Artificial Intelligence in Medicine 74 (2016) 9–20

Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

journal homepage: www.elsevier.com/locate/aiim

Web-video-mining-supported workflow modeling for laparoscopic surgeries

Rui Liu [a], Xiaoli Zhang [a,*], Hao Zhang [b]

[a] Department of Mechanical Engineering, Colorado School of Mines, Golden, CO 80401, USA
[b] Department of Electrical Engineering & Computer Science, Colorado School of Mines, Golden, CO 80401,USA

ARTICLE INFO

Article history:
Received 24 March 2016
Received in revised form
11 November 2016
Accepted 13 November 2016

Keywords:
Topic modeling
Sentiment analysis
Web video mining
Surgical workflow modeling
Laparoscopic surgery

ABSTRACT

Motivation: As quality assurance is of strong concern in advanced surgeries, intelligent surgical systems are expected to have knowledge such as the knowledge of the surgical workflow model (SWM) to support their intuitive cooperation with surgeons. For generating a robust and reliable SWM, a large amount of training data is required. However, training data collected by physically recording surgery operations is often limited and data collection is time-consuming and labor-intensive, severely influencing knowledge scalability of the surgical systems.
Objective: The objective of this research is to solve the knowledge scalability problem in surgical workflow modeling with a low cost and labor efficient way.
Methods: A novel web-video-mining-supported surgical workflow modeling (webSWM) method is developed. A novel video quality analysis method based on topic analysis and sentiment analysis techniques is developed to select high-quality videos from abundant and noisy web videos. A statistical learning method is then used to build the workflow model based on the selected videos. To test the effectiveness of the webSWM method, 250 web videos were mined to generate a surgical workflow for the robotic cholecystectomy surgery. The generated workflow was evaluated by 4 web-retrieved videos and 4 operation-room-recorded videos, respectively.
Results: The evaluation results (video selection consistency n-index $\geq$0.60; surgical workflow matching degree $\geq$0.84) proved the effectiveness of the webSWM method in generating robust and reliable SWM knowledge by mining web videos.
Conclusion: With the webSWM method, abundant web videos were selected and a reliable SWM was modeled in a short time with low labor cost. Satisfied performances in mining web videos and learning surgery-related knowledge show that the webSWM method is promising in scaling knowledge for intelligent surgical systems.

1. Introduction

1.1. Laparoscopic surgery & surgical workflow modeling

The increase of technological complexity in surgery as well as the constant concerns for quality assurance have created an urgent need for intelligent surgical agents, such as surgical robots [1]. A surgical robot is expected to adapt itself to the on-going circumstances in the operation room (OR) and appropriately cooperate with the surgeons [1,2]. To develop such intelligent surgical systems, the key step is endowing them with reliable knowledge such as knowledge of a surgical workflow model (SWM), which can be used to analyze and evaluate surgery operations and also be called as surgical process modeling (SPM). SWMs have the potentials to support clinical applications, ranging from detecting abnormal operations to reducing intraoperative errors [3,4], predicting next surgical phase [5,6] and triggering surgery-related reminders [7,8], improving surgeons' motion ability [9,10], assisting the surgeons' decision making [1,11], documenting surgery procedures [12], and delivering the needed instruments [13,14].

Efforts have been made on building reliable SWMs by analysing the surgical procedures and their internal correlations. Some researchers identified a single surgery procedure by tracking the surgical instrument usage. They proved that the instrument usage was a valuable cue in understanding surgical workflow since specific tools were usually involved in specific surgical phases [15,16].

* Corresponding author at: BB350, 1500 Illinois St., Golden, CO, USA.
E-mail address: xlzhang@mines.edu (X. Zhang).

http://dx.doi.org/10.1016/j.artmed.2016.11.002
0933-3657/

Researchers modeled a typical surgical phase sequence for a specific surgery by merging different surgery procedures with a dynamic time warping (DTW) method [17]. To obtain comprehensive understanding about a surgery, the semantic meaning of surgical phase and phase transitions were interpreted from the low-level sensor data including instruments' working statuses [18], surgeons' motion trajectory/gestures/locations [18], and surgical instrument involvements [19,20].

1.2. Limitations of existed workflow modeling methods

Surgical workflow modeling in real surgical procedures is challenging, due to high complexity and variability of surgery methods, surgery procedures, surgical instruments, surgeons' operational styles, and patient-specific properties in anatomy and pathology [16]. Due to these complexity and variety, a sufficient amount of training data is commonly required to gain accurate SWM knowledge for a given type surgery. Unfortunately, although generating SWMs from surgery demonstrations has been shown to be a feasible solution, current methods, which collect the training data by physically recording surgery operations, are suffering the data shortage. This is because in practical situations, the training data that can be collected from physical operations is relatively limited by time consumption and labor cost. This paucity of training data limits the scalability of SWM knowledge of the intelligent surgical systems. In addition, the cost of time and labor in collecting large amount of training data is high [21,22].

1.3. Web-video-mining-supported surgical workflow modeling

To scale up SWM knowledge and at the same time reduce the time and labor costs in the SWM generation, we developed a novel web-video-mining-supported surgical workflow modeling (webSWM) method, shown in Fig. 1. Our contributions in this paper include:

- An automatic video-mining method is developed to retrieve abundant highly-qualified web videos as the training data, with which a SWM for a specific surgery can be generated. To retrieve qualified videos from the enormous web pool, three standards, including video production quality, video content matching degree and surgeon skill, are set to evaluate videos' qualification degree. The retrieved videos are ranked according to these standards, and the top-scored videos are selected as training data for surgical workflow modeling.
- A reliable SWM is generated to assist a surgical system with surgery understanding. Transitions among different surgery phases are modeled in a probability manner. SWM-related knowledge, such as surgery phase segmentations and instrument-phase correlations, is extracted. This reliable and valuable knowledge is learned to support an intelligent surgical system in performing tasks such as abnormal surgery operation detection and instrument recommendation. Given that the amount of web surgery videos available is enormous, surgery operation manners in the videos are various, and many web videos are free to access, it is feasible for an intelligent surgical system to learn reliable SWM knowledge in an economic way.

1.4. Related work

Some work has been done to overcome the challenges in data scarcity and time/labor intensity.

To enrich the information data for surgical workflow modeling, various types of information sources have been explored. Digital documents, including surgery videos [21–23], surgical instrument/system logs [24–26], and communication records in a clinical team [27,28], were involved to analyze the correlations among surgery phases, instruments, and surgeons. Surgeons' professional experiences, such as surgery methods, tool usages, and surgical ontology were extracted from their oral descriptions [29,30].

To reduce the time/labor cost, one strategy is the replacement of the observer-based data collection method with the sensor-based data collection method [31]. The observer-based method, which means a human observer is involved to collect surgical data manually, is usually conducted selectively for limited targeted surgeries, with high time/labor costs [31–33]. The sensor-based method, which means different types of sensors such as RFID sensors are used to collect surgical data automatically, is usually integrated with surgical systems such as surgical robots to record all the surgeries, with low time/labor costs [31,34,35]. Another strategy is the crowdsourcing-based data collection method. Surgical information, such as surgeons' professional skill levels and surgical tool usages, is collected by using paper/electronic surveys or the online crowdsourcing platform Amazon Mechanical Turk [36–38]. Given the price of the survey is usually less than 5 dollars per hour and the time cost is about 1–2 weeks for each set of survey, the

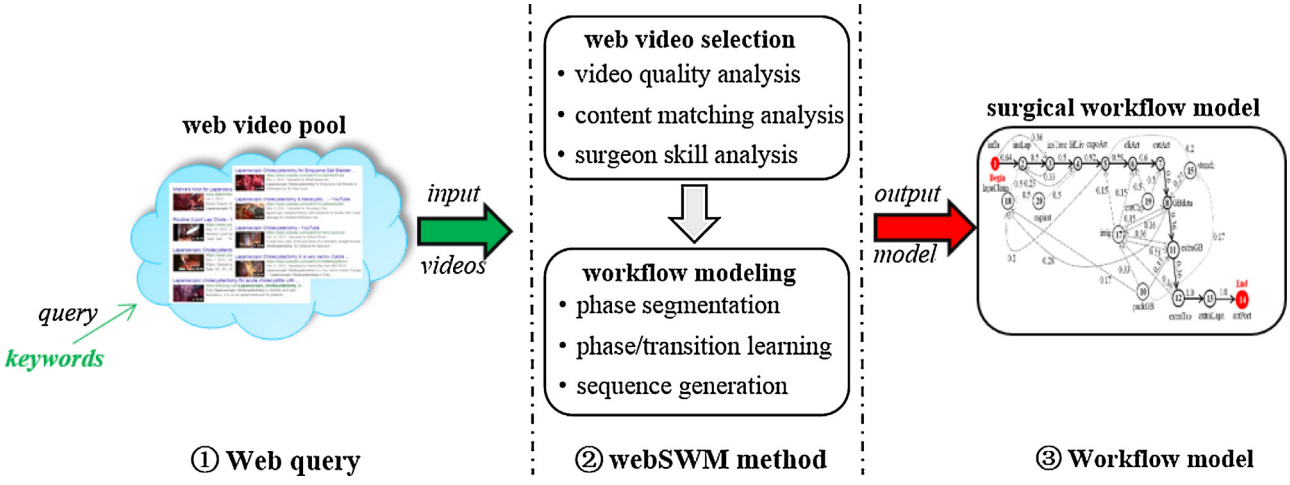

Fig. 1. Overview of the webSWM method. The inputs are web videos and the output is a SWM for a targeted surgery. The webSWM includes two main parts: 1) web video selection and 2) workflow modeling. In 1), surgery videos and video-related information are collected by querying the web with the keywords of the targeted surgery. Then the qualification degree for each video is calculated and the qualified videos were selected. In 2), a probabilistic graphical structure is adopted to describe the surgery phase involvements and their transitions.

crowdsourcing-based method can reduce the time/labor cost of data collection to some degree.

The aforementioned methods could enrich the information sources and meanwhile reduce the time/labor cost for surgical data collection. However, these methods did not solve the scalability problem satisfactorily. First, data in all these methods are constrained by source availability and variety. It is challenging to acquire specific types of data from the available expertise-qualified workers. It is also challenging to get the desired variety of data, given the digital documents in the abovementioned methods are usually recorded in one or several medical institutes in one region, in which the surgery processes are standardized. Second, none of these methods evaluates the data's quality automatically. Human manual evaluation is necessarily involved, increasing the time/labor cost. Therefore, the existed solutions are still expensive and time-consuming, limiting data enrichment and knowledge extraction.

With current technologies introduced in [38,39], which use high-technology systems such as Google Glass, many surgery videos could be collected with relatively low time/labor costs. However, for videos recorded by one hospital or hospitals in certain areas, their operation variety is limited. One of our focuses in the webSWM method is to increase the data variety from a higher number of surgical procedures, which could be reflected by videos contributed by different surgeons all over the world.

In the recent decade, open information extraction (OIE) techniques have been developed and widely used in robotics [40–42]. For example, by processing natural descriptions on web, people's daily experiences were extracted as common sense to support robots' interactions with a human [43–46]. Thus far, the OIE techniques are mainly implemented on home-serving and industrial robots. For medical robots, OIE still hasn't been widely studied. Since that many world-wide surgeons share their professional experiences in the form of surgery video tutorials on the web [47,48], the OIE techniques for automatic, scalable, and low-cost SWM generation are in urgent need.

2. Materials and methods

2.1. Web video selection

Collecting video candidates by web query. To collect video candidates from web, a keyword-based query strategy is used. Given the expressions for describing the same surgery are typically various, different keywords are used. Take "Cholecystectomy Surgery" for example, in the video titles this can be described by {"cholecystectomy surgery", "robotic laparoscopic cholecystectomy surgery", "laparoscopic cholecystectomy", ...}. These keywords are predefined by surgeons and are saved in a local database. For each keyword expression; a query is launched in google search engine [49]. Seven types of video-related information; such as {publication date; length; resolution; title; like count; dislike count; reviewers' comment}; are retrieved. Given that the videos returned by the search engine have already been ranked according to their content relevance; we assume that the top-ranking videos are more likely to be the videos of interest. Therefore; we only collect the top-ranking videos as video candidates.

Video Production Quality Analysis. The most important aspect of video qualification is the video quality ψ_1 defined by the publication date φ_1, length φ_2 and resolution φ_3, shown in Eq. (1) where w is the feature weight. If a video is too outdated or too short, the video may not provide valuable information for workflow modeling. If the video has a low degree of image resolution, computer vision techniques cannot identify the surgery phase and involved instruments. To support video quality analysis, information retrieval tool Pafy [50] is used. The weight for video quality in video qualification evaluation is W_1. Detailed learning process will be described in Section 3.1.

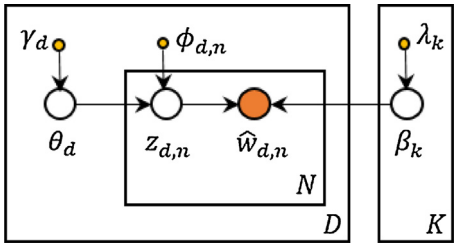

Fig. 2. Plate diagram for a LDA-based topic model.

$$\psi_1 = \Sigma_{i=1}^{3} w_i \varphi_i \tag{1}$$

Video content matching analysis. We considered each video's title as a topic describing the content of the corresponding video. To select the videos with a high content matching degree, we adopt the topic modeling technique. Given that all the videos are collected for one targeted surgery, the retrieved video titles are classified into two topic types: targeted-surgery-relevant topic (sur-Rele) and targeted-surgery-irrelevant topic (sur-Irre). The sur-Rele titles describe the same type of surgery. Therefore they are more likely to be clustered together than being clustered with other sur-Irre titles. Topic modeling is conducted by using the Latent Dirichlet allocation (LDA) algorithm in an unsupervised learning manner. LDA is a generative model that is good at evaluating the mutual similarities among observations and widely used in unsupervised topic modeling [51]. The LDA-based topic modeling process is expressed by Eq. (2) and Fig. 2. K denotes the total topic number and in this paper K = 2 (sur-Rele and sur-Irre). D denotes the total number of involved titles. N denotes the total involved unique words in all the D titles. The video topics are denoted by $\beta_{1:K}$ where K is equal to 2 and each topic β_k is a probabilistic distribution over the title vocabularies. The video topic proportion for the d_{th} title are θ_d, where $\theta_{d:k}$ is the topic proportion for topic k in title d. The topic assignments for the d_{th} title are z_d, where $z_{d,n}$ is the topic assignment for the n_{th} word in title d. The words for title d are $\hat{w}_d$ where $\hat{w}_{d,n}$ is the n_{th} word in the title d. γ_d is the topic proportion on each video title. λ_k is the overall topic distribution on the whole title corpus. $\phi_{d,n}$ is the topic distribution parameter on word $\hat{w}_{d,n}$. The unsupervised learning process is shown in Algorithm 1 and the detailed process of unsupervised topic modeling was introduced in [52,53]. Two clusters were generated. Based on the keyword features, topic types of the two clusters were identified as {sur-Rele, sur-Irre}. Then based on the topic type, the video content matching degree ψ_2 and weight W_2 were defined. Specific topic weight calculating methods are described in Section 3.1.

$$p\left(\beta_{1:K}, \theta_{1:D}, z_{1:D}, \hat{w}_{1:D}\right) = \Pi_{i=1}^{K} \mathrm{p}\left(\beta_i\right) \Pi_{d=1}^{D} \mathrm{p}\left(\theta_d\right) \left(\Pi_{n=1}^{N} \mathrm{p}\left(z_{d,n}|\theta_d\right) \mathrm{p}(\hat{w}_{d,n}|\beta_{1:K}, z_{d,n})\right) \tag{2}$$

Surgeon skill evaluation. As SWMs are used in highly risky environments, their accuracies are critical. Therefore, not only should video quality and video content be considered, but also the surgeon's professional skill level. Typically, surgeon skill evaluation is done by experts manually. However, as the number of web videos could be potentially large, it is impractical to conduct manual evaluation. A new reliable and efficient method is needed. In this paper, we use a sentiment analysis method for automatic surgeon skill evaluation. Based on video comments, layman reviewers' sentiments in watching the surgeons' operations are analyzed. It has

been validated that when reviewers give positive comments, the surgeon's performance is satisfying; when the reviewers give negative comments, the surgeon's performance is unsatisfying [36–38]. Given that a video could have hundreds of comments, we only collect the top-ranking comments. The support-vector-machine (SVM)-based semi-supervised learning algorithm [54] is adopted for video sentiment analysis. To initially train the classifier for sentiment analysis, the Stanford movie review dataset [55] is adopted. A binary SVM is used, shown in Eqs. (3) and (4). In Eq. (3), w_{svm} is the linear combination weight of input data (reviews' comments), ξ_i is the acceptable classification error, and C is the tradeoff parameter between the error and margin. After SVM optimization with a QP solver, the optimal combination parameter w^*_{svm} and the intercept b^* of the optimal hyperplane are solved. With Eq. (4), the sentiment for each comment is classified. We defined that when the discrimination value is negative, the sentiment label for a sentence is negative, otherwise the sentiment label is positive. The sentiment analysis process is described in Algorithm 2. When comments for a video are existing ($M \neq 0$), we use comments to analyze video sentiment; else ($M = 0$) we use the dislike/like hit number instead. After each analysis, the training set is updated to train a new classifier. When more than 50% percent of the reviewer comments for a video are positive, the overall result for the surgeon skill evaluation is positive, otherwise the overall result is negative. The sentiment factor ψ_3 and weights W_3 are defined according to the video's sentiment label.

$$\min \ \frac{1}{2}||w_{svm}||^2 + C\Sigma_{i=1}^{K}\xi_i \tag{3}$$

$$f(x) = sign\left\{{w^*_{svm}}^T \cdot x + b^*\right\} \tag{4}$$

Overall video qualification degree. The overall video qualification degree ψ_0 is defined as Eq. (5). The video production quality ψ_1 is the precondition of content matching ψ_2 and surgeon skill level ψ_3. Only videos meeting requirements of production quality could be considered as being informative for workflow modeling. According to specific application requirements, a threshold value will be set to select the qualified videos.

$$\psi_0 = W_1\psi_1\left(W_2\psi_2 + W_3\psi_3\right) \tag{5}$$

$$w_{lk}(V_2) \longleftarrow w_{lk}(V_1) + \mu\Delta w_{lk}(V_1, V_2) \tag{6}$$

2.2. Workflow modeling

The goal of webSWM is to build a robust and reliable SWM model from web videos. With a steepest-descent algorithm [56,57], which is good at finding the optimal transition path along the temporal direction, the transition probabilities of surgical phases are learned, and the most probable temporal phase sequence along a 'steepest-descent direction' can be generated.

For surgical phase segmentation, given the instrument recognition is technically-ready [58,59], we did not implement computer vision techniques to detect the surgical instrument/phase since the focus of this paper is the feasibility evaluation of modeling workflow by mining web videos. In this paper, we recruited a surgical team composed of surgeons and residents to manually segment the selected videos into predefined surgical phases. Details about the phases were introduced in Section 3.2.

The phase occurrence probability is calculated as the ratio of a phase's occurrence frequency to all the phases' occurrence frequencies in all training samples. The phase transition probability is calculated as the ratio of a phase's one type of transition frequency to this phase's all possible transition frequencies. As the dynamic involvement of the training videos, the phase occurrence and transition probabilities are dynamically updated. The updating process for learning the transition probability is shown in Eq. (6), where w_{lk} denotes the transition probability from phase q_l to phase q_k, V_1 denotes the training sample set, μ is the updating step, and Δw_{lk} is the transition probability adjustment based on the old training set V_1 and the new training set V_2. The instrument-phase correlation and the phase occurrence probability are updated in a similar manner.

The objective function for learning the temporal sequence (workflow) Q is shown in Eq. (7), where 'Begin' denotes the first phase and 'End' denotes the final phase of a surgical phase sequence, respectively. $Q_{Begin-End}$ denotes the temporal main sequence, which starts from 'Begin' and ends at 'End'. q_k denotes phase k. A phase set $\{q_1, q_2, \ldots, q_K\}$ is generated from the training sample set V_1, and K is the total number of phases. d denotes a possible phase sequence between 'Begin' and 'End' and D denotes all likely sequences embedded in the phase set $\{q_1, q_2, \ldots, q_K\}$. e_k denotes the occurrence probability of phase k, and Z is the normalization parameter that is defined by Eq. (8), and Q_1 denotes the existing main temporal sequence based on the old training set V_1. We define 'Begin' as the important phase that merely launches transitions to other phases without receiving any transition, and 'End' as the important phase that merely receives transitions without launching. With Eq. (7), likelihoods for all the possible sequences starting from 'Begin' and ending at 'End' are calculated and the sequence with the maximum likelihood is considered as the temporal main sequence. The basic idea for sequence learning is exhaustive searching for the most likely sequence path with the highest probability. Based on the searching function in Eq. (7) where the denominator is the product of the main phases' probabilities in a selected sequence, the higher percentage of the main phases is contained in the selected sequence, the greater is the sequence likelihood. When a sequence is the main sequence, the likelihood is the highest (1.00). To generate a complete SWM, we also integrate local main sequences, which are launched by phases with a relatively high occurrence probability but are excluded by the global main sequence.

The sequence matching degree is defined as the portion of given testing sequences whose main sequences are matched with the learned main sequence. When the matching degree is smaller than threshold ∂_0, the surgical workflow is considered as 'abnormal'. The instrument-usage consistency is defined as the portion of the phases whose instruments are correctly defined by the learned SWM-related knowledge.

3. Experiments and results

To evaluate the effectiveness of our webSWM method in generating SWM from web videos, experiments were conducted on modeling the SWM for the robotic cholecystectomy surgery, a widely adopted procedure for gallbladder removal. For web query, the experts defined 5 keyword expressions {"robotic cholecystectomy", "cholecystectomy surgery", "robotic laparoscopic cholecystectomy surgery", "laparoscopic cholecystectomy"}. For each query, we collected 50 top ranking videos (250 videos for the total). After automatically removing the repeated ones, we ended with 185 video candidates.

$$Q_{Begin-End}(q_1, q_2, \ldots, q_K) = argmax_{d\in D}\left(\Pi_{q_l,q_k\in d} e_k w_{lk}\right)/Z \tag{7}$$

$$Z = \Pi_{q_l,q_k\in Q_1} e_k w_{lk} \tag{8}$$

With experiments based on the 185 videos, we try to evaluate (1) the effectiveness of the webSWM method in automatically retrieving highly-qualified targeted-surgery-relevant videos from the web, (2) the accuracy of the generated surgical workflow and workflow-related knowledge, and (3) the effectiveness in saving time, money and labor in SWM generation.

Table 1
Evaluation of topic modeling.

	precision		recall		F1-score		title number
	LDA	LSI	LDA	LSI	LDA	LSI	
targeted-surgery-relevant	0.86	0.85	0.52	0.55	0.69	0.70	122
targeted-surgery-irrelevant	0.62	0.68	0.26	0.38	0.44	0.53	63
average	0.74	0.77	0.39	0.47	0.57	0.62	185

3.1. Evaluation of method effectiveness in videos mining

The 185 video candidates were ranked based on the video qualification degree calculated by Eq. (5). In the video quality analysis, publication date, length and resolution were equally important. Therefore the weights w_i(i = 1,2,3) were 0.33. The motivation for considering publication date is to filter out the outdated videos which highly-likely contain unreliable surgical information. With the suggestions from surgical experts, we set the time range as {past ~2005.01.01 (unreliable), 2005.01.01–2010.01.01 (median-reliable), 2010.01.01 ~ now (reliable)}. Given the date is only one of the features deciding the overall video qualification degree, using a threshold date could achieve the expectation on video filtering. Feature values φ_i(i = 1,2,3) were defined as publication date{ >=2010.01.01, φ_1 = 1.0. <2010.01.01 & >=2005.01.01, φ_1 = 0.5. <2005, φ_1 = 0.0}, length {> = 600 s, φ_2 = 1.0. <600 s & >=300 s, φ_2 = 0.5. <300 s, φ_2 = 0.0}, and resolution {>320 × 240, φ_3 = 1.0. <=320 × 240, φ_3 = 0.0}. In content matching analysis, the video with the sur-Rele topic had the content matching degree 1, and the video with sur-Irre topic had the content matching degree 0. In sentiment analysis, when the video sentiment was positive, ψ_3 was defined as 1, and when the video sentiment was negative, ψ_3 was defined as 0. The weight of the video quality W_1 was defined as 1.0. Video content matching and reviewer sentiment were considered to be equally important Therefore both their weights, W_1 and W_2, were defined as 0.5. With the threshold of the overall qualification degree defined as 0.7, 35 videos with the qualification degree greater than the threshold were selected as sample videos to learn and evaluate the SWM.

To prove the reliability of the topic modeling method in analysing videos' content matching degree, we used the latent semantic indexing (LSI) algorithm as the baseline. To prove the reliability of the sentiment analysis method in analysing surgeons' professional skill, we used the Naïve Bayesian network (NB) algorithm as the baseline. LSI and NB are widely used in topic modeling and sentiment analysis, respectively [60]. For evaluations, the precision, recall, and F1 scores (defined as the average of precision and recall) for all algorithms were calculated. To evaluate the effectiveness of webSWM in video selection, we asked a domain expert team (surgeons and residents) to select 35 most-qualified surgery videos from the 185 video candidates.

For topic analysis on both LDA and LSI algorithms, shown in Table 1, the average precision is greater than 0.74, average recall is greater than 0.39, and the F1 score is greater than 0.57. This F1 score performance is in the typical topic-modeling range 0.50–0.80 [61], showing that this method is effective in distinguishing the sur-Rele topics from the sur-Irre topics for the 185 video candidates. Even though only a relatively-small proportion of topics (52%) could be correctly involved (denoted by recall 0.52), most of the involved topics (86%) could be correctly predicted (denoted by precision 0.86), making the top-ranking videos reliable. Given we only select the top-ranking videos, this topic-modeling performance is sufficient for our video-based surgical knowledge collection. Semantic features for the two video topics {sur-Rele, sur-Irre} are shown in Fig. 3. For sur-Rele, the typical features are {"cholecystectomy", "laparoscopic", "surgery", "gallbladder", "removal", "gall", "difficult", "robotic", "case", "video"}, which are consistent with our expression manners and therefore are reasonable.

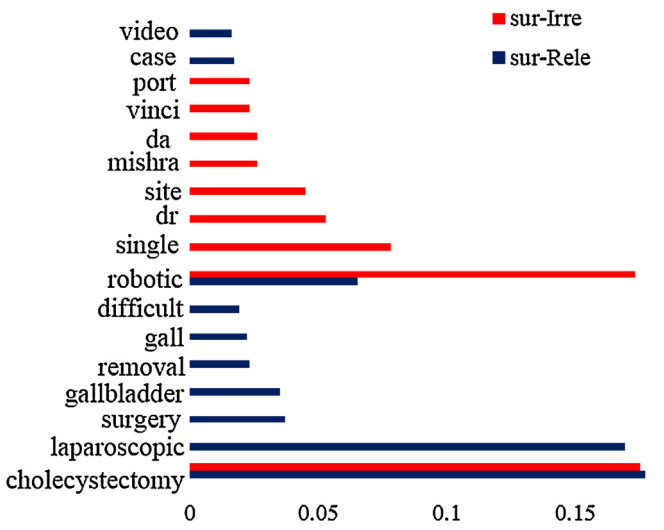

Fig. 3. Semantic features for topics. sur-Irre: targeted-surgery-irrelevant topic. sur-Rele: targeted-surgery-relevant topic. Vertical axis denotes semantic feature and horizontal axis denotes the feature weight given a topic.

Table 2
Evaluation of reviewer sentiment analysis.

	precision	recall	F1-score
SVM	0.63	0.82	0.73
NB	0.71	0.68	0.69
average	0.67	0.75	0.71

*neg: negative comment; pos: positive comment.

Table 3
Comment samples for sentiment analysis.

Comment	Label	cmt num
'There may be another problem in addition to the gall stones'	neg	388
'So elegantly done! Thank you!'	pos	513

neg: negative comment; pos: positive comment. cmt num: comment number.

In surgeon skill evaluation based on sentiment analysis, the initial classifier was trained by Stanford movie review and then as the involvements of surgeon-related comments, the classifier was updated and meanwhile new comments were labeled. To evaluate the accuracy of the sentiment analysis conducted by the webSWM, the reviewers' comments were also labeled by two volunteers manually. Taking the volunteer-labeled comments as the baseline, the webSWM-labeled comments were evaluated. The evaluation results are shown in Table 2, both the SVM-supported classifier and NB-supported classifier achieve an average precision of 0.67, an average recall of 0.75, and an average F1 score of 0.71. Results show that our method could effectively evaluate the surgeons' professional skills by analysing viewers' sentiments in watching the surgery operations. The sentiment analysis sample is shown in Table 3. In total there are 388 negative comments and 513 positive comments involved in the sentiment analysis.

For video selection, we employed two expert volunteers to rank all the 185 video candidates and selected the top N videos according to the qualification standards defined in the webSWM method. The two experts worked together to make the agreement on a single ranking result. A n-index factor is defined to evaluate the consistency of selected videos between the webSWM method and the manual selection method. n-index means the proportion of the top n webSWM-selected videos that are included by the top n expert-

Table 4
Evaluation of webSWM's performance in video selection.

n	average n-index	
	webSWM	Baseline
35	0.60	0.51
70	0.74	0.77
92	0.88	0.91

n-index: the proportion of the top n machine-selected videos that are included by the expert-selected n videos.

selected videos. For example, when n-index is equal to 60% at n = 35, it means in the 35 webSWM-selected videos, 60% of the videos (21 videos) are also selected by the experts. When n was 70 n-index was 74%, and when n was 92 n-index was 88%, as shown in Table 4. The relatively high consistency of video selection proved that our webSWM method could effectively retrieve the highly-qualified web videos.

The baseline for the webSWM method is the method based on LSI and NB (details are in Section 3.1 paragraph 2). As Table 4 Baseline shows, the performance of the webSWM is consistent with that of the baseline method, proving that the effectiveness in video selection is

brought by the method which is consistent with production quality analysis, content matching analysis and surgeon skill evaluation, instead of being brought by the adopted algorithms. When n increases the n-index increases, indicating that webSWM' video selections are becoming more consistent with experts' selections when the standards are relaxed.

3.2. Evaluation of the effectiveness of the webSWM method in workflow generation

The mean duration time for the 158 videos are 11.75 min. Web videos usually are edited therefore incomplete with some procedures removed. To extract reliable information from the piecemeal videos, agreement-confirmed, observation-based, and statistical-learning-supported principles are made in the video segmentation procedure. The agreement-confirmed principle means two volunteers worked together that only the surgery phases identified by both volunteers were counted. The observation-based principle means only the phase occurrences/transitions that were observed in the videos were counted. The phase occurrences/transitions, which were logically involved but cannot directly observed from videos, were ignored. The statistical-learning-supported principle means the phase occurrence/transition frequencies were recorded from the collected incomplete videos in a statistical learning manner. Then the phase occurrence/transition probabilities were calculated according to the method described in Section 2.2. To segment the 35 selected surgery videos, 20 possible phases were predefined by two volunteers, as shown in Table 5. Based on these predefined phases, each video was segmented, shown in Fig. 4. From these videos, we learned the SWM model in each iteration of the cross-validation (Please see Section 3.3 for details). The phase occurrence probabilities in one sample iteration (with first 31 videos) are shown in Fig. 5. The phases {GBdeta2, hemos} with the occurrence probabilities less than 0.01 were ignored.

Table 5
General surgical phases in cholecystectomy.

No.	Surgical Phases
1	CO2 inflation (infla)
2	Inserting a laparoscope (insLap)
3	Inserting a trocar (insTroc)
4	Lifting liver (lifLiv)
5	Exposing artery (expoArt)
6	Clipping artery (cliArt)
7	Cutting artery (cutArt)
8	Gallbladder(GB) detaching (GBdeta)
9	HF cautery-based GB detaching (GBdeta2)
10	Packaging GB (packGB)
11	Extracting GB (extraGB)
12	Extracting trocar (extraTro)
13	Extracting laparoscope (extraLapa)
14	Suturing ports (sutPort)
15	Stanching (stanch)
16	Applying a hemostatic (hemos)
17	Irrigating (irrig)
18	Extracting laparoscope for cleaning (lapaClean)
19	Cutting clip (cutClip)
20	Aspiration of GB (aspirat)

Fig. 4. Phase segmentation.

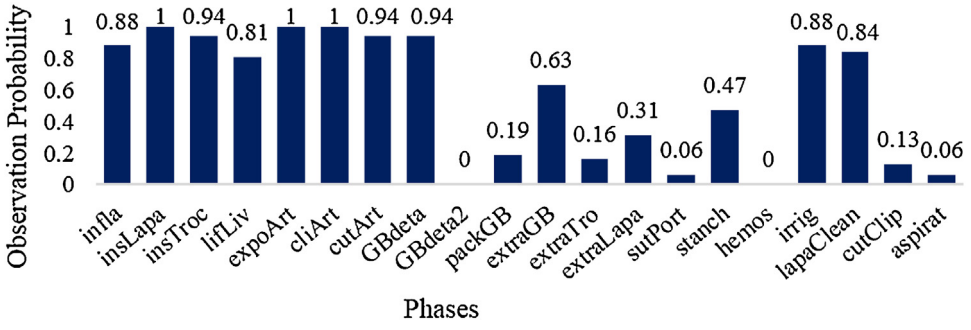

Fig. 5. Phase occurrence probabilities.

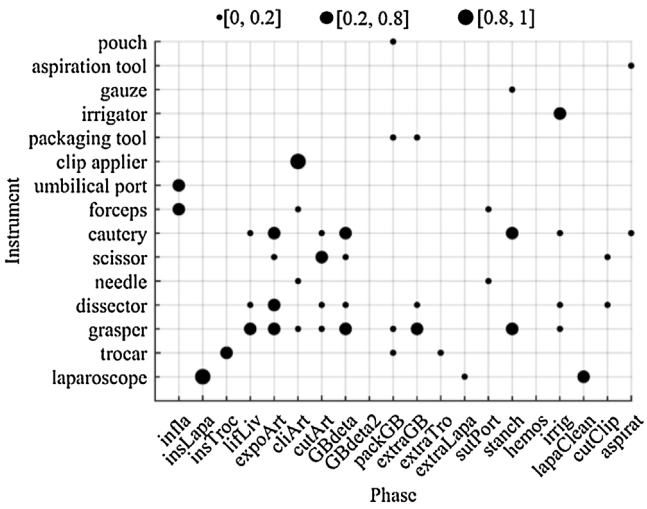

Fig. 6. Phase-instrument correlation.

The phase-instrument correlations for the sample iteration are shown in Fig. 6. Each column lists the involved instruments for a particular phase. The involved instruments are represented as dark dots. The size of dots indicates the correlation intensity. As shown in Fig. 6, most phases own unique combinations of instrument involvements, providing valuable evidences for automatic phase recognition in future SWM-based applications.

The phase transition probabilities for the sample iteration are shown in Fig. 7. In the occasion with 34 training samples, 'infla' was determined as the 'Begin' phase and 'sutPort' was selected as the 'End' phase. The temporal main sequence for robotic cholecystectomy in the sample iteration was generated as shown in Fig. 8. The global main sequence, which is the most typical workflow, is shown by black solid lines. Local sequences, which may be flexibly involved to improve the operation performance, are shown by gray dashed lines. Values between each two main phases denote the phase transition probabilities. It is noteworthy that only transitions with high probabilities are demonstrated in Fig. 8 and others with low probabilities were eliminated. We intended to do so to show the temporal main sequence for a robotic cholecystectomy in real situations. The performance is good for generating the SWM knowledge with the sequence matching degree greater than 0.85 (shown in Tables 6 and 7) when comparing the learned workflows with the observed workflows in both the web/OR surgery operations.

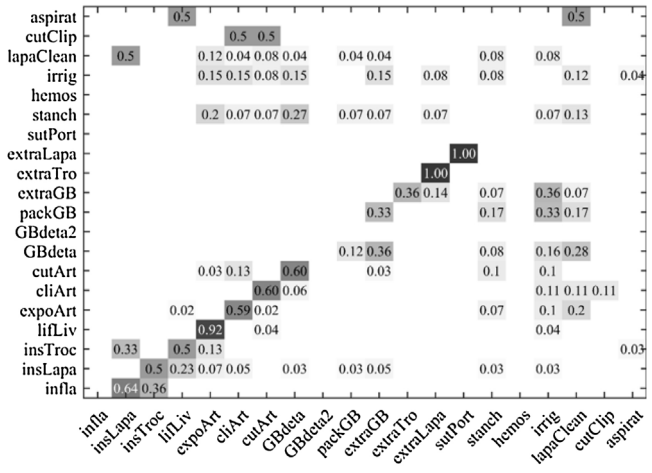

Fig. 7. Mutual phase transition probability matrix.

3.3. Evaluation of the accuracy of the generated SWM and SWM-related knowledge

To evaluate the accuracy of the generated SWM knowledge, cross validation was conducted. 35 videos were divided into 9 groups (one group included 3 videos and each of the other groups included 4 videos). Eight groups of videos were iteratively selected to train the model, and the remaining one group (web video validation stage) and the 4 OR videos (OR video validation stage) were used to test the model. This learning-testing process has been conducted iteratively according to the standard cross-validation process. These 4 OR videos are about the cholecystectomy surgeries and were recorded from the operation rooms at the University of Nebraska Medical Center. Their average length is about 39 min, the average resolution is 1280 × 720, and the average number of phases of a surgery operation is 21.

Web videos are various in surgery procedure/instrument involvements for that the videos are from different surgeons with different professional skill levels. In addition, web videos are usually slightly different from real surgery processes, for that some of them have been edited to remove trivial steps. On the other hand, OR videos are unedited and can accurately reveal real surgery processes which are relatively complicate in surgery procedure/instrument transitions, while they are less various in surgery procedure/instrument involvements for that the surgeons performing the surgeries in the videos are usually from the same hospitals and follow similar surgery standards. Therefore with both types of videos, we aim to validate the accuracy of generated knowledge (surgical workflow model and the instrument usage).

1) Web video validation

An abnormal sequence is defined as the sequence with the probability lower than an acceptable threshold 0.5. Usually the low probability of a sequence is caused by inaccurate/incorrect phase occurrences and transitions. To validate the anomaly-detection ability of the learned SWM knowledge, different levels of noise have been added into 4 testing videos (web videos), respectively (4 videos with 10%-abnormal sequence, 4 videos with 50%-abnormal sequence, and 4 videos with 100%-abnormal sequences). A x%-abnormal sequence is generated by reordering x% of surgical phases of an expert video randomly. An abnormal sequence was confirmed if the matching degree of a sequence to that provided by an expert surgeon (ground truth) was less than 0.50. By comparing the anomaly-detection results provided by the SWM knowledge and

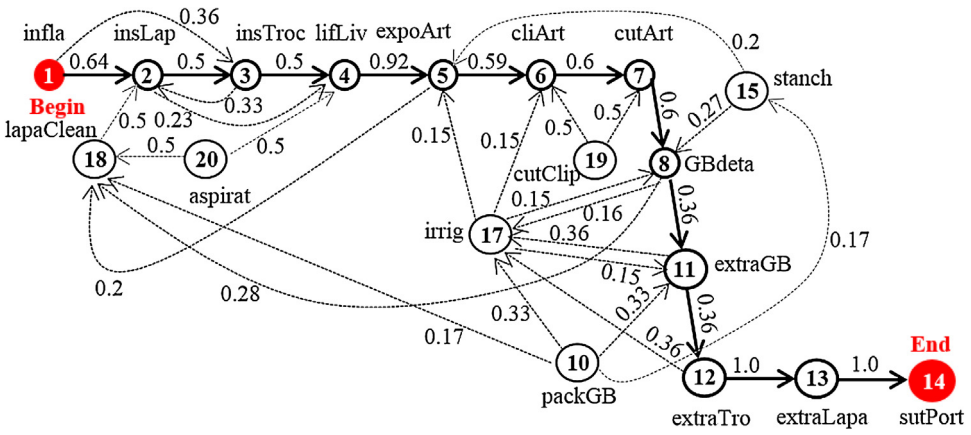

Fig. 8. Temporal main sequence of robotic cholecystectomy.

Table 6
Results for web video validation.

Samples	Ground Truth	Evaluation by SWM	Sequence Matching	Instrument Usage Consistency
expert sequence	normal	normal	0.84	0.95
expert sequence (10% anomaly)	abnormal	abnormal	0.44	0.95
expert sequence (50% anomaly)	abnormal	abnormal	0.27	0.95
expert sequence(100% anomaly)	abnormal	abnormal	0.04	0.95

Table 7
Results for OR video validation.

Samples	Ground Truth	Evaluation by SWM	Sequence Matching	Instrument Usage Consistency
average value	normal	normal	0.86	0.84

Table 8
[{(Algorithm 1)}] Video content matching analysis.

Input: video titles D={d_1, d_2, …}
Output: topic distribution λ_k, topic proportion γ_d on each document, joint topic distribution $p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, \hat{w}_{1:D})$

1. Define topic updating parameter ρ_t (t=1,2, …) for the future time sequence
2. Initialize topic distribution **λ** randomly
3. **for** t=0 to ∞ **do**
4. Choose random video title d
5. Initialize topic proportion γ_{tk}(k=1, 2) for title d
6. **repeat**:
7. Update word distribution $\phi_{d,n}$ for topic k
8. Update topic proportion γ_d for title d
9. **until** title objective
$\frac{1}{k}\sum_k |adjustment\ of\ \gamma_{tk}|$<0.0001
10. Calculate the topic distribution $\tilde{\lambda}_k$ given title d,
$\tilde{\lambda}_k = \eta + \sum_n \hat{w}_{t,n}\,\phi_{t,n}$
where η is Dirichlet distribution parameters
11. Update the overall topic distribution
$\lambda_k \leftarrow (1-\rho_t)\lambda_k + \rho_t\tilde{\lambda}_k$
12. Calculate joint topic distribution
$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, \hat{w}_{1:D})$ [52]
13. end for
14. **return** λ_k,γ_d, $p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, \hat{w}_{1:D})$

Table 9
[{(Algorithm 2)}] Surgeon skill evaluation.

Abbreviations: comment x, comment's sentiment label y, a video's comment set X, a video's sentiment label Y, neg/pos: negative/positive sentiment, n count for a sentiment labels, q dislike/like hit number of a video

Input: training comments (X_K, Y_K)={$(x_{1:K}, y_{1:K})$}
unlabeled comments $X_M = \{x_{K+1:M}\}$
Output: sentiment label Y_v for a video

1. Let training set T=(X_K, Y_K), learn sentiment classifier f_0 where argmax $f_0(X_K)= Y_K$
2. classifier f=f_0
negative comment number n_{neg}=0
positive comment number n_{pos}=0
3. **if** M≠0:
4. **for** $x_m \in X_M$ **do**
5. new label y_m =argmax $f(x_m)$
6. $n_{y_m} \leftarrow n_{y_m} + 1$
7. new training set $(X_K, Y_K) \leftarrow (X_K, Y_K) + (x_m, y_m)$
8. new classifier f: argmax $f(X_K)= Y_k$
9. $Y_v = \underset{y_m \in \{neg,pos\}}{\text{argmax}}\ n_{y_m}$
10. **end for**
11. **else**:
12. $n_{neg} = q_{dislike}, n_{pos} = q_{like}$
13. $Y_v = \underset{y_m \in \{neg,pos\}}{\text{argmax}}\ n_{y_m}$
14. **end if**
15. **return** Y_v

those provided by the expert surgeons, the method's effectiveness in abnormal workflow detection was evaluated.

The workflow evaluation results are shown in Table 6. All the sequences with probability lower than 0.5 were successfully detected as 'abnormal'. The average sequence matching degree for the 4 expert videos in the full cross validation was as high as 0.90. The average result of the sequences with 10% abnormal transitions was around 0.44. The results for the sequences with 50% abnormal transitions and the sequences randomly generated were lower than 0.27.

These results with corresponding sequence-matching-degree declining show that the modeled SWM is accurate. It also shows our webSWM method is effective in generating a reliable SWM. Average instrument-usage consistency is larger than 0.95, showing the likely involved instruments in the web videos were accurately defined in learned SWM. In the noise generation, there is no new phases involved. Therefore the instrument usage consistency is unchanged when the noise levels are changed.

2) OR video validation

All the surgical workflows performed by the volunteer expert surgeons in the OR were identified as 'normal' by using the web-learned SWM. The operations were successful in all the recorded videos. As Table 7 shows, after the full cross validation the sequence matching degree is as high as 0.86 and instrument-usage consistency is high as 0.84, proving that our web-learned SWM is accurate and can reflect and evaluate the workflow in actual surgeries.

In terms of surgery-related video collection, the webSWM method is superior to traditional methods with respect to time and cost. To collect the surgery-related videos, the general methods collected videos by employing a group of surgeons to record a certain amount of videos (typically $5 \sim 10$ videos) from the OR. While the webSWM method mines the surgery-related videos from the web. The traditional methods may take days to collect a certain amount of videos (for example 250), while the webSWM method could mine the same amount of surgery-related videos in a short time (in our experiments, the processing time for 250 videos is 10 min). The reliability of the webSWM method demonstrated from the experimental results and the method's potential to reduce time and cost for video collection show its promise to overcome the current problem of limited data sources for surgical workflow modeling. Combined with other technologies, such as automatic instrument tracking and phase segmentation technologies presented in [58,59], the webSWM method has the potential to enable fully automatic, efficient, and reliable surgical workflow modeling procedures.

4. Discussions

1) Bias

First, the web data has stereotypes that the data with popular methods is more abundant than that with less-popular methods. The probabilistic description manner, which is supposed to remain the influential features and ignore the uninfluential features, usually extracts the popular features and ignores the unpopular features. Popularity is not equal to influence, for that some features could be equally influential but be less popular. For a specific website, some content and specific styles of data are relatively more abundant/popular. But the importance of the data is not always consistent with their popularity. The less-popular data could be equally important with the more-popular data. Therefore the bias always exists in the public data, which could reduce the degree of accuracy of the learned surgical workflow models.

2) Critical cases vs exceptions

The consideration of critical cases and exceptions is a key problem to use public available data for modeling. The critical cases contain relative important and typical information, deciding the main structure and stability of a SWM. The exceptions contain useful but untypical information, improving the reliability of a SWM. Given usually the critical cases are popular with abundant web videos, the main structure of a SWM could be correctly learned, ensuring the model's stability. Given our webSWM method collects videos from different surgery procedures recorded from all over the world, the variety could be ensured to some degree, increasing the model's reliability. Therefore, even though the critical cases and exceptions affect the accuracy of a SWM, extending the training samples could alleviate the negative influence of data bias on SWM learning to some extent.

One potential solution in the future could be to use a prior knowledge such as a general workflow "lift the liver – clip the artery – detach the gallbladder" [62] to guide the SWM learning by specifying the relative phase transition sequences and increasing the importance of phases underestimated/ignored by the biased data.

3) Effects of web video selection

First, public-available web videos such as Youtube videos are usually those with successful procedures, containing the typical operations. With the accumulation from the large quantity of web videos, the typical main sequences are emphasized, and meanwhile the less-typical local sequences are underestimated. Second, the cold-start problem, which denotes a system cannot perform all the likely inferences for that the collected data is not sufficient enough to model any given potential situation [63], is unavoidable given that not all the possible operations for a given type of surgery are covered in web videos. This problem is a common one that not only affects our method's performances but also the traditional methods' performances.

4) Performance discussion of sentiment analysis

The current sentiment analysis performance is limited but acceptable. The typical F1-score range for sentiment analysis, using traditional machine learning methods such as NB or SVM, is 0.6–0.8 [64,65]. Our current performance (F1-score 0.71) is within this range, showing that our method could be relatively effective in sentiment analysis. Given the limitation of traditional learning methods, our method could be relatively effective in sentiment analysis. In the future, we could improve the sentiment analysis performance by adopting the state-of-the-art algorithms such as deep neural networks, which already have brought obvious improvements in natural language processing [66].

The correlation between the viewers' comment sentiment and the surgeon's skill level is existed but not strong. We aim to prove that it is feasible to capture this correlation by using the information techniques in this paper and then in the future make the correlation-capturing method practical by using more sophisticate information techniques. At the current stage, our goal is initially achieved and the final result is reasonable with the F1-score greater than 0.69, shown in Table 2.

Viewers' comments such as "I guess the total surgery cost would be unacceptable" are not relevant with the surgeon skill. The sentiment analysis on these comments is meaningless. Fortunately a large part of the comments are about the surgeon's performance or influenced by the surgeon's performance. Sentiment analysis on these comments can generally reflect their assessments on the surgeon's skills.

The potential method for improving the performance of sentiment analysis is to combine both topic modeling and sentiment analysis. With this method, the topics of the comments can be analyzed, and only the surgeon-skill-related comments will be collected. Then sentiment analysis can be conducted on these topic-specific comments to perform more accurate surgical skill evaluation. Given the comments are usually short with limited information, the topic analysis on these comments will be challenging and also will be a good research topic worthy further efforts.

5) Methods for video sample extension

Training videos could be extended by directly collecting more top-ranking videos from the video candidates, which could be extended by launching more web querying. All the videos have been ranked according to their qualification degrees, and videos with higher ranks indicate they have more reliable information for SWM learning. More video involvements will benefit SWM learning by providing the variety of surgical procedures.

6) Video sufficiency analysis

For the specific surgery in this paper, 35 is sufficient given the goal of feasibility study. First, given the laparoscopic cholecystectomy surgery is relatively simple, and some of the phase transitions are standardized as "lift the liver – clip the artery – detach the gallbladder" [62], we do not need a large number of videos for SWM learning. Moreover, according to our experimental experience, the percentage in top 15%–25% is recommended to maintain both the knowledge relevance and correctness. 35 videos are ranked in the top 20% of the 185 videos, which is consistent with the recommended range. Last, the performances in both web video validation (Table 6) and OR video validation (Table 7) also proved that a functioning model could be learned from the 35 videos.

7) Method improvement for multiple operation approaches of one surgery

With different types of instruments and standardized methods, surgical workflow models for a given type of surgery are different. If the major phases of these different models are mutually distinctive, multiple models could be built for the same type of surgery, and then by instrument/phase-based classification the appropriate model will be selected to support a specific surgical operation. If these models are not distinctive enough with most major phases overlapped, only the mean SWM model with its variety being covered is needed to build.

For building different models for different approaches of a given type of surgery, we need to first classify the data into different types. Then based on a specific type of data and the webSWM method, a model is created for representing one approach of a surgery.

The data classification method can be supported by typical classification methods such as SVM, shown in Eq. (9), where X_a denotes the matrix of features including the involved instruments and the surgical phases, W_a denotes the corresponding feature weights, b_a denotes intercept of the optimal hyperplane, and Y_a denotes the identified surgery operation approach. The method for learning parameters W_a and b_a is shown in Eq. (3).

$$Y_a(X_a) = sign\left\{W_a{}^T \cdot X_a + b_a\right\} \tag{9}$$

8) The alternative algorithm for model learning

The surgical workflow is a temporal sequence. The basic requirement towards the selected algorithm is the capability of modeling a directed model. Given the Hidden Markov Model (HMM) is good at temporal sequence modeling [67], the SWM learning algorithm could also be HMM.

9) Solutions for handling incomplete videos

At the current stage, to ensure the data reliability for accurate SWM learning, we only collect information based on observations, instead of logic inference. However, logic inference could dig more information that is true while unobservable. For the sake of this consideration, in the future we could use a weighted logic method such as Markov Logic Network to explore the useful but unobservable information in one incomplete video and integrate the piecemeal but correlated information from multiple incomplete videos [68].

10) Method selection

Compared with the baseline method (LSI + NB), the adopted method (LDA + SVM) performed better in higher-range (top 0%–20%) videos and worse in lower-range (top 0%–50%) videos (please refer the data in Table 4 of our paper. 20% of the total 185 videos is approximately 35, and 50% of the total videos is approximately 92). The goal of setting the baseline is to prove that the web-mining-supported video selection is feasible. Good performances in both baseline and adopted methods validated the feasibility of our web-video mining-supported surgical workflow modeling concept.

As to which method is more reasonable, it depends on practical application requirements. In our work the videos were selected for learning surgery operations, in which the safety of patients was concerned and the reliability of surgery knowledge was emphasized. We believe that the higher consistency do the machine-selected videos have, more reliable information the machine-selected videos can carry. For example, 35 videos with 0.6 consistency is reliable than the 35 videos with 0.51 consistency. Therefore we choose to use the current method, which is more reliable in the higher-range videos, instead of using the baseline method.

5. Conclusion

In this paper, we proposed a web-video-mining-supported surgical workflow modeling (webSWM) method. An accurate and reliable surgical workflow model (SWM) for robotic laparoscopic surgery was generated by mining web videos. Reliable methods for mining internet videos were developed. A steepest-descent algorithm was used to build a robust and reliable SWM. The experiment was conducted using both web videos and OR videos. Satisfied results validated the accuracy of the generated SWM and SWM-related knowledge.

The webSWM method is feasible and effective in extracting SWM-related knowledge from web videos. Satisfied performances in mining web videos and learning surgery-related knowledge show that webSWM is promising in scaling knowledge for intelligent surgical systems.

In the future, we will develop the computer-vision-based method to segment the surgical video with full automation. We will also mine web videos to build a SWM knowledge database for various surgeries. On the application level, we will try to implement the generated SWM knowledge on phase prediction, tool recommendation, and surgery remaining time prediction to realize further benefits for robotic surgeries.

References

[1] Way L, Stewart L, Gantert W, Liu K, Lee C, Whang K, et al. Causes and prevention of laparoscopic bile duct injuries: analysis of 252 cases from a human factors and cognitive psychology perspective. Ann Surg 2003;237:460–9.

[2] Loukas C, Georgiou E. Performance comparison of various feature detector-descriptors and temporal models for video-based assessment of laparoscopic skills. Int J Med Robot Comput Assist Surg 2015;12(3):387–98, http://dx.doi.org/10.1002/rcs.1702.

[3] Seymour N, Gallagher A, Roman S, O'Brien M, Andersen D, Satava R. Analysis of errors in laparoscopic surgical procedure. Surg Endosc 2004;18:592–5.

[4] Sarker S, Chang A, Vincent C, Darzi A. Technical skills errors in laparoscopic cholecystectomy by expert surgeons. Surg Endosc 2005;19:832–5.

[5] Padoy N, Blum T, Ahmadi S. Statistical modeling and recognition of surgical workflow. Med Image Anal 2012;16:632–41.

[6] Blum T, Feußner H, Navab N. Modeling and segmentation of surgical workflow from laparoscopic video. Med Image Comput Comput Assist Interv 2010:400–7.

[7] National Guideline Clearinghouse. [homepage on the internet]. AHRQ-Agency for Health Care Research and Quality [cited 2016 Mar 1]. Available from: http://www.guideline.gov/.

[8] Scientific Medical Societies in Germany. [homepage on the internet]. AWMF-Arbeitsgemeinschaft Der Wissenschaftlichen Medizinischen Fachgesellschaften E.V. Science-based Guidelines for Diagnostics and Therapy [cited 2016 Mar 1]. Available from: http://www.awmf-leitlinien.de/.

[9] Peters J, Fried G, Swanstrom L, Soper N, Sillin L, Schirmer B, et al. Development and validation of a comprehensive program of education and assessment of the basic fundamentals of laparoscopic surgery. Surgery 2004;135:21–7.

[10] Swanstrom L, Fried G, Hoffman K, Soper N. Beta test results of a new system assessing competence in laparoscopic surgery. J Am Coll Surg 2006;202:62–9.

[11] Neumuth T, Krauss A, Meixensberger J, Muensterer O. Impact quantification of the daVinci telemanipulator system on surgical workflow using resource impact profiles. Int J Med Robot Comput Assist Surg 2011;7:156–64.

[12] Blum T, Navab N, Feußner H. Methods for automatic statistical modeling of surgical workflow. Proc Meas Behav 2010:64–5.

[13] Meyer M, Levine W, Egan M, Cohen B, Spitz G, Garcia P, et al. Computerized perioperative data integration and display system. Int J Comput Assist Radiol Surg 2007;2:191–202.

[14] Miyawaki F, Masamune K, Suzuki S, Yoshimitsu K, Vain J. Scrub nurse robot system −intraoperative motion analysis of a scrub nurse and timed-automata-based model for surgery. IEEE Trans Ind Electron 2005;52:1227–35.

[15] Neumuth T, Jannin P, Schlomberg J, Meixensberger J, Wiedemann P, Burgert O. Analysis of surgical intervention populations using generic surgical process models. Int J Comput Assist Radiol Surg 2011;6:59–71.

[16] Ahmadi S, Sielhorst T, Stauder R, Horn M, Feussner H, Navab N. Recovery of surgical workflow without explicit models. Med Image Comput Comput Assist Interv 2006;42:0–428.

[17] Blum T, Padoy N, Feußner H, Navab N. Modeling and online recognition of surgical phases using hidden markov models. Med Image Comput Comput Assist Interv 2008;62:7–635.

[18] Paalvast M. Real-time estimation of surgical procedure duration. Ph.D dissertation. Delf: Delft Univ.; 2015.

[19] Franke S, Meixensberger J, Neumuth T. Multi-perspective workflow modeling for online surgical situation models. J Biomed Inform 2015;54:158–66.

[20] Padoy N, Blum T, Feussner H, Berger M-O, Navab N. On-line recognition of surgical activity for monitoring in the operating room. The AAAI conference on artificial intelligence 2008.

[21] Katic D, Wekerle A, Gärtner F, Kenngott H, Muller-Stich B, Dillmann R, et al. Model-based formalization of medical knowledge for context-aware assistance in laparoscopic surgery. Int Soc Optics Photonics SPIE Med Imaging 2014, http://dx.doi.org/10.1117/12.2042240, 903603-903603.

[22] Lin H, Shafran I, Murphy T, Okamura A, Yuh D, Hager G. Automatic detection and segmentation of robot-assisted surgical motions. Med Image Comput Comput Assist Interv 2005;8:802–10.

[23] Neumuth T, Meißner C. Online recognition of surgical instruments by information fusion. Int J Comput Assist Radiol Surg 2012;7:297–304.

[24] Zeng Q, Sun S, Duan H, Liu C, Wang H. Cross-organizational collaborative workflow mining from a multi-source log. Decis Support Syst 2013;54:1280–301.

[25] Aalst W, Dongen B, Herbst J, Maruster L, Schimm G, Weijters A. Workflow mining: a survey of issues and approaches. Data Knowl Eng 2003;47:237–67.

[26] Song Y, Wang Q, Jiang X, Liu S, Zhang Y, Bai S. Fully automatic volumetric modulated arc therapy plan generation for rectal cancer. Radiother Oncol 2016;119(3):531–6.

[27] Vankipuram M, Kahol K, Cohen T, Patel V. Toward automated workflow analysis and visualization in clinical environments. J Biomed Inform 2011;44:432–40.

[28] Jalote-Parmar A, Badke-Schaub P. Workflow integration matrix: a framework to support the development of surgical information systems. Des Stud 2008;29:338–68.

[29] Neumuth D, Loebe F, Herre H, Neumuth T. Modeling surgical processes: a four-level translational approach. Artif Intell Med 2011;51:147–61.

[30] Reiter R. The frame problem in situation the calculus: a simple solution (sometimes) and a completeness result for goal regression. Artif Intell Math Theory Comput 1991;35:9–80.

[31] Lalys F, Jannin P. Surgical process modelling: a review. Int J Comput Assist Radiol 2014;9(3):495–511.

[32] MacKenzie L, Lbbotson J, Cao C, Lomax A. Hierarchical decomposition of laparoscopic surgery: a human factors approach to investigating the operating room environment. Minim Invasive Ther Allied Technol 2001;10(3):121–8.

[33] Neumuth T, Durstewitz N, Fischer M, Strauss G, Dietz A, Meixensberger J, et al. Structured recording of intraoperative surgical workflows. Med Imaging 2006, http://dx.doi.org/10.1117/12.653462, 0:6145–61450A.

[34] Agarwal S, Joshi S, Finin T, Yesha Y, Ganous TA. Pervasive computing system for the operating room of the future. Mobile Netw Appl 2007;12(2–3):215–28.

[35] James A, Vieira D, Lo B, Darzi A, Yang G-Z. Eye-gaze driven surgical workflow segmentation. International conference on medical image computing and computer-assisted intervention 2007;10(2):110–7.

[36] Goh A, Goldfarb D, Sander J, Milest B, Dunkin B. Global evaluative assessment of robotic skills: validation of a clinical assessment tool to measure robotic surgical skills. J Urol 2012;187(1), 274–52.

[37] Chen C, White L, Kowalewski T, Aggarwal R, Lintott C, Comstock B, et al. Crowd-sourced assessment of technical skills: a novel method to evaluate surgical performance. J Surg Res 2014;187:65–71.

[38] Rahimy E, Garg S. Google glass for recording scleral buckling surgery. JAMA Ophthalmol 2015;133(6):710–1.

[39] Birnbaum F, Wang A, Brady C. Stereoscopic surgical recording ssing gopro cameras: a low-cost means for capturing external eye surgery. JAMA Ophthalmol 2015;133(12):1483–4.

[40] Esparza S, O'Mahony M, Smyth B. Mining the real-time web: a novel approach to product recommendation. Knowl Based Syst 2012;29:3–11.

[41] Tenorth M, Klank U, Pangercic D, Beetz M. Web-enabled robots. IEEE Robot Autom Mag 2011;18:58–68.

[42] Liu R, Zhang X, Webb J, Li S. Context-specific intention awareness through web query in robotic caregiving. IEEE international conference on robotics and automation (ICRA) 2015:1962–7.

[43] Liu R, Zhang X, Li S. Use context to understand user's implicit intentions in activities of daily living. IEEE international conference on mechatronics and automation (ICMA) 2014:1214–9.

[44] Liu R, Zhang X. Context-specific grounding of web natural descriptions to human-centered situations. Knowl Based Syst 2016;111:1–16, http://dx.doi.org/10.1016/j.knosys.2016.07.037.

[45] Samadi M, Kollar T, Veloso MM. Using the web to interactively learn to find objects. AAAI Conf Artif Intell 2012:2074–2080.

[46] Liu R, Zhang X. Understanding human behaviors with an object functional role perspective for robotics. IEEE Trans Cognit Dev Syst 2016;8(2):115–27.

[47] Sebelik M, Wilson C. Sources and quality of online surgical videos: using YouTube to learn thyroid surgery. Otolaryngol Head Neck Surg 2014;151:180–1.

[48] Bezner S, Hodgman E, Diesen D, Clayton J, Minkes R, Langer J, et al. Pediatric surgery on YouTube™: is the truth out there? J Pediatr Surg 2014;49:586–9.

[49] Google. [homepage on the internet]. Google Search Engine [cited 2016 Mar 1]. Available from: www.google.com.

[50] Python Software Foundation. [homepage on the internet]. pafy [cited 2016 Mar 1]. Available from: https://pypi.python.org/pypi/pafy/0.4.2.

[51] Steyvers M, Griffiths T. Probabilistic topic models. In: Landauer TK, MxNamara Dnnis DSS, Kintsch W, editors. Handbook of latent semantic analysis. New Jersy: Lawrence Erlbaum Associates, Inc.; 2011. p. 424–40.

[52] Hoffman M, Blei D. Online learning for latent dirichlet allocation. Adv Neural Inf Process Syst 2012;91:7–925.

[53] Blei DM. Probabilistic topic models. Commun ACM 2012;55:77–84.

[54] Garla V, Taylor C, Brandt C. Semi-supervised clinical text classification with laplacian SVMs: an application to cancer case management. J Biomed Inform 2013;46:869–75.

[55] Maas A, Daly R, Pham P, Huang D, Ng AY, Potts C. Learning word vectors for sentiment analysis. The annual meeting of the association for computational linguistics (ACL) 2009.

[56] Forney G. The viterbi algorithm. Proceedings of the IEEE 1973;61:268–78.

[57] Paulo S. Steepest-descent algorithm. In: Paulo SRD, editor. Adaptive filtering algorithms and practical implementation. 3rd ed. New York, USA: Springer; 2010. p. 49–50.

[58] Oropesa I, Sanchez-Gonzalez P. Laparoscopic instrument tracking based on endoscopic video analysis for psychomotor skills assessment. Surg Endosc 2013;27:1029–39.

[59] Kumar S, Sovizi J, Narayanan MS. Surgical tool pose estimation from monocular endoscopic videos. IEEE international conference on robotics and automation 2015;59:8–603.

[60] Wei X, Croft WB. LDA-based document models for ad-hoc retrieval. Proceedings of the annual international ACM SIGIR conference on research and development in information retrieval 2006:178–85.

[61] Hong L, Davison B. Empirical study of topic modeling in twitter. ACM proceedings of the first workshop on social media analytics 2010:80–8.

[62] Padoy N, Blum T, Ahmadi S, Feussner H, Berger M, Navab N. Statistical modeling and recognition of surgical workflow. Med Image Anal 2012;16:632–41.

[63] Cold Start, 2016. [Online][cited 2016 June 20]. Available: https://en.wikipedia.org/wiki/Cold_start.

[64] Pak A, Paroubek P. Twitter as a corpus for sentiment analysis and opinion mining. Language resources and evaluation conference 2010:1320–6.
[65] Pang B, Lee L. A sentimental education: sentiment analysis using subjectivity summarization based on mining cuts. Proceedings of the annual meeting on association for computational linguistics 2004:271.
[66] Schmidhuber J. Deep learning in neural networks: an overview. Neural Netw 2015;8:5–117.
[67] Tang K, Li F-F, Koller D. Learning latent temporal structure for complex event detection. IEEE conference on computer vision and pattern recognition (CVPR) 2012:1250–7.
[68] Richardson M, Domingos P. Markov logic networks. Mach Learn 2006;62(1–2):107–36.