Automated Surgical-Phase Recognition Using Rapidly-Deployable Sensors

Robert DiPietro[1], Ralf Stauder[2,3], Ergün Kayis[2], Armin Schneider[3], Michael Kranzfelder[3], Hubertus Feussner[3], Gregory D. Hager[1], and Nassir Navab[1,2]

[1] Department of Computer Science, Johns Hopkins University, Baltimore, MD, USA
[2] Computer Aided Medical Procedures, Technische Universität München, Germany
[3] MITI, Klinikum rechts der Isar, Technische Universität München, Germany

Abstract. Surgical-phase recognition is important to many future applications in clinical care, from building context-aware operating rooms to automatically providing feedback to surgeons in training. In this work, we focus on learning-based phase recognition in laparoscopic gallbladder removals (cholescystectomies). Using data from 15 sensors across 42 surgeries, we 1) compare performance using support vector machines, hidden Markov models, and conditional random fields and 2) demonstrate that it is possible to achieve 74% accuracy using only 8 rapidly-deployable sensors.

1 Introduction

Being able to detect the current surgical workflow phase during an intervention is a critical step for many applications, such as context-aware assistance systems, automated triggering of peripheral actions, and prediction of upcoming surgical steps. These features are often mentioned in conjunction with the concept of an "operating room of the future" [6].

Analysis and modeling of surgical workflow has therefore been a field of growing interest for the last few years. Several methods have been developed to model and segment different levels of an intervention, from very detailed surgical activities up to surgical phases and even full surgeries [12]. Various statistical and machine learning approaches have been employed to that end, including hidden Markov models [4], support vector machines [13, 14], random forests [17], and semantic rules [9]. Most methods rely on instrument data to be available, but other approaches directly analyze the laparoscopic video, either to detect the used instruments [1], or to directly recognize workflow information [8].

In this work, the primary objective is to predict surgical phases using sensors that require little to no installation or maintenance time.

2 Data

2.1 Medical procedure

The type of surgery we chose for this study is a laparoscopic cholecystectomy (minimally invasive removal of the gallbladder). It is performed regularly, is a

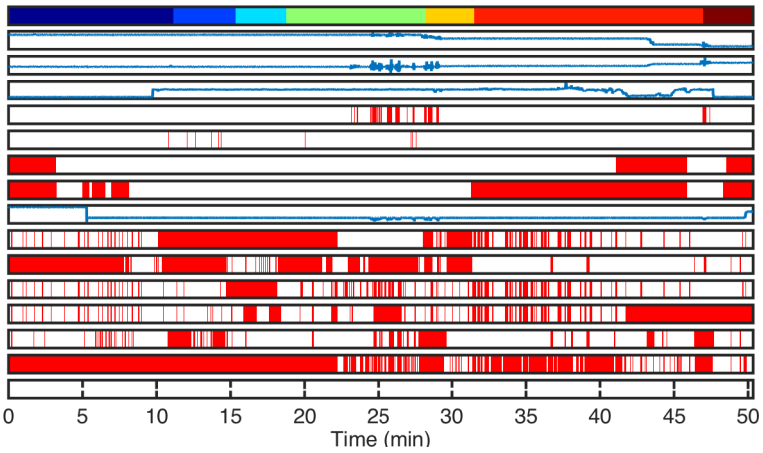

Fig. 1. All sensor data from one exemplary surgery. Best viewed in color. From top to bottom: surgical phase (color-coded ground truth), irrigation weight, suction weight, intra-abdominal pressure, HF coagulation, HF cutting, OR room light, surgical lamp, table inclination, alligator forceps, PE forceps, plastic-clip applicator, laparoscopic scissors, irrigation rod, suction rod, metal-clip applicator. Continuous signals are shown in blue, and binary signals in red (when active). Note that the last signal is blank in this example because the metal-clip applicator was not used in this surgery.

highly standardized surgery, and is well known in the field of surgical workflow analysis.

Under general anesthesia and after inflating the abdomen with inert gas, four trocars are inserted as working channels through minimal incisions in the abdominal wall (phase 1). Next, the area known as *Calot's triangle* is prepared so that the cystic duct and cystic artery are exposed (phase 2). The cystic duct and artery are then sealed off using plastic or metal clips and cut (phase 3). Next, the gallbladder is completely detached from the liverbed (phase 4), and then the liver and surrounding tissue is checked for bleeding (phase 5). The detached gallbladder is then packaged in a plastic bag and retrieved from the body (phase 6). Finally the trocars are removed, the gas is drained, and the incisions are closed (phase 7). Occasionally the orders of phases 5 and 6 can be swapped, but this does not occur in our dataset.

2.2 Sensors

We equipped an operating room at (Medical department, Anonymous Hospital, Some university), with sensors to capture various intraoperative signals. Comparable to the work in [10], we were able to record a total of 15 signals, of which 4 were continuous and the remaining 11 were binary. The continuous signals include the intra-abdominal pressure applied by the insufflator, the weight of the irrigation and suction bags respectively, and the inclination of the surgical table

as it is adjusted throughout the intervention. Two of the binary signals are the state of the OR light and the surgical lamp, two further signals depict the mode of the high-frequency (HF) generator used for coagulating or cutting tissue by applying monopolar current. The remaining 7 signals indicate usage of the up to 7 laparoscopic instruments.

The first half of the described signals, namely the continuous measurements, the light signals, and the HF modes, can be recorded by attaching suitable sensors to corresponding places of the surgical table or on status indication LEDs, resulting in minimal installation effort. These rapidly-deployable sensors should extend easily to other hospitals.

The instrument-usage signals require more elaborate sensors. In our case radio frequency identification (RFID) tags have been attached to the instruments and an array of antennas was placed on the instrument table. An instrument is considered to be in use (with its signal being active) if the corresponding RFID tag cannot be detected within range of the antennas. This generally requires more infrastructure and therefore cannot be made available within a short timeframe in most operating theaters. Also, we note that the RFID signals are much more noisy than the other signals; this can be seen clearly in Fig. 1.

3 Methods

3.1 Feature Extraction

As a first step, we extract features at every time step for every input signal. The features are based on raw signal values, windowed means, windowed standard deviations, and slopes of windowed linear fits. Window-based features are computed using various durations (4 s, 16 s, 64 s, and 256 s). Cross validation was not used here; we simply used powers of 4 in order to capture both short- and long-term trends. Furthermore, the windows always end at the current time step; this avoids using information from the future (for online applications). This results in $4 \times 3 + 1 = 13$ features per time step per signal. Finally, to also capture behavior that was confined to the past, we copy all features from earlier times (4 s, 16 s, 64 s, and 256 s) over to the current time step. This results in a total of $13 \times 5 = 65$ features per signal per time step.

3.2 Classification

We will compare three different classification methods to classify each time step into 1 of 7 possible phases. These methods are based on support vector machines (SVMs) [5,7], hidden Markov models (HMMs) [15], and conditional random fields (CRFs) [11, 19].

In the first method, we take a one-vs-one [2] approach to multi-class learning with linear SVMs[1], resulting in $7 \times 6 \div 2 = 21$ classifiers. At test time, the

[1] Gaussian-kernel SVMs were also considered, also using a logarithmic grid search to determine parameters, but performance did not improve over linear SVMs.

prediction for each time step is the phase with the most votes among the 21 classifiers. This results in classification with no regard to temporal consistency and is used as a baseline. Next, to provide further input to the next two methods, we extract SVM scores for a different set of training data.

In the second method, SVM scores are used as observations in a Gaussian HMM with diagonal covariance. This generative approach models the distribution over both observations and unobserved phases. Training occurs using a training set separate from that of the SVMs'; this process is simple because phases are observed during training, and there is therefore no need for expectation maximization. At test time, phases are predicted on a sequence-by-sequence basis; for each sequence, the most-likely phase configuration over all time steps is determined using a standard forward-backward algorithm [15].

In the third method, SVM scores are used as features in a linear-chain conditional random field. This discriminative approach models the distribution over phases given observations. Unary potentials for each phase are computed using the 21 SVM scores and 1 bias term, resulting in a total of $7 \times 22 = 154$ unary weights (shared across time). Pairwise potentials for each (phase, phase) pair are computed using just 1 bias term, resulting in $7 \times 7 = 49$ pairwise weights (again shared over time). Training is performed by maximizing the (convex) L2-regularized conditional likelihood using standard nonlinear-optimization techniques. At test time, phases are predicted on a sequence-by-sequence basis; for each sequence, the most-likely phase configuration over all time steps is determined. (Marginals during training and most-likely phase configurations during testing are both computed efficiently using standard message-passing techniques [3].)

For all methods, all input features are normalized, and free parameters (the SVMs' soft-margin parameter and the CRF's L2-regularization parameter) are fixed using a logarithmic grid search and 4-fold cross validation using their respective training sets. Varying class sizes were accounted for during SVM training by scaling slack parameters and during CRF training by including bias terms.

4 Experiments, Results, and Discussion

4.1 Experimental Setup

We use data from 42 surgeries, each annotated by a medical expert using associated laparoscopic videos. Fig. 1 shows data from an exemplary surgery. The top row indicates ground-truth surgical phase; the next eight rows show continuous and binary data from the rapidly-deployable sensors; and the bottom seven rows show binary data from the RFID sensors. One RFID signal is not shown and was never used to record data in our experiments.

We aim to evaluate performance for each of the three methods, denoted SVM, SVM-HMM, and SVM-CRF, under two difference scenarios: using data from all available sensors and using data from rapidly-deployable sensors only. The following process was carried out once for each scenario.

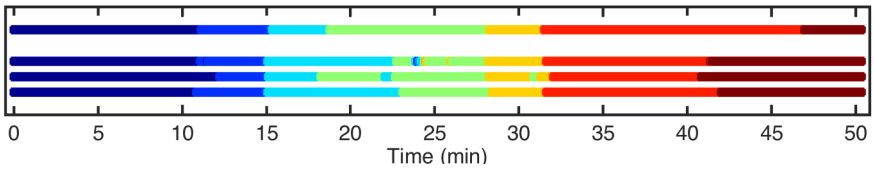

Fig. 2. Example predictions from the same sequence shown in Fig. 1, using all available sensors. Best viewed in color. From top to bottom: ground truth, SVM prediction, SVM-HMM prediction, SVM-CRF prediction. Accuracies are 78.4%, 81.1%, and 79.8% respectively.

	SVM			SVM-HMM			SVM-CRF		
	Precision	Recall	Jaccard	Precision	Recall	Jaccard	Precision	Recall	Jaccard
PT	89.8	91.6	82.4	92.5	89.0	82.7	89.0	88.8	78.7
P	77.5	70.2	58.2	78.5	60.8	51.5	83.8	71.8	62.8
C	71.5	71.7	55.7	61.5	79.6	53.5	67.1	79.2	57.3
DG	64.3	62.4	46.3	68.3	59.3	46.4	70.0	71.7	53.9
SB	70.5	82.0	61.0	72.7	84.9	64.1	81.6	67.7	57.3
RG	81.5	77.9	66.0	72.9	81.3	61.4	69.4	89.8	62.8
F	74.9	77.6	60.7	79.2	59.3	47.8	87.6	47.9	39.9

Table 1. Precision, recall, and Jaccard index for each method and surgical phase using data from all sensors. All metrics were computed by averaging results over four randomly-chosen but non-overlapping test sets, each consisting of 10 surgeries. The overall accuracies for each method are 75.9%, 73.1%, and 74.4% respectively. Phases are indicated by their abbreviations: Place Trocar, Preparation, Clipping, Detaching Gallblader, Stop Bleeding, Retrieve Gallbladder, and Finalization.

We begin our experiments by randomly shuffling the order of the 42 surgeries. Next we select surgeries 1–16 for SVM training, 17–32 for HMM and CRF training, and 33–42 for testing. Training and testing is then carried out as explained in Sec. 3.2 and performance is evaluated. Performance metrics include precision, recall, and Jaccard index, all computed on a per-phase basis, and accuracy, computed globally as the percentage of correct predictions. This process is repeated for 4 trials with randomly-chosen but non-overlapping test sets, and final performance metrics are averaged across the trials.

4.2 Results

We first present results using data from all sensors. Fig. 2 shows results for one exemplary surgery, and Table 1 shows all performance metrics averaged over all trials, as described in Sec. 4.1. The accuracies for the SVM, SVM-HMM, and SVM-CRF, averaged over all trials, are 75.9%, 73.1%, and 74.4% respectively.

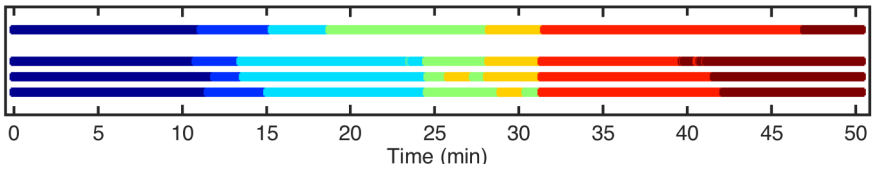

Fig. 3. Example predictions from the same sequence shown in Fig. 1, using rapidly-deployable sensors only. Best viewed in color. From top to bottom: ground truth, SVM prediction, SVM-HMM prediction, SVM-CRF prediction. Accuracies are 70.2%, 69.5%, and 74.2% respectively.

	SVM			SVM-HMM			SVM-CRF		
	Precision	Recall	Jaccard	Precision	Recall	Jaccard	Precision	Recall	Jaccard
PT	93.0	95.5	89.0	91.0	89.6	82.1	88.3	92.9	82.2
P	77.1	61.3	51.5	73.7	51.3	42.3	76.3	61.3	51.0
C	62.1	83.9	55.2	62.3	80.5	54.3	65.0	82.2	57.3
DG	69.9	50.2	41.5	61.1	39.9	31.8	60.4	57.1	41.3
SB	65.9	80.7	56.8	58.5	85.8	52.7	67.4	62.1	47.1
RG	81.1	74.1	62.2	71.7	82.4	61.4	68.5	88.7	61.5
F	71.6	75.7	56.4	76.6	61.0	49.8	82.3	45.2	37.1

Table 2. Precision, recall, and Jaccard index for each method and surgical phase using data from rapidly-deployable sensors only. All metrics were computed by averaging results over four randomly-chosen but non-overlapping test sets, each consisting of 10 surgeries. The overall accuracies for each method are 73.9%, 69.6%, and 70.4% respectively.

Next we present results using data from rapidly-deployable sensors only. Fig. 3 shows results for the same exemplary surgery, and Table 2 shows all performance metrics averaged over all trials, as described in Sec. 4.1. The accuracies for the SVM, SVM-HMM, and SVM-CRF, averaged over all trials, become 73.9%, 69.6%, and 70.4% respectively.

4.3 Discussion

One takeaway from these experiments is that the temporal component introduced by the HMM and CRF does not improve performance. We believe that this is because the HMM and the linear-chain CRF only capture extremely-local temporal dependencies. This essentially results in a smoothed version of the SVM results, depending (indirectly through learned parameters) on the relative abundances of same-phase transitions and different-phase transitions. Confusion matrices for SVM-only results are included in Table 3.

In future work we plan to experiment with more complex models which do a better job at capturing temporal relationships. Some models we plan to

	PT	P	C	DG	SB	RG	F
PT	91.8	3.9	0.0	0.9	0.6	0.0	2.8
P	6.9	69.6	10.8	11.5	1.1	0.0	0.0
C	3.1	10.0	71.8	14.4	0.8	0.0	0.0
DG	1.1	6.6	14.1	62.1	16.0	0.0	0.0
SB	0.0	0.2	2.2	13.0	82.0	2.6	0.0
RG	0.9	0.0	0.4	0.7	3.0	78.1	16.8
F	0.4	0.0	0.1	0.0	0.0	22.3	77.2

	PT	P	C	DG	SB	RG	F
PT	95.6	2.9	0.8	0.1	0.5	0.0	0.1
P	4.3	60.4	27.5	5.6	0.8	0.0	1.4
C	0.5	10.2	83.7	5.4	0.1	0.0	0.0
DG	0.1	5.1	21.1	50.5	23.1	0.0	0.0
SB	0.0	0.6	1.1	15.7	80.7	1.9	0.0
RG	1.8	0.0	0.1	0.2	2.2	74.1	21.5
F	0.5	0.0	0.0	0.0	0.0	24.2	75.2

Table 3. SVM classification confusion matrices using all sensors (left) and rapidly-deployable sensors only (right). Rows correspond to ground truth, columns to predictions. Entries were computed by accumulating ground truth and predictions over four randomly-chosen but non-overlapping test sets, each consisting of 10 surgeries.

consider are semi-Markov CRFs [16], which model segment-level phases and transitions rather than time-step-level phases and transitions, and skip-chain CRFs [18], which include longer-term pairwise connections, typically at the cost of performing approximate inference rather than exact inference.

Finally we note that 74% accuracy is obtainable using only the 8 of 15 sensors that are rapidly deployable.

5 Conclusions

The first contribution of this work is a performance comparison using support vector machines, hidden Markov models, and conditional random fields, applied to automated surgical-phase recognition. We found that SVMs alone achieve accuracy rates of approximately 75%. Furthermore, we found that the temporal component introduced by HMMs and linear-chain CRFs does not improve performance.

The second contribution consists of demonstrating that 74% accuracy can be achieved using only 8 rapidly-deployable sensors. This is a reduction of only 2% accuracy from using all 15 sensors, including those that take much more time and effort to install.

References

1. Allan, M., Ourselin, S., Thompson, S., Hawkes, D.J., Kelly, J., Stoyanov, D.: Toward detection and localization of instruments in minimally invasive surgery. IEEE transactions on bio-medical engineering 60(4), 1050–8 (Apr 2013)
2. Allwein, E.L., Schapire, R.E., Singer, Y.: Reducing multiclass to binary: A unifying approach for margin classifiers. The Journal of Machine Learning Research 1, 113–141 (2001)
3. Bishop, C.M.: Pattern recognition and machine learning. Springer (2006)

4. Blum, T., Navab, N., Feußner, H.: Methods for Automatic Statistical Modeling of Surgical Workflow. Proceedings of Measuring Behavior 2010, 64–65 (2010)
5. Burges, C.J.: A tutorial on support vector machines for pattern recognition. Data mining and knowledge discovery 2(2), 121–167 (1998)
6. Cleary, K., Kinsella, A., Mun, S.K.: OR 2020 workshop report: Operating room of the future. International Congress Series 1281, 832–838 (May 2005)
7. Cortes, C., Vapnik, V.: Support-vector networks. Machine learning 20(3), 273–297 (1995)
8. Haro, B.B., Zappella, L., Vidal, R.: Surgical Gesture Classification from Video Data. In: MICCAI. pp. 1–8 (2012)
9. Katić, D., Wekerle, A.L., Gärtner, F., Kenngott, H., Müller-Stich, B.P., Dillmann, R., Speidel, S.: Knowledge-Driven Formalization of Laparoscopic Surgeries for Rule-Based Intraoperative Context-Aware Assistance. In: 5th International Conference, IPCAI. pp. 158–167. Fukuoka, Japan (2014)
10. Kranzfelder, M., Schneider, A., Fiolka, A., Schwan, E., Gillen, S., Wilhelm, D., Schirren, R., Reiser, S., Jensen, B., Feußner, H.: Real-time instrument detection in minimally invasive surgery using radiofrequency identification technology. Journal of Surgical Research 185, 1–7 (Jul 2013)
11. Lafferty, J., McCallum, A., Pereira, F.C.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. International Conference on Machine Learning (ICML) (2001)
12. Lalys, F., Jannin, P.: Surgical process modelling: a review. International journal of computer assisted radiology and surgery 9(3), 495–511 (May 2014)
13. Lalys, F., Riffaud, L., Morandi, X., Jannin, P.: Surgical Phases Detection from Microscope Videos by Combining SVM and HMM. In: Menze, B., Langs, G., Tu, Z., Criminisi, A. (eds.) Medical Computer Vision. Recognition Techniques and Applications in Medical Imaging, Lecture Notes in Computer Science, vol. 6533, pp. 54–62. Springer, Berlin, Heidelberg (2011)
14. Neumuth, T., Jannin, P., Schlomberg, J., Meixensberger, J., Wiedemann, P., Burgert, O.: Analysis of surgical intervention populations using generic surgical process models. International journal of computer assisted radiology and surgery 6(1), 59–71 (2011)
15. Rabiner, L.: A tutorial on hidden markov models and selected applications in speech recognition. Proceedings of the IEEE 77(2), 257–286 (1989)
16. Sarawagi, S., Cohen, W.W.: Semi-markov conditional random fields for information extraction. In: Advances in Neural Information Processing Systems. pp. 1185–1192 (2004)
17. Stauder, R., Okur, A., Peter, L., Schneider, A., Kranzfelder, M., Feuß ner, H., Navab, N.: Random Forests for Phase Detection in Surgical Workflow Analysis. In: 5th International Conference on Information Processing in Computer-Assisted Interventions (IPCAI) (2014)
18. Sutton, C., McCallum, A.: Collective segmentation and labeling of distant entities in information extraction. Tech. rep., DTIC Document (2004)
19. Sutton, C., McCallum, A.: An introduction to conditional random fields. arXiv preprint arXiv:1011.4088 (2010)