Modeling and Segmentation of Surgical Workflow from Laparoscopic Video

Tobias Blum[1], Hubertus Feußner[2], and Nassir Navab[1]

[1] Computer Aided Medical Procedures (CAMP), Technische Universität München, Germany

[2] Department of Surgery, Klinikum Rechts der Isar, Technische Universität München, Germany

Abstract. Modeling and analyzing surgeries based on signals that are obtained automatically from the operating room (OR) is a field of recent interest. It can be valuable for analyzing and understanding surgical workflow, for skills evaluation and developing context-aware ORs. In minimally invasive surgery, laparoscopic video is easy to record but it is challenging to extract meaningful information from it. We propose a method that uses additional information about tool usage to perform a dimensionality reduction on image features. Using Canonical Correlation Analysis (CCA) a projection of a high-dimensional image feature space to a low dimensional space is obtained such that semantic information is extracted from the video. To model a surgery based on the signals in the reduced feature space two different statistical models are compared. The capability of segmenting a new surgery into phases only based on the video is evaluated. Dynamic Time Warping which strongly depends on the temporal order in combination with CCA shows the best results.

1 Introduction

Automatic analysis of surgical workflow is an important topic for assessment of surgical skills, analysis of surgical workflow and intelligent systems that need to be aware of the current state of an ongoing surgery. Work in this area usually involves signals that can be obtained in an automatic way. This can be video images, information about tools that are currently used, signals from robotic systems or additional sensors like force sensors, that are installed on surgical tools. These signals are used as input data for machine learning techniques or statistical modeling. For surgical skills assessment often simulators are used, where it is possible to attach sensors to tools or phantoms [1] or use tracking systems to record motion [2]. Other work uses signals from surgical robots [3] where sensors are often built-in and the data is easily accessible. However, in non-robotic surgery, the acquisition of signals is a more challenging problem. In laparoscopic surgery, video images are one important source of information. In [4] instrument segmentation and tracking, tissue deformation and changes in specular highlights are detected from laparoscopic video. This data has been used to classify four different states. In [5] five different laparoscopic tools have been recognized

T. Jiang et al. (Eds.): MICCAI 2010, Part III, LNCS 6363, pp. 400–407, 2010.

based on color and shape using a stereo endoscope in a simulated setup. Both methods have not been used on whole surgeries, where a lot of different instruments are used that often only have subtle differences. Recognition of surgical phases for a whole surgery has been shown in [6], where video was used to detect the presence of surgical clips and whether the endoscopic camera is inserted or not. However, additional information about the use of instruments has been used, which has been obtained manually. In [7] four operating room states have been detected from a video camera mounted on the ceiling of an OR. In a simulate setup a more fine grained detection of surgical workflow has been shown by [8] using nine external cameras. In this work we present an approach to detect phases of a full real minimally invasive surgery (MIS) only from laparoscopic video. Instead of training classifiers for specific instruments, we use a supervised dimensionality reduction on simple image features. By using additional data for the dimensionality reduction, we extract features from the video that contain semantic information. First we will describe the signals that are used, then we will discuss statistical modeling based on these signals.

2 Method

2.1 Signals

The method we describe is applicable to every kind of laparoscopic surgery. For the experiments we used data from a laparoscopic cholecystectomy. This is a very common surgery that is performed minimally invasive in most of the cases. The surgery has a fixed workflow that we have split up into 14 phases that occurred in every instance of the surgery. Especially in MIS the workflow is strongly correlated with the instruments that are used. For every phase the ending point has been defined based on the use of a certain instrument or a combination of instruments. The starting point of a phase corresponds to the ending point of the previous phase. We have recorded the laparoscopic video and additional video from external cameras for ten surgeries. We present a method that tries to detect the 14 phases only based on the laparoscopic video. The external videos and the laparoscopic video have been used to manually annotate which instrument is used at which time. The information about the instrument use is taken for the dimensionality reduction that is described below but not for the detection itself. There are 17 different signals that have been obtained for every surgery. Most of them represent the use of surgical instruments. But also high-frequency coagulation and cutting, which is performed by applying current to an instrument, and the information which trocar is placed, are used. Every surgery i is represented by a multidimensional time series $O_i \in \mathbb{R}^{17 \times l_i}$, where l_i is the length of operation i in seconds. While the values of O_i will only be 0 and 1 in our case, we use $\mathbb{R}$ in the formula, as the described methods are also applicable to real valued data without modifications. Such a representation of the surgical workflow by instrument vectors has been used before for segmenting a surgery into phases by [9]. An example of these signals can be seen in figure 1.

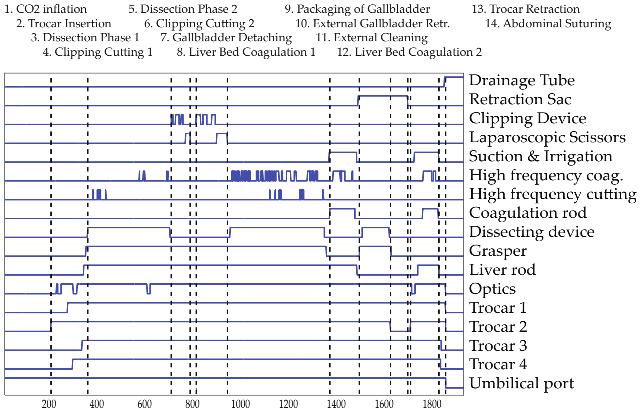

Fig. 1. The instrument use over time during one exemplary surgery. The time is given in seconds and the dotted lines indicate the phases.

From the video images a range of simple image features are computed for every image. The features are horizontal and vertical gradient magnitudes, histograms and the pixel values of a 16x16 version of the image. All of these features have been computed for all three RGB and all three HSV channels, resulting in a 1932-dimensional feature vector for each image. Sampling the features at 1 Hz we obtain the time series $V_i \in \mathbb{R}^{1932 \times l_i}$ for every surgery i.

Most machine learning methods do not perform well with high dimensional feature spaces. There are several ways to deal with this problem. One way is to design classifiers that detect certain instruments or aspects of a surgery as for example done by [4,5]. When developing such a classifier, the feature space is usually reduced by manually choosing features that work well for a certain instrument. While these methods work, it is tedious to design them and often they are only applicable to one certain kind of procedure. Other approaches that have been used in the domain of workflow analysis are unsupervised dimensionality reduction methods like PCA [3] which performs dimensionality reduction in a way to maintain the maximum variability in the data or feature weighting methods like Boosting [10] which select features based on their capability to discriminate between two classes. We use another approach that makes use of the additional information about the use of instruments. In contrast to the features that are extracted from the video, the instruments have an obvious strong semantic meaning. By using Canonical Correlation Analysis (CCA) the visual features are weighted based on their correlation with the manually annotated signals. By using CCA we perform a dimensionality reduction such that the resulting signals are correlated with semantic meaningful signals and thus also have an expressive power.

CCA takes two time series $\mathbb{O} \in \mathbb{R}^{o \times l}$ and $\mathbb{V} \in \mathbb{R}^{r \times l}$ and computes two projection matrices A and B where A_i, respectively B_i denote the ith row of the matrix. The two matrices project both time series to a new space with dimensionality $d = min(o, r)$. This is done such that the correlation between every

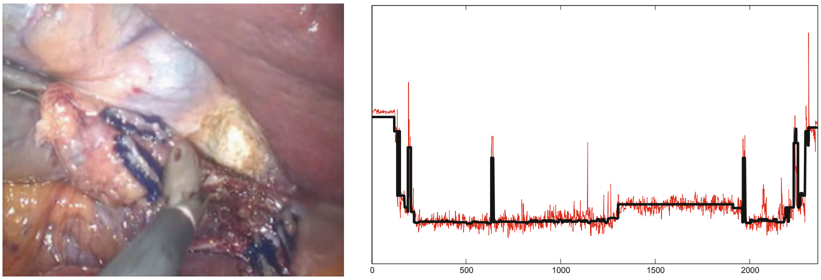

Fig. 2. Left: Example of a laparoscopic image. Right: Instruments (bold) and image features (fine) projected to a common space using CCA. The first dimension which has the highest correlation is shown here.

pair, $corr(A_i\mathbb{O}, B_i\mathbb{V})$, is maximized while every linear combination A_i is orthogonal to all linear combinations A_j, $j < i$. The same condition holds for B. CCA can be seen as a method that takes two views of the same semantic object in order to extract a representation of the semantics [11]. It has been used e.g. for alignment of human behavior [12] and text based image retrieval [11].

By applying CCA to our data, we reduce the dimensionality of the image features to 17 and obtain a new 17-dimensional representation of the instrument use. The first dimension of the image features and the instrument signals projected to a common space is shown in figure 2. For the statistical modeling and the detection of phases that is described below, we completely discard $A\mathbb{O}$ and only use $B\mathbb{V}$. The correlation in the new space is decreasing with every dimension. Therefore we have chosen only to use dimensions, where the correlation $corr(A_i\mathbb{O}, B_i\mathbb{V})$ is > 0.50. We compare this method to a standard dimensionality reduction using PCA.

2.2 Modeling

For the segmentation we are using a 14-state left-to-right Hidden Markov Model (HMM), where each state represents one phase. To segment a surgery O_i into phases, we compute the Viterbi-path, that assigns one of the 14 states to each time step of O_i. As each HMM state corresponds to one phase, we can directly use this to segment the surgery. The HMM transition probabilities are simply estimated from the length of each phase in the training data. Defining the observation symbol probabilities i.e. the probability that one feature vector has been generated in one phase, is an important choice for an HMM. As we have real valued data, standard methods like counting the observation symbol frequency are not applicable. To be able to compare different advanced methods, we have chosen to use WEKA [13], a library that implements a wide range of standard machine learning approaches many of which can output probabilities that can be used as observation symbol probabilities. A first test was done using nine surgeries for training and one for testing. The segmentation results for a 14-state HMM using different classifiers have been computed. The best results have been achieved using Support Vector Machines and the meta-classifiers RotationForest, Bagging and LogitBoost. These have been included in a full cross-validation that is described later.

While a 14-state left-to-right HMM takes into account the temporal order of the phases, it does not capture the whole underlying semantics of the workflow. Especially for the signals that are obtained using CCA the model should be able to represent as much semantic information as possible. One option would be constructing a HMM that has many states, modeling each surgical step. However constructing such a HMM is difficult. Instead we have chosen to use Dynamic Time Warping (DTW) to build a model of an average surgery that captures the underlying semantics. DTW is a method that warps one time series onto another one. This is done by generating a warping path that maps every time step i of one surgery to a time step j in the other surgery while minimizing the sum of distances between corresponding points. Similar as done by [10] we construct a model of an average surgery by warping all surgeries to a common timeline and averaging the signals for each time step. As we know the phase for every time step in the training surgeries, we can label the phase for every time step of the average surgery. To segment a new surgery we warp it to the average model using DTW and carry over the phase labels. Building the DTW average and warping a surgery to this average is done using the features obtained after applying CCA, respectively using PCA.

3 Results

For comparing the methods we have used data from ten surgeries. For three of the surgeries, parts of the surgery have not been recorded due to technical problems. These surgeries have been used only for training. We have performed a leave-one-out cross-validation always using one of the seven complete surgeries for testing and all other nine for training. Four different methods have been compared. DTW using the features obtained using CCA, DTW on the features after PCA, and HMM on the data from CCA and PCA. For the HMM observation probability distribution we have used RotationForest, Bagging, LogitBoost and SVM. For the three meta-classifiers which build a classifier based on simpler classifiers we have performed a full cross-validation using several choices of simple classifiers. We only provide the results obtained with the best classifier, which was in both cases Bagging using C4.5 decision trees. For the methods that use PCA, we tried different numbers of principal components. The results are presented in table 1. It can be seen that the standard deviation is very high. We think that this is a result of the small training set and is an indicator that results can be improved when working with more data. The confusion matrix of DTW + CCA can be seen in figure 3. Most errors are along the diagonal. This is because a strong temporal model is used.

Recording laparoscopic and additional external video images during a surgery and labeling the instrument use is difficult and tedious. Therefor we could only acquire a data set of limited size. One must be careful to draw conclusions from

Table 1. Means and standard deviations of the number of time-steps where the phase was classified correctly

	PCA	CCA
HMM + Bagging	47.12%(±12.67)	53.46%(±14.51)
DTW	62.90%(±18.30)	76.81% ± (12.42)

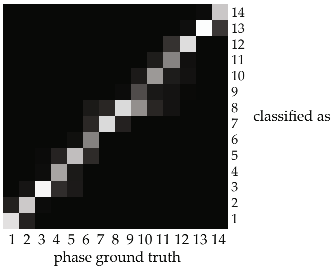

Fig. 3. This images shows the confusion matrix from the whole cross-validation visualized by the image brightness. It can be seen that in most cases of misclassification a phase is classified as neighboring phase.

such a data set. To be able to interpret the outcome, we compared the results from the different methods using the Wilcoxon signed-rank test, a statistical hypothesis test that can handle small data sets. The differences between the compared methods and the corresponding p-values are provided in table 2. The PCA data does not contain a large amount of semantic information. Therefore we did not expect improvement from using DTW compared to a 14-state HMM (1). While there is a difference of 15.8%, this results can not be considered as significant ($p > 0.03$) as the difference mainly results from only two surgeries. Also from using HMMs with PCA or CCA (2) we did not expect a big difference, as the HMM does not fully take advantage of the characteristic of the CCA data. While the significance here is high, the difference is only low. For using CCA + DTW we expected a significantly better result as PCA + DTW (3) or CCA + DTW (4), as only the combination CCA and DTW makes full use of the semantic information that is added by the dimensionality reduction using CCA. This assumption is supported by the results of our comparison.

Table 2. Comparison of the methods that have been used. Difference in percentage points (pp) and p-value is given

compared methods	difference in pp	p-value
(1) PCA + HMM \ PCA + DTW	15.8%	0.188
(2) PCA + HMM \ CCA + HMM	6.3%	0.016
(3) CCA + DTW \ PCA + DTW	29.7%	0.023
(4) CCA + DTW \ CCA + HMM	23.4%	0.008

One way to improve the performance would be to add additional information. As discussed before, other work on laparoscopic video used classifiers for special instruments or aspects of the surgery. In [6] we have presented classifiers to detect surgical clips or whether the camera is inserted into a trocar or not. We have added these two signals to the ones obtained using CCA. Using DTW we obtained a classification result of 79.12%. Other information that could be added with limited technical efforts are signals that are obtained from devices or machines used in the OR. As example we added two signals representing the use of high frequency coagulation and cutting. Using these signals we could further improve the classification to 81.36%.

4 Discussion

In this work we have presented a method that allows segmenting a laparoscopic surgery into phases, using only information from laparoscopic video. We have used a supervised dimensionality reduction method that makes use of additional semantic meaningful information to extract a new representation of the image features that also includes semantic information. In combination with a statistical model that can represent the semantics of time series, we have shown that this method performs better than standard machine-learning and dimensionality reduction methods. It has been shown that especially the combination of the supervised dimensionality reduction and an appropriate statistical model leads to better results. One shortcoming of this work is that the segmentation can only be performed after the whole video has been recorded as DTW requires the whole time-series. HMMs are capable of estimating the current state while the time series is not complete yet. However the results of this work have shown that a simple HMM topology can not achieve good results. One way to handle this would be to use more complex HMM topologies that take into account more of the semantics of the data. This could be achieved using HMMs that derive their topology from data as done by [14]. One advantage of DTW is that a warping path is obtained that assigns every time step of a surgery to a time step of the average model. By taking the warping paths of two surgeries their video can be synchronized e.g. to compare different surgeries or to show a set of synchronized surgeries for training. It can also be used to automatically search for a certain phase in the video.

We believe that methods like CCA will play in important role for workflow analysis. The amount of data that can be obtained from the OR is increasing. There is a growing number of cameras, data can be gathered from anesthesia devices and signals from robots, instrument tracking or people localization systems become available. To be able to combine data from several sources and to do sophisticated modeling and analysis, methods like CCA are well suited. An important advantage of the method that was presented here is that we only need the video to detect the current phase. By taking the approach of performing a supervised dimensionality reduction we add the additional information about instrument use while being able to segment a new surgery without needing this additional information. A future goal is to extend the method to online use. This would allow monitoring, prediction the remaining duration of a surgery or offering context-sensitive user interfaces only by using the data from the laparoscopic video.

References

1. Mackel, T., Rosen, J., Pugh, C.: Markov Model Assessment of Subjects' Clinical Skill Using the E-Pelvis Physical Simulator. IEEE Transactions on Biomedical Engineering 54(12) (2007)
2. Leong, J., Nicolaou, M., Atallah, L., Mylonas, G., Darzi, A., Yang, G.: HMM assessment of quality of movement trajectory in laparoscopic surgery. In: Larsen, R., Nielsen, M., Sporring, J. (eds.) MICCAI 2006. LNCS, vol. 4190, pp. 752–759. Springer, Heidelberg (2006)
3. Lin, H., Shafran, I., Murphy, T., Okamura, A., Yuh, D., Hager, G.: Automatic detection and segmentation of robot-assisted surgical motions. In: Duncan, J.S., Gerig, G. (eds.) MICCAI 2005. LNCS, vol. 3749, pp. 802–810. Springer, Heidelberg (2005)
4. Lo, B., Darzi, A., Yang, G.: Episode classification for the analysis of tissue/instrument interaction with multiple visual cues. In: Ellis, R.E., Peters, T.M. (eds.) MICCAI 2003. LNCS, vol. 2878, pp. 230–237. Springer, Heidelberg (2003)
5. Speidel, S., Benzko, J., Krappe, S., Sudra, G., Azad, P., Müller-Stich, B., Gutt, C., Dillmann, R.: Automatic classification of minimally invasive instruments based on endoscopic image sequences. In: SPIE Medical Imaging, vol. 7261 (2009)
6. Padoy, N., Blum, T., Feußner, H., Berger, M.O., Navab, N.: On-line recognition of surgical activity for monitoring in the operating room. In: Innovative Applications of Artificial Intelligence (2008)
7. Bhatia, B., Oates, T., Xiao, Y., Hu, P.: Real-Time Identification of Operating Room State from Video. In: Innovative Applications of Artificial Intelligence, pp. 1761–1766 (2007)
8. Padoy, N., Mateus, D., Weinland, D., Berger, M.O., Navab, N.: Workflow monitoring based on 3d motion features. In: ICCV Workshop on Video-oriented Object and Event Classification (2009)
9. Bouarfa, L., Jonker, P., Dankelman, J.: Surgical context discovery by monitoring low-level activities in the OR. In: MICCAI Workshop on Modeling and Monitoring of Computer Assisted Interventions, M2CAI (2009)
10. Padoy, N., Blum, T., Essa, I., Feußner, H., Berger, M.O., Navab, N.: A boosted segmentation method for surgical workflow analysis. In: Ayache, N., Ourselin, S., Maeder, A. (eds.) MICCAI 2007, Part I. LNCS, vol. 4791, pp. 102–109. Springer, Heidelberg (2007)
11. Hardoon, D., Szedmak, S., Shawe-Taylor, J.: Canonical correlation analysis: an overview with application to learning methods. Neural Computation 16(12), 2639–2664 (2004)
12. Zhou, F., de la Torre, F.: Canonical time warping for alignment of human behavior. In: Neural Information Processing Systems Conference, NIPS (2009)
13. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.: The WEKA Data Mining Software: An Update. SIGKDD Explorations 11(1) (2009)
14. Varadarajan, B., Reiley, C., Lin, H., Khudanpur, S., Hager, G.: Data-Derived Models for Segmentation with Application to Surgical Assessment and Training. In: Yang, G.-Z., Hawkes, D., Rueckert, D., Noble, A., Taylor, C. (eds.) MICCAI 2009. LNCS, vol. 5761, pp. 426–434. Springer, Heidelberg (2009)