CrossMark

ORIGINAL ARTICLE

Automatic data-driven real-time segmentation and recognition of surgical workflow

Olga Dergachyova[1,2] · David Bouget[1,2] · Arnaud Huaulmé[1,2,3] ·
Xavier Morandi[1,2,4] · Pierre Jannin[1,2]

Abstract

Purpose With the intention of extending the perception and action of surgical staff inside the operating room, the medical community has expressed a growing interest towards context-aware systems. Requiring an accurate identification of the surgical workflow, such systems make use of data from a diverse set of available sensors. In this paper, we propose a fully data-driven and real-time method for segmentation and recognition of surgical phases using a combination of video data and instrument usage signals, exploiting no prior knowledge. We also introduce new validation metrics for assessment of workflow detection.

Methods The segmentation and recognition are based on a four-stage process. Firstly, during the learning time, a Surgical Process Model is automatically constructed from data annotations to guide the following process. Secondly, data samples are described using a combination of low-level visual cues and instrument information. Then, in the third stage, these descriptions are employed to train a set of AdaBoost classifiers capable of distinguishing one surgical phase from others. Finally, AdaBoost responses are used as input to a Hidden semi-Markov Model in order to obtain a final decision.

Results On the MICCAI EndoVis challenge laparoscopic dataset we achieved a precision and a recall of 91 % in classification of 7 phases.

Conclusion Compared to the analysis based on one data type only, a combination of visual features and instrument signals allows better segmentation, reduction of the detection delay and discovery of the correct phase order.

Keywords Computer-assisted surgery · Surgical workflow · Surgical Process Modelling · AdaBoost · Hidden semi-Markov Model

✉ Olga Dergachyova
olga.dergachyova@univ-rennes1.fr

Pierre Jannin
pierre.jannin@univ-rennes1.fr

[1] INSERM, U1099, Rennes 35000, France

[2] Université de Rennes 1, LTSI, Rennes 35000, France

[3] Université Joseph Fourier, TIMC-IMAG UMR 5525, Grenoble 38041, France

[4] CHU Rennes, Département de Neurochirurgie, Rennes 35000, France

Introduction

Over the past decade, the scientific medical community has been working towards the next generation of intelligent operating room (OR) [2,6]. This concept includes various directions such as image-guided and robotic surgical systems, augmented reality and visualization, sensing devices and context-aware systems in computer-assisted interventions (CA-CAI). The research of the present work joins the last direction. The main goal of CA-CAI systems is to ensure OR situation awareness and assist the surgeon and surgical team in preventing medical errors. Needed support can be provided through multiple clinical applications, e.g. resource management [11], surgeon evaluation, detection and prevention of adverse events, aid in decision-making process and robotic assistance [7].

Creation of a CA-CAI system requires a large informative learning set of clinical data. A variety of sensors enables the extraction of diverse information. In the area of surgical workflow detection, the literature describes methods exploiting data from instruments usage binary signals [16],

RFID tags and sensors [1], different kinds of tracking devices (tracking of instruments [8], surgical staff trajectories [14] and eye motions [9]) or even simple measurements of the surgical table inclination and lights state [20]. Last but not least, video data represent a great source of information while being the most challenging one. The preferable option is to leverage data coming from sensors already used as part of the clinical routine, thus not requiring additional equipment installation. It makes endoscopic [13] and microscopic [5,12] videos ones of the best data providers. Sometimes multiple data signals are mixed to complete information lacks [21].

The second important aspect of CA-CAI systems is recognition of surgical workflow. Depending on targeting application, this can be done in different granularity levels: stages, phases, steps, actions or gestures, starting from the highest going to the lowest [11]. The approach we propose in this paper is dedicated to surgical phase discovery from video data and instrument signals. Usually the main strategy of visual-based methods for surgical workflow detection is to extract visual features such as RGB and HSV histograms, Histograms of Oriented Gradients, optical flow [17], SIFT/SURF and STIP [5] points or their combinations [12]. Extracted features are then used to classify images with machine learning techniques, e.g. support vector machine (SVM), bag-of-visual-words [12], K-nearest neighbours [5], Bayesian networks [13] or conditional random fields [17]. Knowing that the surgical procedure is a continuous process, its temporal aspect can be exploited by Hidden Markov Models (HMM) and Dynamic Time Warping (DTW) methods [3,12]. The same strategy can be applied to other sensor input including surgical tool signals. For instance, interesting results are provided by a combination of instrument usage signals with HMM or DTW [16]. Furthermore, the set of

extracted features can be reduced by selecting the most relevant ones, as it was done in [15] using AdaBoost to define the most representative surgical instruments for phase detection.

The methods for surgical workflow detection described in the literature have some drawbacks. The most essential applications of CA-CAI systems require a real-time analysis. A half of the methods are not suitable for on-line applications, especially those using DTW which are required to have a complete sequence. Besides, some of the methods have a priori assumptions on surgery type, particular instruments or regions of interest [5,12,17] constricting their application to a specific intervention only. To resolve these problems our paper proposes a completely data-driven method working in real time and independent of surgery type and its specificities. It takes as input endoscopic, microscopic or operative field videos and instrument usage binary signals.

Finally, the workflow detection method used in CA-CAI systems has to be validated. A common way to do that is to compute accuracy, precision and recall scores. However, such global scores usually not informative enough and underestimate performances as regards to the actual clinical needs. For such purpose, we propose 3 novel validation metrics and a new application-dependent approach for error estimation, and use them within the following targeted applications: estimation of remaining time, relevant information display or device triggering.

Methods

The workflow detection of our approach is performed in four stages using the data from "Data" section, as shown in Fig. 1. Firstly, a Surgical Process Model is constructed as described in "Surgical Process Modelling" section. The

Fig. 1 Scheme of the proposed four-stage segmentation and recognition approach, outlining processing at training and testing time. *D* stands for descriptions of all training samples, *S* for signatures of all training samples, *d* for test sample description and *s* for test sample signature
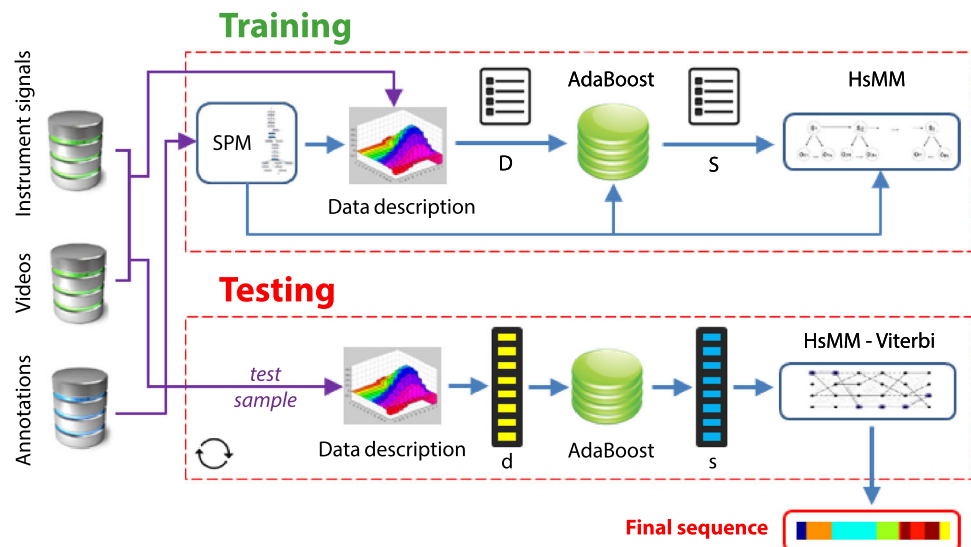
Table 1 Mean values and standard deviations of phase durations and of total operation time (min:s)

	P1	P2	P3	P4	P5	P6	P7	Total
Mean	06:30	09:23	02:50	05:36	03:00	06:59	06:31	40:49
SD	03:13	07:16	02:08	01:02	01:58	03:35	04:06	10:21

next stage is to create descriptions from images and instrument signals (see "Data description" section). "AdaBoost classification" section presents intermediate AdaBoost classification done in the third stage. The last stage, detailed in "Hidden semi-Markov Model" section, provides final classification results. Finally, in "Validation metrics" section we propose new validation metrics dedicated to workflow recognition.

Data

The dataset we used to validate our method comes from the subchallenge of the MICCAI 2015 EndoVis challenge dedicated to the surgical workflow detection.[1] It contains 7 endoscopic videos of laparoscopic cholecystectomies. The videos are in full HD quality (1920×1080) at 25 fps. Additionally, for each video frame the dataset includes usage information about the 10 instruments used during the intervention. There are 26 possible combinations of their simultaneous use.

The laparoscopic cholecystectomy operation passes through 7 phases: (P1) Placement of trocars, (P2) Preparation, (P3) Clipping and cutting, (P4) Dissection of gallbladder, (P5) Retrieving of gallbladder, (P6) Haemostasis and (P7) Drainage and closing. One operation is executed in linear order without any comebacks. In the other 6 operations the phase Haemostasis is performed before the retrieving phase is completed. Therefore it can pause the phase 5. The phase 6 is performed again after the end of the phase 5. Duration statistics are presented in Table 1, where the mean values and standard deviations were computed on the 7 surgeries.

Surgical Process Modelling

The main idea of our approach is to remain generic and independent towards the data and apply the same algorithm to different datasets. To implement this idea we rely on Surgical Process Modelling [10]. A surgical procedure can be represented with an individual Surgical Process Model (iSPM) describing its order of surgical task execution. The union of several iSPMs gives a generic model called gSPM showing different possible ways to perform the same procedure. In addition, it is possible to extract from the gSPM some useful information such as a list of all tasks and their durations. In this work the gSPM is represented in the form of a directed weighted graph automatically computed from data annotations. To construct it iSPMs are parsed in order to extract all unique surgical phases which play the role of vertices of the graph. Then, we derive all edges, meaning transitions between phases. When the structure of the graph is in place, we assign an attribute to each edge and vertex. The edge attributes define the probability of continuing by performing the phase pointed out by the edge. The vertex attributes indicate phase durations. We use the model and its statistics to manage the following segmentation and recognition process in all stages.

Data description

The second stage of the approach is to describe input data. For this we construct a visual description of each video frame and a corresponding vector of instrument signals. First, to maintain the generic aspect of the method, the images from the videos are described with no attachment to particular areas or objects and no assumption on surgery type. Only standard global visual cues are used for that purpose. This choice is driven by two reasons: 1) standard image descriptors are generally computationally light, and 2) they allow successful classification of all kinds of images unlike other more complex descriptors designed for specific goals. Three main image aspects are examined: colour, shape and texture. The colour is represented in form of histograms from RGB, HSV and LUV colour spaces. The shape is described by Discrete Cosine Transform (DCT) and Histograms of Oriented Gradients (HOG). The texture is transmitted through Local Binary Patterns (LBP) histograms. The image description is computed as follows. The RGB histogram of the entire image contains 16 bins and their sum for each colour component. The same is done for the HSV, except that the Hue component has 18 bins. For the LUV colour space we extract only L component in 10 bins, and we also take their sum. For the DCT representation only 25 values, corresponding to the highest frequencies, are taken. We compute 6 HOG of 9 bins on 6 different areas of the image. Four histograms are computed from 4 rectangular sectors of equal size going from the image centre to its corners. The fifth sector of the same size is focused on the centre of the frame. The last

sector contains the entire image. For the LBP histogram we use 58 uniform patterns only. All computed values are concatenated into a visual description vector of total 252 values $(3 \times 16 + 3 \times 1 + 18 + 2 \times 16 + 3 \times 1 + 10 + 1 + 25 + 6 \times 9 + 58 = 252)$.

Along with the visual information we incorporate the signals of instrument usage. Each analysed data sample can be described as a set of instrument binary signals: 1 if instrument is in use, 0 if not. In this way we have an instrument vector of length M, where M is a total number of surgical tools ($M = 10$ in our case). The visual description vector and instrument vector are concatenated together to be used as input for the following AdaBoost classification.

AdaBoost classification

In order to classify each analysed sample into surgical phases, the third stage of our method relies on AdaBoost classifier using descriptions from "Data description" section as input. The boosting approach, underlying concept of AdaBoost, is a machine learning technique using a great number of weak classifiers assembled into a cascade to form one strong classifier. The AdaBoost algorithm [19] finds features that separate positive data samples from negatives the best. The advantage of AdaBoost applied to our problem is its capacity to analyse each data aspect separately to find the most discriminant ones unlike SVM, for example, which takes all features in a scope. To be more concrete, visually each surgical phase can possibly differ from all others only by one or couple of features (e.g. specific colour component or gradient direction), so it can have its own particularities. The same is applicable to the instruments. Thereby, in our case each surgical phase needs a proper classifier of type one-vs-all distinguishing it from all others.

During training time, for each surgical phase, we create a set of positive (i.e. belonging to this phase) and negative (i.e. belonging to any other phase) samples represented by description vectors. Then the AdaBoost classifier specific to this phase is trained on the built set. At the end of the training we have N AdaBoost classifiers, where N is the number of surgical phases. For our dataset $N = 7$, but practically it is defined by the computed gSPM.

At testing time each sample passes through all classifiers and obtains a signature of $2 \times N$ length consisting of N positive/negative responses (1 or -1) and N confidence scores from 0 to 100 indicating classifiers certainty. Each ith response shows if the ith classifier recognizes the sample as belonging to its phase or not. The values of the signature containing confidence scores are divided into k intervals. Here k is empirically set to 20. This is needed to obtain a discrete signature which is used as input for the next stage of our method.

Hidden semi-Markov Model

In this final stage we construct a predictive model taking into account the temporal aspect of the procedure to improve the detection capacity of our method. Hidden Markov Model is a powerful tool for modelling time series. Based on observed data, the model seeks to recover the sequence of hidden states that the process comes through. Applicable to surgical workflow segmentation, the surgical phases would be the hidden states we want to discover and the data samples (signatures from the last stage in our case) would be the observations. An HMM is formalized as $\lambda = (S, O, A, B, \pi)$, where S is a finite set of states, O is a finite set of observations (sometimes called vocabulary), A is a set of probabilities of transition from one state to another, B is a probability distribution of observation emissions by states, π is a probability distribution over initial states (see [18] for more details).

Classical HMM has a major drawback: exponential state duration density due to auto-transitions. This is poorly suited for modelling of physical signals and processes with various state durations. An extension of the classical model exists, and it is called explicit-duration HMM or Hidden semi-Markov Model (HsMM). An efficient implementation was presented by Yu in [22]. A state duration probability distribution P is added to the model $\lambda = (S, O, A, B, \pi, P)$, and the probabilities of all auto-transitions in A are set to zero.

The last stage of our algorithm is HsMM training on signatures made of AdaBoost responses used as observations. First of all, the gSPM constructed at the first stage automatically defines set of phases S and initializes A, P and π. Then a finite observation vocabulary O is built from all unique signatures found in the training data. B is initially computed by counting the number of occurrences of all signatures from O in each phase. The model is then refined thanks to modified forward–backward algorithm.

At testing time, the sequence of signatures representing the surgical procedure is decoded one by one with modified Viterbi algorithm inspired from [22] in order to get a sequence of phase labels attributed to each sample as final result.

Validation metrics

In order to validate the performance of our method, we propose 3 novel metrics and an error estimation approach. These metrics are less severe, more informative and more representative of real application-based requirements. The notions of transitional moment, transitional delay, transition window and noise mentioned below are illustrated in Fig. 2.

Average transitional delay (ATD) This metric characterizes the average time delay between the real phase-to-phase transitional moments TM and predicted ones. ATD is defined as follows:
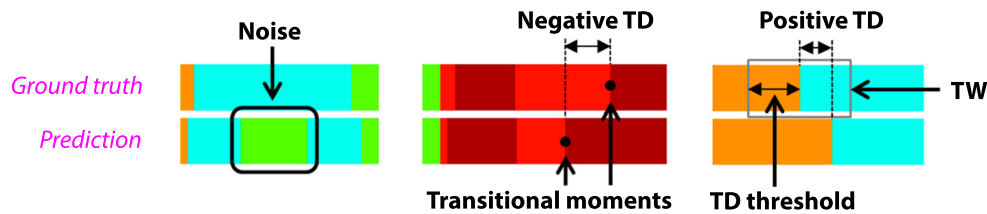
Fig. 2 Examples of transitional moments, negative and positive transitional delays (TD), transitional delay threshold, transition window (TW) and noise

$$\text{ATD} = \frac{1}{s} \sum_{k=1}^{s} abs(t_{\text{P}}(\text{TM}_k) - t_{\text{GT}}(\text{TM}_k)), \qquad (1)$$

where s is the number of TMs belonging both to the ground truth sequence GT and to the predicted sequence P, $t_{\text{GT}}(\text{TM}_k)$ is the time stamp of k-th TM in GT, and $t_{\text{P}}(\text{TM}_k)$ is the time stamp of the same closest TM in P. Transitional delay can be positive which tells that the system is late compared to the ground truth, so the detected transition is made after the real one. Negative values show that system switches too early. The average delay is computed from absolute values. Sometimes some TMs of GT are completely missed in P, and this is mostly due to a fault of the classification. The proposed metric, in the other hand, measures the reaction time of the system. Thereby, ATD does not account for such skipped transitions.

Noise level (NL) The discovered sequence can contain some noise, meaning short time separate misclassifications that create false transitional moments. This metric measures the rate of noise present in predicted sequences. NL is computed as $\text{NL} = \frac{N}{T}$, where N is the number of misclassified samples not belonging to any transitional delay, and T is the total number of samples in the predicted sequence.

Coefficient of transitional moments (C_{TM}) This coefficient is computed as $C_{\text{TM}} = \frac{\text{TM}_{\text{P}}}{\text{TM}_{\text{GT}}}$. It is a ratio of the number of detected transitional moments TM_{P} in the predicted sequence to the number of real ones TM_{GT} in the ground truth. It implicitly reflects the robustness of the workflow recognition method. For example, predicted sequences with a repeated high-frequency noise (e.g. high C_{TM}) point to classification inconsistency.

Application-dependent scores (AD-scores) For some applications a certain time delay in detection, which does not influence overall efficiency, can be acceptable. We propose to re-estimate standard performance scores using acceptable delay thresholds for transition windows. To do so, we redefine the notion of "true positives" TP, replacing it by "application-dependent true positives" TP′. The sample $S \in \text{TP}'$ if $S \in$ TP or if $S \in \text{TW}_{\text{tr}}(i, j)$ and (label$(S) = i$ or label$(S) = j$),

where TW_{tr} is one of the transition windows with the allowed threshold tr, centred on the transitional moment between phases i and j in the ground truth. The AD-accuracy, AD-precision and AD-recall are computed as usual but using TP′ instead of TP.

Results

We used metrics from "Validation metrics" section to validate the performance of our approach through three separate studies using data from "Data" section. In order to estimate the impact of each data type and to get a better understanding of the method capacities we first tested each data type (i.e. visual and instrumental) separately, and then, their combination. The first study (VO) estimated method performances given videos as the single data input. During the second study (IO) instrument usage signals were the only input. The third study (VI) explored the performances of the method given the combination of visual and instrument information. Manual annotations were used as ground truth. The leave-one-surgery-out cross-validation protocol was used.

The same set of tests was run for all studies. The first test estimated the standard accuracy, precision and recall scores using sample-by-sample comparison between the ground truth and predicted sequences. The second test evaluated ATD, the third test measured NL, and the fourth estimated C_{TM}. The results of these tests are presented in Fig. 3 and "First study: visual information only (VO)", "Second study: instrument information only (IO)", "Third study: combination of visual and instrument information (VI)" sections. After that, we re-estimated the standard scores for two possible applications. The first application (A1) is device triggering and relevant information display. It allows a relatively small time delay, so we fixed the transitional delay threshold to 15 s, meaning that all misclassifications caused by a transitional delay of 15 s or less are counted as good classifications. The second application (A2) is estimation of remaining surgical time, where a high accuracy is less important than in other applications. The transitional delay threshold can be fixed at 1 min with no essential impact on the estimation. Obtained AD-scores are presented in Table 2.
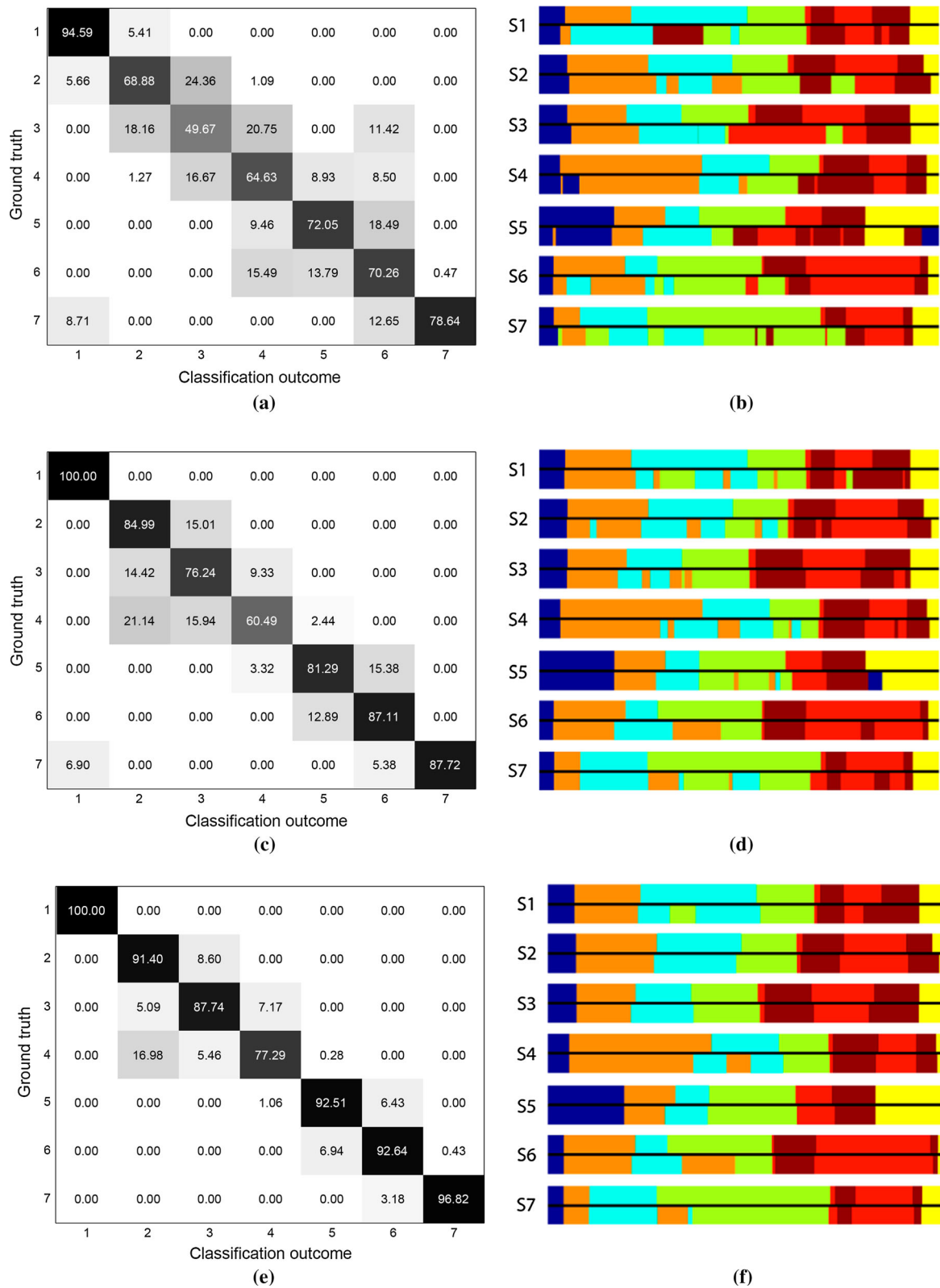
Fig. 3 Global confusion matrices (in %) and visual representations of sequences. The confusion matrices are normalized by the ground truth. For each sequence representation S the ground truth is at the top, and the prediction is at the bottom. **a** Confusion matrix for VO study. **b** Visual sequences of VO study. **c** Confusion matrix for IO study. **d** Visual sequences of IO study. **e** Confusion matrix for VI study. **f** Visual sequences of VI study

Table 2 Application-dependent scores (in %) for all study–application couples

Score	VO (A1)	VO (A2)	IO (A1)	IO (A2)	VI (A1)	VI (A2)
AD-accuracy	71.65	77.96	81.82	84.53	91.59	93.23
AD-precision	75.40	81.00	85.34	87.63	93.54	95.10
AD-recall	74.75	80.34	85.50	87.74	93.70	95.17

Table 3 Standard precision and recall scores (in %) for all conducted studies

Score	Study	P1	P2	P3	P4	P5	P6	P7	Mean ± SD
Precision	VO	81.49	76.68	49.74	67.58	74.47	55.87	98.84	72.10 ± 16.39
	IO	93.88	70.00	66.11	87.33	84.79	80.85	100.00	82.85 ± 12.75
	VI	100.00	76.83	83.69	93.12	93.60	91.05	99.14	91.06 ± 8.29
Recall	VO	94.59	68.88	49.67	64.63	72.05	70.26	78.64	71.25 ± 13.64
	IO	100.00	84.99	76.24	60.49	81.29	87.11	87.72	82.55 ± 12.15
	VI	100.00	91.40	87.74	77.29	92.51	92.64	96.82	91.20 ± 7.28

First study: visual information only (VO)

Standard scores The global accuracy of this method reached 68.10 %. Table 3 shows mean precision and recall scores for all phases and the whole procedure. Mean values were computed from the global confusion matrix (Fig. 3a) constructed by the addition of confusion matrices of all 7 surgeries.

ATD The average transitional delay, computed from all transitional moments of all surgeries, was 1 min 6 s (1648 frames) with a standard deviation of 23 s (572 frames) which resulted in 2.69±0.93 % of the total time for a mean intervention duration of 40 min 48 s.

NL The average noise level in all predicted sequences constituted 16.54 % of all analysed samples, which represented 53.53 % of all misclassifications. Supposing that there are only two types of error, more than a half of all misclassifications were noise and 46.47 % were caused by transitional delays.

C_{TM} Because of the noise presence in all predicted sequences, none of them had the same workflow (i.e. order of phases) that the ground truth. This resulted in a relatively high average C_{TM} equal to 1.70.

Second study: instrument information only (IO)

Standard scores Using surgical tool signals only, the method achieved a global accuracy of 78.95 %, mean precision of 82.85 % and mean recall of 82.55 % (Table 3). Figure 3c illustrates the global confusion matrix.

ATD The average transitional delay was twice lower than in the VO study. ATD was equal to 32 s (804 frames) with a

standard deviation of 20 s (510 frames), i.e. 1.31±0.83 % of the total time.

NL The noise level stayed relatively high (13.17 %). Its part in the overall detection error was 64.18 %. The remaining 35.82 % were due to a transitional delay.

C_{TM} The IO method had a high coefficient of transitional moments equal to 1.95. As in the VO study, no sequence was correctly predicted in terms of phase order.

Third study: combination of visual and instrument information (VI)

Standard scores Thanks to the data combination, our approach gained in predictive capacity and showed a global accuracy of 88.93 %, mean precision of 91.06 % and mean recall of 91.20 % (Table 3). The global confusion matrix is in Fig. 3e.

ATD A slight improvement was made in terms of transitional delay compared to the IO study. The *ATD* decreased to 26 s (643 frames) with a standard deviation of 21 s (535 frames), or 1.05±0.87 % of the total time.

NL The mean noise level in predicted sequences was four times smaller than in the VO study. On average the noise level reached only 4.16 % of the total classifications and constituted 53.21 % of all errors. Thus, the other part of the errors (46.79 %) occurred when transitioning to the next phase too early or too late.

C_{TM} In this study, the predicted sequences had less false transitions. Suchwise, 3 sequences had no noise in them and had a correct phase order. The average C_{TM} decreased to 1.13, showing an enhancement of robustness.

Error analysis

In order to better understand the source of misclassifications we conducted an additional analysis. The results showed that 42.11 % of all errors in the VO study occurred because the HsMM observation vocabulary constructed from the training data did not contain signatures of tested samples. In 51.34 % of the errors test samples from different phases had the same signature. In the IO study, only 4.15 % of all errors were caused by the signature absence and 87.88 % by the signatures identity. For the VI study the error distribution was 78.01 and 18.57 %, respectively. Other errors were probably caused by the HsMM itself, when the phase duration in the test sequence differed a lot from those in the training sequences.

Discussion

Detection of surgical workflow is an important issue requiring a relevant approach and a rigorous validation. In this paper we proposed and validated a novel method dedicated to the real-time segmentation and recognition of surgical phases. The results of the work are discussed below.

Input data impact

A large set of possible signals collected in the OR can be used for workflow detection. Different signal combinations should be investigated. In this work we proposed a method based on visual or/and surgical instrument information. As shown by our experiments, the union of both data types strongly enhances the performance of the method. It allows a better recognition (i.e. higher accuracy), more homogeneous detections (i.e. lower standard deviation) and a better prediction of transitional moments significantly reducing the time delay. This also eliminates a large part of the noise which gives a more correct sequential order. From "Error analysis" section it is clear that information combination enables the construction of descriptions better reflecting the difference between phases (only 18.57 % of error are caused by the descriptions similarity). Video can capture particularities of each phase which can be expressed in terms of features, but the transitional moments are hard to recognize due to a strong resemblance of border frames. That is why the ATD in the VO study is twice bigger compared to the IO study (1 min 6 s vs 32 s). Instrument signals enable a better detection of transitional moments, because the beginning of a new phase is often defined by a change of employed instruments. However, use of identical sets of instruments during different phases leads to false detections. Videos and binary signals of surgical tools usage turned out to be a complementary information correcting weaknesses of each other.

Although the IO method gives better results than VO (78.85 % of accuracy over 68.10 %), a small set of possible instrument combinations and their multiple use in different phases limits its individual application regardless of the chosen classifier. Developing new image features, in contrast, can extend discriminative force of the method. Furthermore, in case of the IO method the phase classification result depends on the instrument information validity. In this work information about instrument use was provided within manual annotation files, meaning the ground truth. For real on-line applications it should come from sensors installed on the instruments, or the surgical tools must be automatically segmented and recognized from images using approaches like [4].

Methodology improvements

The data description process has a major impact on the results. According to our experiments, a global visual description is not enough (giving a precision of 72 % in the VO study) to confidently classify images from videos into surgical phases. In this work we applied only standard visual features allowing fast computation in order to save the generic side of the approach and real-time speed. More elaborate features might improve the recognition capacity. In the case of the IO method, data descriptions in form of instrument vector are not diverse enough to be solely used as input. A large part of classification errors (87.88 %) is related to this factor. Thus, the instrument information should be always complemented with another source of data to create more complex phase descriptions. The overall approach stays fairly flexible to be adapted to a new input data type or data description mechanism. Thereby, other signals from the OR and their combinations should be tested as well. Moreover, an enlargement of the training dataset would reduce classification errors by expanding the observation vocabulary of the HsMM. Also, for off-line applications a part of the noise can be eliminated in post-filtering.

Application to context-aware systems

When developing methods which end-goal is an integrated use inside the OR, defining a sufficient level of accuracy can be problematic. Always being able to perform a perfect segmentation is almost impossible knowing that no definite or objective border between phases exists. The more important question in this case is "What detection delay is acceptable for a particular application?". In our paper we demonstrated that the performance of the approach could be estimated differently depending on the targeted application and existing needs. This shows how close the system is to a real clinical application.

Finally, we would like to emphasize the strongest advantage of this approach: its generality and absence of any a priori knowledge or assumptions. There is no need to segment a particular region of interest or anatomical structure. It requires videos and/or instruments usage signals only, making it purely data-driven. Both can be acquired regardless of the surgery type. The method can be used for on-line detection as it can deliver a classification decision each second in real time. All this allows believing that the method could be applied to any surgical operation as a part of a context-aware system.

Conclusion

In order to better assist the surgical stuff inside the OR, a need of context-aware systems, accurately identifying surgical workflow, appeared. The method proposed in this paper automatically segments and recognizes surgical phases from videos and instrument usage signals in real time. It was validated on endoscopic data using standard performance scores and novel metrics measuring detection time delay, noise and coefficient of transitional moments. With visual information only the system reached 68.10 % of accuracy, whereas instrument usage signals provided 78.95 %. The accuracy increased to 88.93 % when using a combination of both signals as input. That shows a strong need to combine input signals in order to improve method performances.

Compliance with ethical standards

Conflict of interest Olga Dergachyova, David Bouget, Arnaud Huaulmé, Xavier Morandi and Pierre Jannin declare that they have no conflict of interest.

Ethical approval For this type of study formal consent is not required.

Informed consent Not required. Used data were anonymously available through the MICCAI EndoVis challenge.

References

1. Bardram J, Doryab A, Jensen R, Lange P, Nielsen K, Petersen S (2011) Phase recognition during surgical procedures using embedded and body-worn sensors. In: IEEE international conference on pervasive computing and communications, pp 45–53
2. Bharathan R, Aggarwal R, Darzi A (2013) Operating room of the future. Best Pract Res Clin Obst Gynecol 27(3):311–322
3. Blum T, Feuner H, Navab N (2010) Modeling and segmentation of surgical workflow from laparoscopic video. In: Medical image computing and computer-assisted interventions, vol. 6363, pp 400–407
4. Bouget D, Benenson R, Omran M, Riffaud L, Schiele B, Jannin P (2015) Detecting surgical tools by modelling local appearance and global shape. IEEE Trans Med Imaging 34(12):2603–2617
5. Charriere K, Quellec G, Lamard M, Coatrieux G, Cochener B, Cazuguel G (2014) Automated surgical step recognition in normalized cataract surgery videos. In: IEEE international conference on engineering in medicine and biology society, pp 4647–4650
6. Cleary K, Kinsella A (2005) Or 2020: the operating room of the future. J Laparoendosc Adv Surg Tech 15(5):495–497
7. Despinoy F, Bouget D, Forestier G, Penet C, Zemiti N, Poignet P, Jannin P (2015) Unsupervised trajectory segmentation for surgical gesture recognition in robotic training. IEEE transactions on biomedical engineering. doi:10.1109/TBME.2015.2493100
8. Holden MS, Ungi T, Sargent D, McGraw RC, Chen EC, Ganapathy S, Peters TM, Fichtinger G (2014) Feasibility of real-time workflow segmentation for tracked needle interventions. IEEE Trans Biomed Eng 61(6):1720–1728
9. James A, Vieira D, Lo B, Darzi A, Yang GZ (2007) Eye-gaze driven surgical workflow segmentation. In: medical image computing and computer-assisted interventions, pp 110–117
10. Jannin P, Morandi X (2007) Surgical models for computer-assisted neurosurgery. NeuroImage 37(3):783–791
11. Lalys F, Jannin P (2014) Surgical process modelling: a review. Int J Comput Assist Radiol Surg 9(3):495–511
12. Lalys F, Riffaud L, Bouget D, Jannin P (2012) A framework for the recognition of high-level surgical tasks from video images for cataract surgeries. IEEE Trans Biomed Eng 59(4):966–976
13. Lo B, Darzi A, Yang GZ (2003) Episode classification for the analysis of tissue/instrument interaction with multiple visual cues. In: Medical image computing and computer-assisted interventions, vol 2878, pp 230–237
14. Nara A, Izumi K, Iseki H, Suzuki T, Nambu K, Sakurai Y (2011) Surgical workflow monitoring based on trajectory data mining. In: New frontiers in artificial intelligence, vol 6797, pp 283–291
15. Padoy N, Blum T, Essa I, Feussner H, Berger MO, Navab N (2007) A boosted segmentation method for surgical workflow analysis. In: Medical image computing and computer-assisted interventions, vol 4791, pp 102–109
16. Padoy N, Blum T, Ahmadi SA, Feussner H, Berger MO, Navab N (2012) Statistical modeling and recognition of surgical workflow. Med Image Anal 16(3):632–641
17. Quellec G, Lamard M, Cochener B, Cazuguel G (2014) Real-time segmentation and recognition of surgical tasks in cataract surgery videos. IEEE Trans Med Imaging 33(12):2352–2360
18. Rabiner LR (1989) A tutorial on hidden Markov models and selected applications in speech recognition. Proc IEEE 77(2):257–286
19. Schapire RE (2003) The boosting approach to machine learning: An overview. In: Nonlinear estimation and classification, pp 149–171
20. Stauder R, Okur A, Navab N (2014) Detecting and analyzing the surgical workflow to aid human and robotic scrub nurses. In: The Hamlyn Symposium on Medical Robotics, p 91
21. Weede O, Dittrich F, Worn H, Jensen B, Knoll A, Wilhelm D, Kranzfelder M, Schneider A, Feussner H (2012) Workflow analysis and surgical phase recognition in minimally invasive surgery. In: IEEE international conference on robotics and biomimetics, pp 1080–1074
22. Yu SZ, Kobayashi H (2003) An efficient forward-backward algorithm for an explicit-duration hidden markov model. IEEE Signal Process Lett 10(1):11–14