, 0.5]$. The dashed parts are not applicable, with $p_{\mathrm{bg}} + p_\# > 1$

remain steady for P5, P6. The evaluation of the proposed score S (Proposed) and four competitors, namely Ξ [14], SD_1 [8], RDE_1 [8] and $1 - \mathrm{MF}$ [6,15] are shown in Fig. 2. We observe that P1 and P2 are passed by all scores. However, P3, P5 and P6 defeat all scores but $1-\mathrm{MF}$ and S (Proposed). Finally, only the proposed score S passes P3 and is thus compliant with all performance criteria.

Detector and loss

Architecture and training overview We propose the first end-to-end OC2D method. We take care to comply with C1–C5, especially with C3, namely to return a single response per true contour pixel. This is probably the toughest criterion as response thickness is one of the main limitations of current CNN-based semantic edge and occlusion boundary detectors. The problem is also well-known in edge detection from image gradient. These detectors trigger, for instance, if the gradient magnitude is larger than a threshold. A low threshold thus leads to overdetection and violates C3, whereas a high threshold leads to high FN rates and violates C1. Finding a threshold to comply with both C1 and C3 is generally not possible. The popular Canny edge detector [3] solves this problem using a low threshold and performs morphological operations to thin the responses.

The proposed detector takes inspiration from the Canny detector but uses a CNN and an end-to-end training process. The key idea is to design new penalties to integrate thinning in the loss. We chose a U-Net architecture because it performs well for semantic segmentation with a limited amount of training images. We output three probability maps $\mathcal{P} = \{p_{\mathrm{oc}}, p_{\mathrm{ob}}, p_{\mathrm{bg}}\} \in [0, 1]^3$ for the occluding contours, the occlusion boundaries and the background (see Fig. 1) and use a softmax layer to ensure $p_{\mathrm{oc}} + p_{\mathrm{ob}} + p_{\mathrm{bg}} = 1$. We propose a three-step training procedure, gradually integrating two new advanced structural penalties in the loss: the Binarising Penalty (BiP) and the Thinning Penalty (TiP).

First training step: initial task learning The first training step specialises the model to the OC2D task using a mere cross-entropy loss:

$$\mathcal{L}_1(\mathcal{P}, \mathcal{Y}) = \sum_{* \in \{\mathrm{oc},\mathrm{ob},\mathrm{bg}\}} \mu_* \mathcal{L}_{\mathrm{CE}}(p_*, y_*), \tag{1}$$

where $*$ simply runs over the three classes, $\mathcal{Y} = \{y_{\mathrm{oc}}, y_{\mathrm{ob}}, y_{\mathrm{bg}}\} \in \{0, 1\}^3$ are the true labels with $y_{\mathrm{oc}} + y_{\mathrm{ob}} + y_{\mathrm{bg}} = 1$, $\mu_{\mathrm{oc}} = 1$, $\mu_{\mathrm{ob}} = 1.5$ and $\mu_{\mathrm{bg}} = 0.01$ are fixed weights, and $\mathcal{L}_{\mathrm{CE}}$ is cross-entropy. We stop training when the model stalls.

Second training step: binarising The second training step fine-tunes the model to binarise its outputs, as in image binarisation. It combines cross-entropy with a new Binarising Penalty (BiP) B designed to encourage binary outputs:

$$\begin{aligned}\mathcal{L}_2(\mathcal{P}, \mathcal{Y}) &= \mathcal{L}_1(\mathcal{P}, \mathcal{Y}) \\ &\quad + \sum_{* \in \{\mathrm{oc},\mathrm{ob},\mathrm{bg}\}} \alpha\big((1-\beta)K + \beta \mathrm{B}(p_*)\big).\end{aligned} \tag{2}$$

B is affinely combined with a constant K, and involves two hyperparameters α and β tuned during training. We propose $\mathrm{B}(x) = x(1-x)$ as the simplest BiP. We propose two training strategies meant to gradually increase the BiP effect, illustrated in Fig. 3. The *amplitude strategy*, denoted BiPα, where α gradually increases from 0 to 20 while β is set to 1. The *frequency strategy*, denoted BiPβ, where β gradually increases from 0 to 1 while α is set to 20. The increase in α and β is 0.05 after each epoch. We stop training when the model stalls and keep the best model with hyperparameters α_{opt}, β_{opt}. The outputs are still in the [0, 1] range but become very close to $\{0, 1\}$.

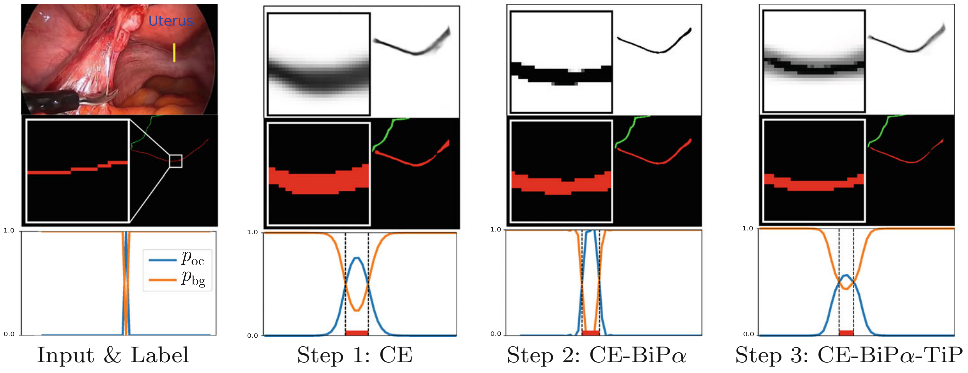


Fig. 4 (Column 1, top) Input image for uterus OC2D with a cross section of the occlusion boundary in yellow. (Column 1, middle) True occluding contours in red and occlusion boundaries in green. (Columns 2–4, top) Output probability map p_{oc} from the proposed OC2D at each of its training steps. CE is cross-entropy, BiP is our binarising penalty, and TiP is our thinning penalty. (Columns 2–4, middle) Results of the proposed OC2D as in column 1, middle. (All columns, bottom) Probabilities p_{oc} and p_{bg} along the selected cross section with transitions between background and occlusion boundary in dashed black

Third training step: thinning The third training step fine-tunes the model to favour thin responses, as in morphological edge thinning. It combines cross-entropy with a new thinning penalty (TiP), penalising pixels whose probability of being an occluding contour is higher than and yet close to those of not being one:

$$\mathcal{L}_3(\mathcal{P},\mathcal{Y}) = \mathcal{L}_1(\mathcal{P},\mathcal{Y}) + \sum_{\#\in\{\mathrm{oc},\mathrm{ob}\}} \gamma \max(0, p_\# - p_{\mathrm{bg}})\sigma(\theta(\lambda - p_\#)), \tag{3}$$

where # simply runs over the two contour classes. In the TiP term, illustrated in Fig. 3, γ is a hyperparameter which we vary in the [0, 40] range, increasing by 0.05 after each epoch. The first factor penalises the pixels for which $p_\# > p_{\mathrm{bg}}$ as a linear function of the probability discrepancy. The second factor penalises the pixels for which $p_\# < \lambda$, where $\lambda \in [0, 1]$ is a fixed threshold which we chose as $\lambda = 0.8$. A value close to 1 means that only those pixels nearby true contours should be detected. It uses a sigmoid σ and a fixed slope $\theta = 15$. We stop training when the model stalls and keep the best model with hyperparameter γ_{opt}. The outputs represent much thinner contours (see Fig. 4).

Dataset of uterus laparoscopy

We propose the first dataset of laparoscopy images with accurate advanced contour labels for 3818 images.

Images The images come from 79 anonymous uterus laparoscopy videos, 29 available from an IRB-approved study in our hospital and 50 from YouTube. These show a variety of procedures including hysterectomies, resections of endometriosis nodules and cysts, salpingectomies, adenomyomectomies and myomectomies. We extracted multiple frames from each video to ensure that our dataset captures the two essential types of variability. The first variability is the intra-patient and within-procedure one, which is due, for instance, to viewpoint change, uterus deformation and colour change, as the procedure goes by. The second variability is the inter-patient and multiple-procedure one, which is due, for instance, to the shape and appearance of the uterus, and specific changes caused by the type of procedure and the disease. We also took care to include various typical events such as occlusion by surgical instruments and blurry images (Fig. 5).

Labels and labelling As shown in Figs. 1 and 5, the labels are the occluding contours, the occlusion boundaries and the connection contours of the uterus. The connection contours typically occur at the junction between the uterus and the fallopian tubes, and at the cervix, where the uterus ends, but there is no occlusion boundary or occluding contour. The connection contours are not used in our OC2D method, but they nonetheless represent valuable information, as together with the occlusion boundary and occluding contours they define the uterus region. The labelling was done by a surgeon using the online platform Supervisely [19].

Evaluation

Evaluation overview We evaluated the proposed OC2D method and its three training steps against a baseline and existing work. We refer to our first training step, namely a U-Net trained with cross-entropy, as the *baseline*. The proposed training from Sect. 4 is CE-BiPx-TiP, where $x \in \{\alpha, \beta\}$. The naming uses '-' between the training steps. In order to

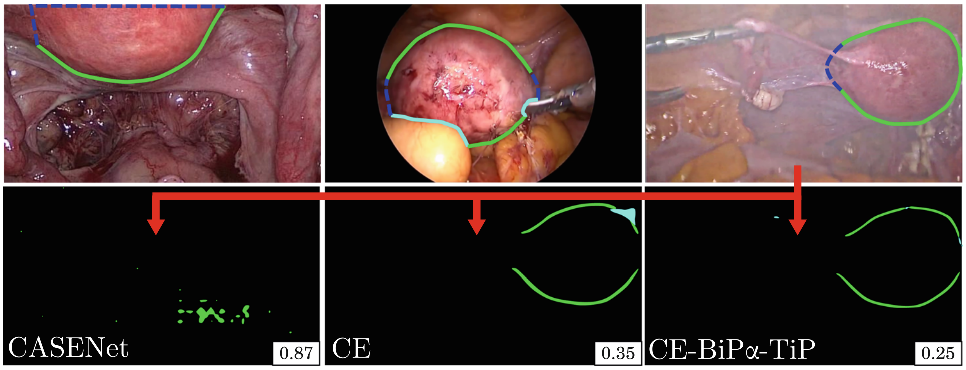


Fig. 5 Top: excerpts from our dataset of 3818 labelled laparoscopy images. The uterus is the main organ of interest. The occluding contours are in green, the occlusion boundaries in cyan and the connection contours in blue. Bottom: responses of OC2D applied on the rightmost image, showing the robustness of the proposed method to surgical smoke. The obtained scores are of 0.87, 0.35 and 0.25 for CASENet, CE and CE-BiPα-TiP, respectively

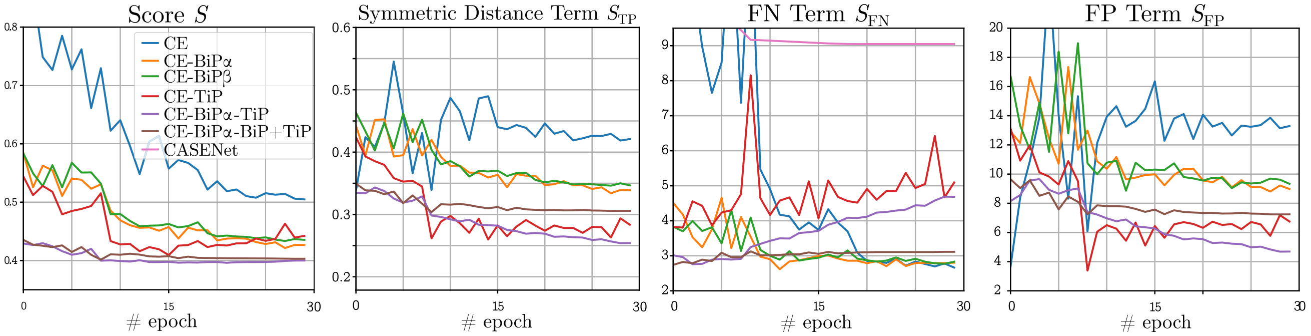


Fig. 6 Quantitative performance evaluated during training of OC2D. From left to right: the overall score $S = \frac{1}{d_{\max}}(S_{TP} + S_{FP} + S_{FN})$ and its three terms S_{TP}, S_{FP} and S_{FN}. The lower the better. CASENet results are out of the graphs ranges for S, S_{TP} and S_{FP}

understand the role of each loss term and of the training steps, we have four alternative scenarios, whose names are self-explaining: CE-BiPx, CE-BiPα-BiP+TiP and CE-TiP, where '+' means an aggregate of loss terms. We compared with CASENet [26].

Implementation We use the implementation of U-Net and CASENet in Pytorch from [17] and [1], respectively. We fine-tuned CASENet on our dataset from pretraining on the Semantic Boundary Dataset [8]. We used stochastic gradient descent and decayed the initial learning rate by 0.1 every 10 epochs. We used a random 72%–13%–15% train-validation-test split of our data et.

Results Several quantitative and qualitative results of the OC2D methods applied on highly challenging cases are shown in Figs. 5 and 7. They show in particular robustness of the proposed OC2D to strong uterus occlusions, presence of smoke and blood. Quantitative performance evaluated during training is shown in Fig. 6, using the proposed score S, and a breakdown of its three terms S_{TP}, S_{FP} and S_{FN}. Apart from CASENet which shows very poor performance, we observe that the baseline CE has the worst performance. The proposed binarising penalty improves performance compared to CE, similarly for both training strategies in CE-BiPα and CE-BiPβ. The full proposed training CE-BiPα-TiP with both penalties obtains the best results, improving in all respects but slightly degrading the FN rate, as thinning increases underdetection. CE-TiP, which skips the second training step, has lower performance. Interestingly, CE-BiPα-BiP+TiP, which includes both penalties in the third training step, performs closely to CE-BiPα-TiP. It improves the FN rate but degrades the TP and FP rates. The very poor performance obtained with CASENet could be partly explained by the limited number of training images (Fig. 7).

User study

We ran a user study to evaluate three OC2D methods, namely our baseline CE, CE-BiPα-TiP of Sect. 4 (the best performing in Sect. 6) and CASENet against manually marked occluding contours in an existing augmented laparoscopy system [4]. We used 10 recorded gynecologic laparoscopies with MRI and preoperative 3D model collected under IRB

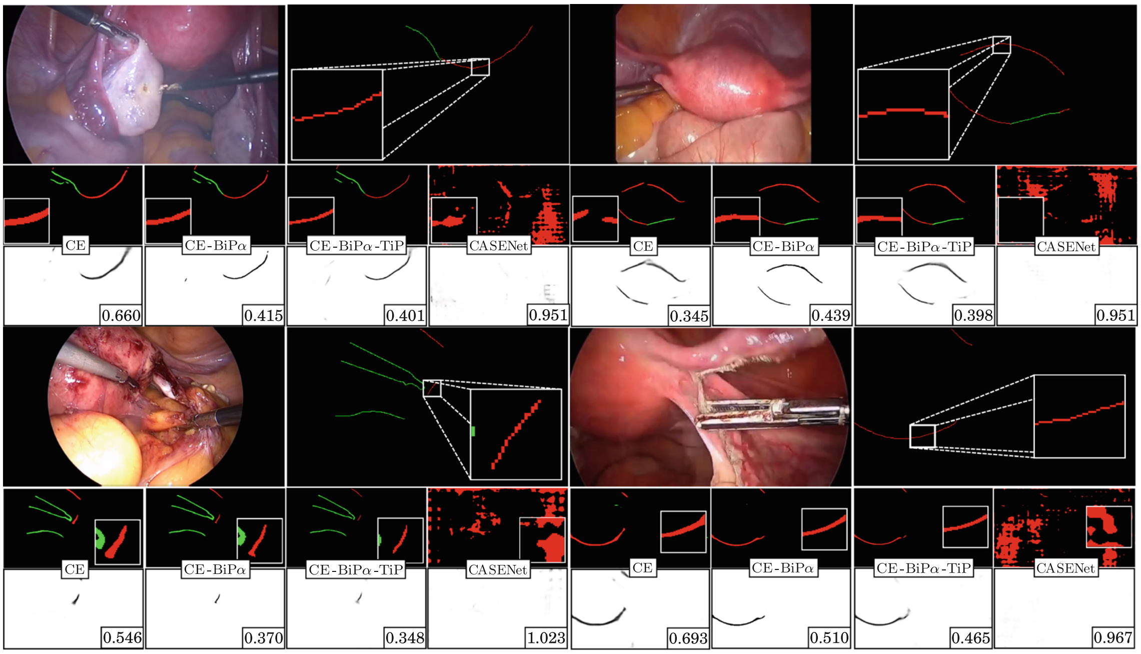


Fig. 7 OC2D responses for four laparoscopy image examples. For each example, the first row represents the input laparoscopy image and the manually marked ground truth. Occluding contours are marked in red and occlusion boundaries in green. The second row corresponds to the detectors responses. The third row corresponds to the output probabilities p_{oc}

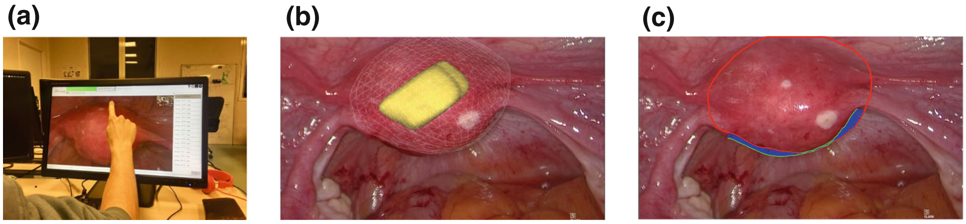


Fig. 8 **a** Labelling of the occluding contours by a surgeon using a tactile screen. **b** Augmentation with the preoperative 3D model registered using the occluding contours from OC2D. A myoma is visualised in yellow and the uterus external surface in light gray. **c** Evaluation of the registration accuracy using the reprojection error, as the distance between the silhouette of the tracked 3D model (red) and manually labelled occluding contours of the uterus (green)

approval in our hospital. We involved 5 surgeons, broken down in 3 juniors and 2 seniors, all of them familiar with augmented reality. The surgeons were asked to label the occluding contours of the uterus using a tactile screen (see Fig. 8a) as in surgery conditions, and marking time was recorded, for 18 images on average. We independently ran the OC2D methods on the same image sets, and running time was recorded. The registration accuracy was then evaluated for each laparoscopy by running [4]. The results of this system directly depend on the occluding contours, as it uses them to constrain preoperative 3D model registration. The system then tracks the uterus to perform live augmentation. We evaluated accuracy by evaluating the reprojection error of the tracked 3D model in a set of 10 independent frames. The reprojection error is defined as the average distance between the tracking-predicted occluding contour and its careful annotation, as shown in Fig. 8b. The frames were selected to ensure viewpoint variability towards the uterus and such that at least 10% of the tracked 3D model reprojects in the image. This procedure was run for the 10 laparoscopies, the 5 surgeons and 3 OC2D methods, which led to a total of 80 cases. The results are shown in Table 1. CE and CE-BiPα-TiP led to nearly identical registration accuracy as manual marking, but to a dramatic reduction of surgeon time of 3 min and 53 s on average, representing 97.4% of augmented reality setup time. Despite showing completely aberrant contour responses, CASENet shows an average error 14 pixels higher than the proposed CE and CE-BiPα-TiP, a difference which is not as significant as we expected. It is due to the use of an *M-estimator* in the occluding contour term of the

Table 1 User study for 10 laparoscopies, averaged over 5 surgeons for manual results

Case	Manual	CE-BiPα-TiP	CE	CASENet	Time OC2D	Time Manual
1	**34.97**	42.13	42.89	71.85	7.7″	4′56″
2	56.41	53.93	**53.14**	60.58	7.8″	5′12″
3	**93.40**	94.83	95.92	127.44	6.7″	4′39″
4	42.13	**40.84**	43.42	50.46	7.8″	5’37″
5	85.13	88.10	**80.33**	93.31	8.5″	4′34″
6	90.47	**90.07**	91.37	100.30	5.1″	3′37″
7	96.34	90.62	92.17	**84.72**	6.6″	3’24″
8	**46.76**	48.56	49.03	54.58	5.5″	3′28″
9	**32.27**	33.92	33.47	49.30	6.1″	3′30″
10	39.58	41.43	**38.39**	67.00	2.2″	1′28″
Average	**61.75**	62.44	62.01	75.95	6.4″	4′02″

Bold highlights the better results for each row (the lowest error). The reprojection error (the lower, the better) is in pixels. The time is in minute (′) and seconds (′/). Time OC2D is evaluated with CE-BiPα-TiP, but other methods present similar values

minimised energy proposed in [4] that makes the registration method highly robust to false contour responses. A stronger consequence of this study is to indicate that fully automatic augmented laparoscopy is feasible. The fact that the surgeon should understand the concept of occluding contour and devote undivided attention to label around 20 images during surgery has been prohibitive for the wide acceptance of augmented reality. With OC2D, this constraint is now dropped, and usability dramatically increased.

Conclusion

We have identified the organ-specific detection of occluding contours as a key missing component in the usability of computer-aided laparoscopy with augmented reality. We have identified this component with OC2D, an open and challenging semantic detection problem, for which we have proposed a complete framework. This includes a distance-based evaluation score, the first to comply with all performance criteria including Canny's, a loss allowing one to train a CNN-based detector, with two new specific penalties, and a dataset of carefully labelled laparoscopy images. Our penalties binarise the response map and thin the response contours. They allow our detector to outperform the baseline and existing work, in terms of response thickness, FN and FP rates. We have conducted a user study to evaluate the impact of automation by OC2D against manually marked occluding contours in augmented laparoscopy. Automation led to a substantial reduction of surgeon time while preserving augmentation accuracy. The surgeons are relieved from the intraoperative labelling task and from understanding the concept of occluding contours, confirming our initial motivation of developing OC2D. As future work, we plan to study self-supervision for OC2D by using silhouette constraints from multiple-view geometry and how the proposed binarising and thinning penalties may improve other detection tasks.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Informed consent Informed consent was obtained from all individual participants included in the study. This article does not contain any studies with animals performed by any of the authors.

References

1. Acuna D, Kar A, Fidler S (2019) Devil is in the edges: learning semantic boundaries from noisy annotations. In: CVPR
2. Adagolodjo Y, Trivisonne R, Haouchine N, Cotin S, Courtecuisse H (2017) Silhouette-based pose estimation for deformable organs application to surgical augmented reality. In: IROS
3. Canny JF (1986) A computational approach to edge detection. TPAMI 8(6):679–698
4. Collins T, Pizarro D, Bartoli A, Canis M, Bourdel N (2014) Computer-assisted laparoscopic myomectomy by augmenting the uterus with pre-operative mri data. In: ISMAR
5. Deng R, Shen C, Liu S, Wang H, Liu X (2018) Learning to predict crisp boundaries. In: ECCV
6. Dubuisson M, Jain A (1994) A modified hausdorff distance for object matching. In: ICPR
7. Grard M, Chen L, Dellandréa E (2019) Bicameral structuring and synthetic imagery for jointly predicting instance boundaries and nearby occlusions from a single image. arXiv
8. Hariharan B, Arbeláez P, Bourdev L, Maji S, Malik J (2011) Semantic contours from inverse detectors. In: ICCV
9. ISCAS: Miccai endoscopic vision challenges (2019). https://endovis.grand-challenge.org
10. Koo B, Ozgur E, Roy BL, Buc E, Bartoli A (2017) Deformable registration of a preoperative 3d liver volume to a laparoscopy image using contour and shading cues. In: MICCAI

11. Leibetseder A, Petscharnig S, Primus MJ, Kletz S, Münzer B, Schoeffmann K, Keckstein J (2018) Lapgyn4: a dataset for 4 automatic content analysis problems in the domain of laparoscopic gynecology. In: Proceedings of the 9th ACM multimedia systems conference, MMSys, pp 357–362
12. Liu Y, Cheng M, Hu X, Bian J, Zhang L, Bai X, Tang J (2019) Richer convolutional features for edge detection. TPAMI 41(8):1939–1946
13. Lopez-Molina C, Baets BD, Sola HB (2013) Quantitative error measures for edge detection. Pattern Recognit. 46(4):1125–1139
14. Magnier B, Abdulrahman H, Montesinos P (2018) A review of supervised edge detection evaluation methods and an objective comparison of filtering gradient computations using hysteresis thresholds. J. Imaging 4(6):74
15. Martin DR, Fowlkes CC, Malik J (2004) Learning to detect natural image boundaries using local brightness, color, and texture cues. PAMI 26(5):530–549
16. Ramamonjisoa M, Lepetit V (2019) Sharpnet: Fast and accurate recovery of occluding contours in monocular depth estimation. arXiv
17. Ronneberger O, Fischer P, Brox T (2015) U-net: Convolutional networks for biomedical image segmentation. In: MICCAI
18. Stauder R, Ostler D, Kranzfelder M, Koller S, Feußner H, Navab N (2016) The TUM lapchole dataset for the M2CAI 2016 workflow challenge. arXiv
19. Supervisely. https://supervise.ly/
20. Török P, Harangi B (2018) Digital image analysis with fully connected convolutional neural network to facilitate hysteroscopic fibroid resection. Gynecol. Obstet. Investig. 83(6):615–619
21. Twinanda AP, Shehata S, Mutter D, Marescaux J, de Mathelin M, Padoy N (2017) Endonet: a deep architecture for recognition tasks on laparoscopic videos. IEEE Trans Med Imaging 36(1):86–97
22. Wang G, Wang X, Li FWB, Liang X (2018) Doobnet: deep object occlusion boundary detection from an image. In: ACCV
23. Wang P, Yuille AL (2016) DOC: deep occlusion estimation from a single image. In: ECCV
24. Yang J, Price BL, Cohen S, Lee H, Yang M (2016) Object contour detection with a fully convolutional encoder-decoder network. In: CVPR
25. Yu Z, Liu W, Zou Y, Feng C, Ramalingam S, Kumar BVKV, Kautz J (2018) Simultaneous edge alignment and learning. In: ECCV
26. Yu Z, Feng C, Liu M, Ramalingam S (2017) Casenet: deep category-aware semantic edge detection. In: CVPR

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.