Machine learning and coresets for automated real-time video segmentation of laparoscopic and robot-assisted surgery

Citation	Volkov, Mikhail, Hashimoto, Daniel A., Rosman, Guy, Meireles, Ozanan R. and Rus, Daniela. 2017. "Machine learning and coresets for automated real-time video segmentation of laparoscopic and robot-assisted surgery."
As Published	10.1109/icra.2017.7989093
Publisher	IEEE
Version	Author's final manuscript
Citable link	https://hdl.handle.net/1721.1/137252
Terms of Use	Creative Commons Attribution-Noncommercial-Share Alike
Detailed Terms	http://creativecommons.org/licenses/by-nc-sa/4.0/

Machine Learning and Coresets for Automated Real-Time Video Segmentation of Laparoscopic and Robot-Assisted Surgery

Mikhail Volkov [1] and Daniel A. Hashimoto [2] and Guy Rosman [1] and Ozanan R. Meireles [2] and Daniela Rus[1]

***Abstract*— Context-aware segmentation of laparoscopic and robot assisted surgical video has been shown to improve performance and perioperative workflow efficiency, and can be used for education and time-critical consultation. Modern pressures on productivity preclude manual video analysis, and hospital policies and legacy infrastructure are often prohibitive of recording and storing large amounts of data.**

In this paper we present a system that automatically generates a video segmentation of laparoscopic and robot-assisted procedures according to their underlying surgical phases using minimal computational resources, and low amounts of training data. Our system uses an SVM and HMM in combination with an augmented feature space that captures the variability of these video streams without requiring analysis of the non-rigid and variable environment. By using the data reduction capabilities of online *k*-segment coreset algorithms we can efficiently produce results of approximately equal quality, in real-time. We evaluate our system in cross-validation experiments and propose a blueprint for piloting such a system in a real operating room environment with minimal risk factors.

I. Introduction

Video-based coaching and debriefing of laparoscopic and robot-assisted minimally invasive surgery (RMIS) has been demonstrated to contribute to enhanced surgical performance [3, 27]. These procedures are typically taught in a stepwise fashion by identifying distinct steps or *phases* or an operation [11]. Context-aware segmentation of surgical video can facilitate education [25], surgical coaching [9, 27], post-operative reviews [11], time-critical consultation, and can improve perioperative workflow efficiency [9, 17] of operating room assignment and turnover.

While recording laparoscopic and robotic surgical procedures is easy, analyzing them is a time-consuming process, usually done manually. Modern pressures on training and productivity preclude spending hours viewing and editing surgical video for the purpose of routine video-based coaching or performance review. Moreover, strict hospital policies and legacy infrastructure are often prohibitive of recording and storing large amounts of video data as a matter of routine. Thus, a key focus of this project was a design that performs well a minimal amount of annotation. First, interviewing surgeons is expensive and time-consuming, so it is important to develop a protocol that is not disruptive to the surgeons' work if the project is to succeed. Second, developing such a system allows us to easily train for new surgical procedures with little effort. Lastly, these system requirements are applicable beyond surgical video to many robotic systems, such as a video summary captured from autonomous vehicles, modeling robotic behavior in a fixtureless assembly operation, remotely supervised robots used in space exploration and emergency response scenarios, etc.

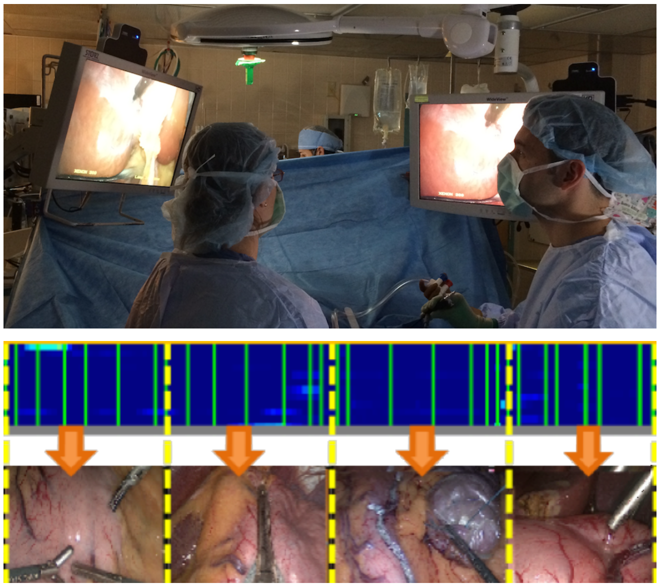


Fig. 1: Laparoscopic cholecystectomy performed at MGH (top). Coreset segmentation of recorded laparoscopic procedure and the corresponding surgical phases detected by our system (bottom).

In this study, we present an online phase recognition system for automatic phase segmentation and identification of laparoscopic surgical video in real-time. Our system is fast, accurate, efficient and scalable, while requiring minimal training data. We focus on laparoscopic video recorded by surgeons instead of robotic video for several reasons. First, human laparoscopic procedures are more common, more highly controlled, and rely on fewer computer-assisted variables, all of which makes them more challenging in principle. Second, the end results of our work are directly applicable to robot-assisted surgery (compare Fig. 4(5) vs Fig. 4(v)), with the additional benefit of being able to rely on multi-sensor data, such as instrument information, from the robotic surgical unit.

We employ coreset algorithms for video segmentation [26] and summarization [31] to reduce large amounts of raw data to a small input to our system. A coreset for the *k*-segment mean problem enables us to compute temporal segmentation of a video stream based on a predetermined feature space representation. The end result is a small subset of video frames that capture the semantic content of the video. We show how coresets allow us to process a large amount

[1] Computer Science and Artificial Intelligence Laboratory, MIT, Cambridge, MA {mikhail,rosman,rus}@csail,mit.edu

[2] Department of Surgery, Massachusetts General Hospital, Boston, MA dahashimoto@partners.org, ozmeireles@mgh.harvard.edu

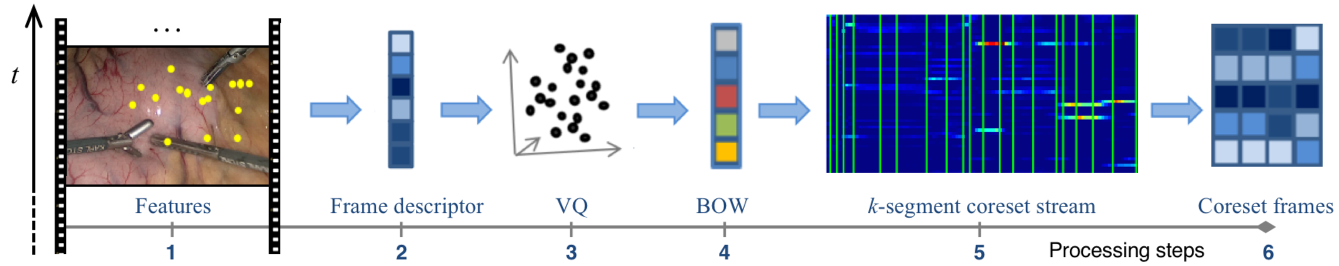


Fig. 2: System overview (left to right) – video stream; feature extraction from video frames; descriptor representation (see Fig. 3); vector quantization; bag-of-words representation; *k*-segment coreset reduction; and finally, the coreset frames presented as input to our system.

of unclassified data with minimal computational overhead, while still providing guarantees about the classification rates.

We use priors from the relevant features and temporal changes, and without relying on methods such as instrument detection, 3D modeling, spatial geometry, etc. To this end we leverage the technical knowledge of expert surgeons to design a feature space that captures the principal axes of variability and other visual discriminant factors for the specific surgical video domain. Using well-established machine learning methods trained on minimal ground truth data, we show that we can segment, summarize, and classify a surgical video according to its constituent phases with a high level of accuracy, on par or better than previous work.

A. *Related Work*

The general problem of automated video segmentation has been researched extensively (see [26], and the references therein). There has been a lot of research on surgical phase recognition [2, 13, 16, 21, 28], but this work has been mostly limited to offline video of entire procedures.

In [18, 22] labeled instrument signals from the OR were used as low-level tasks to infer corresponding surgical high-level tasks and showed that instruments are used for maneuvers other than their primary function. Similarly, [10, 22] focused on investigating manual maneuvers and low-level surgical tasks to infer corresponding surgical high-level tasks by detecting specific hand and instrument maneuvers in laparoscopic surgery

The authors in [29] used instrument signals for phase recognition, such as those provided by da Vinci; [6, 30] investigated instrument detection; [1] looked at instrument pose estimation; [12, 19, 23, 24] investigated tool tracking. Our work is agnostic to specific instruments, focusing instead on lower-level hybrid feature spaces that are based on annotations from experienced surgeons.

In [28] 3D models of instruments were used to train the model. In [5, 20] tools were used as binary signals indicating the progress of laparoscopic procedures, In [32] the authors used metabolizable fluorescent markers attached to the target organ to guide a 2D/3D intra-operative registration algorithm. By contrast, we avoid relying on any ad-hoc detection signals, in order to keep our system general and applicable to multiple types of surgical (and robotic) contexts.

For temporal models, all of [2, 4, 8, 25] all used Hidden Markov Models (HMMs) as the method of choice, although [2] used canonical-correlation analysis (CCA), and [8] also looked at Hidden Semi-Markov Models (HSMM) with AdaBoost. In this work, we used the standard HMM because it is relatively simple and has fewer parameters.

On the clinical side, the studies that motivate the opening paragraph of this Section are comprehensive in their own right and convey the urgent need for the kind of systems that we present in this paper.

B. *Key Contributions*

The main contributions of this paper are:

1) We present a real-time algorithm for surgical phase segmentation and identification, which requires very little training data, and is fast and efficient.
2) We demonstrate the effectiveness and robustness of our algorithms on real medical data. Training on a total dataset of just over 10 hours of video, we are able to achieve a 92.8% prediction accuracy.
3) We show that by using coresets we can reduce the data to a small but informative subset that yields practically identical levels of accuracy.

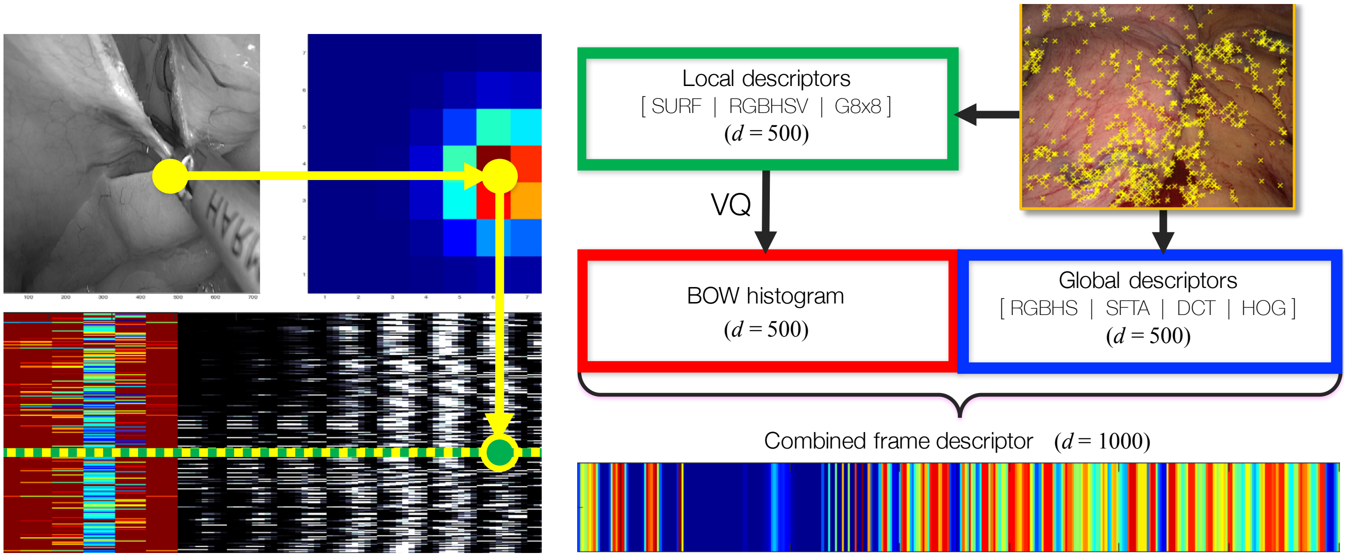
II. Technical Approach

We now describe in detail the technical approach used to enable our end-to-end system.

A. *Features.*

When performing temporal segmentation, the choice of per-frame features is crucial. Since we aim at learning from a limited amount of data and under small computational resource, it is important that our feature space is concise and representative, with a balance between invariance and expressiveness.

Instead of trying to model specific 3D objects in the videos (see for example [23], and references therein), we favor a more generic approach for several reasons. Foremost, semantically important objects have an in-class variability at the geometric level – consider the set of shapes that form of a chair or a cup, all of which have the same function and semantic meaning. While for some objects, large-scale datasets allow direct learning of the appearance space, this is prohibitive in the case of few training examples and/or computational resources, and may result in overfitting. Furthermore, exactly 3D geometry and pose are not always obtainable. Consider specular, metallic, semi-transparent objects, or non-rigid objects – all of which are plentiful inside a patient undergoing surgery. Moreover, videos such as



(a) Clockwise: a SURF keypoint x; Gaussian surface $G(x)$ centered at x; each descriptor encodes SURF coefficients, color values, and location $G(x)$ coefficients for a single keypoint.

(b) Local descriptors (green) are projected using a VQ to produce a fixed-dimension BOW histogram (red); additional global descriptors (blue) are then computed for the entire frame, and concatenated with the BOW to produce the combined frame descriptor.

Fig. 3: Construction of the augmented local descriptors (3a), augmented global descriptors (3b), and the resulting frame representation.

laparoscopic and first-person view videos as are often seen by robots tend to have partial views – this is often encountered in detecting of instruments and objects, as noted by [23].

For these reasons we look for and identified several visual cues in the videos, categorized broadly as *local* and *global* descriptors. These are motivated by the way surgeons deduce the stage of the surgery.

We use these cues to compose a feature space that captures the principal axes of variability and other discriminant factors that determine the phase, and then train a set of classifiers as an intermediate feature. We now describe these visual cues, the augmented descriptor structure, and the final frame representation using the bag-of-words (BOW) model.

(i) *Color.* Histogram intersection has been used in similar work to extract color-oriented visual cues by creating a training image database of positive and negative images [13]. Other descriptor categories for individual RGBHSV channels can be utilized to increase dimensionality to discern features that depend on color in combination with some other property. Pixel values can also be used as features directly [2]. In this work, we use RGB/HSV components to augment both the local descriptor (color values) and global descriptor (color histogram).

(ii) *Position.* Relative position of organs and instruments is an important visual cue. We encode the position of SURF-detected keypoints with an 8×8 grid sampling of a Gaussian surface centered around the keypoint (Fig. 3a). The variance of the Gaussian defines the spatial “area of influence” of a keypoint.

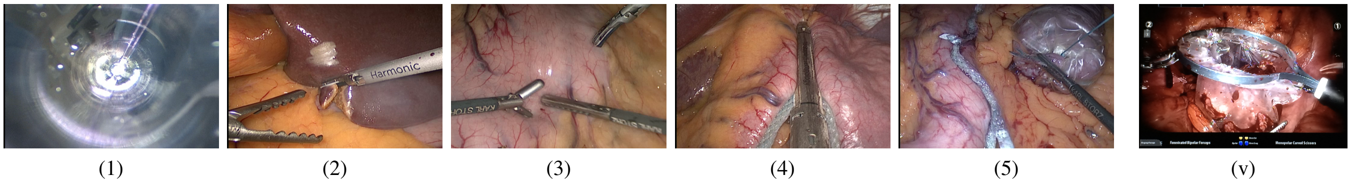
(iii) *Shape.* Shape is important for detecting instruments, which are some of most obvious visual cues for identifying the phase. Shape can be encoded with various techniques, such as the Viola-Jones object detection framework [1], using image segmentation to isolate the instruments and match against artificial 3D models [28], and other methods. For local frame descriptors we use the standard SURF descriptor as a base, and for global frame descriptor we add grid-sampled HOG descriptors [1] and DCT coefficients [1].

(iv) *Texture.* Texture is a crucial visual cue to distinguish vital organs, which tend to exhibit a narrow variety of color. Texture can be extracted using a co-occurrence matrix with Haralick descriptors [14], by a sampling of representative patches to be evaluated with a visual descriptor vector for each patch [13], and other methods. In this work, we use the newer SFTA texture descriptor [7], which has shown better performance than Haralick filter banks.

Finally, we combine the augmented descriptors into a single fixed-dimension frame descriptor. For this we use the BOW model, which is is a simplifying representation commonly used to standardize the dimensionality of features [1]. We compute a representative vector quantization (VQ) by sampling frames using local descriptors only. Any set of local descriptors (Fig. 3a, bottom) can then be represented as a histogram of projections in the fixed VQ dimension ($d_1 = 500$). The final combined frame descriptor is then composed of the BOW histogram and the additional dimensions ($d_2 = 500$) of the global descriptor (Fig. 3b, top), for a combined dimension $d = 1000$ (Fig. 3b, bottom).

B. Coreset Computation

Coresets are compact data reduction constructs that can efficiently produce a problem dependent compression of the data. Specifically, in [26], k-segment coresets have been shown to great aid efficient segmentation of large-scale, online video stream. While in our case the segmentation cost is slightly different, as we show in Section III, we can obtain guarantees on the approximation accuracy afforded by the coresets, and trade off data approximation with computa-



(1) (2) (3) (4) (5) (v)

Fig. 4: Phases (1–5 shown) of the laparoscopic sleeve gastrectomy (LSG) procedure: (1) port, (2) biopsy, liver retraction, (3) omentum removal, dissection, hiatus inspection, (4) stapling, (5) bagging, (6) irrigation, (7) final inspection, withdrawal. Fig. 4(v) shows the bagging phase from a similar surgery performed with a da Vinci Surgical System (compare with Fig. 4(5)).

tional resources. Furthermore, as in our previous work [31], we can augment the coreset by a keyframe compression of the video that enables fast retrieval and anytime access for large visual stream. We therefore we use an online k-segment coreset algorithm to compute a compact representation of the video stream over which we can compute the segmentation. This allows our system to run online, in real-time, using minimal computational resources.

C. Phase prediction.

A binary classifier setup for each phase was used as opposed to a multi-class classifier. Using this approach was preferred in order to decouple *phase transitions*, which is the main goal of the classifier layer, from *phase identification*, which is the goal of the HMM temporal model.

As a first step, we train a series of support vector machines for each phase. Each SVM classifies a phase i by outputting a binary variable $p_i = 1,\ P\backslash\{p_i\} = 0$. This approach was shown to be more accurate than a single multi-class SVM in a similar visual domain [14]. This is an iterative step that involves interviewing surgeons, re-calibrating the feature space, re-training the classifiers, etc. Interviewing surgeons is expensive and time-consuming therefore it is important to repeat this step first until we achieve the desired level of accuracy. The first step is to ensure that the augmented feature space presented in Section II yields an acceptable level of accuracy for this domain with respect to the ground truth phases. Fig. 5a shows the binary outputs produced by the SVM. Fig. 5b shows the rate of correct classification (accuracy) of the predictions compared against ground truth. We note that there are two ambiguous cases: (i) multiple SVM outputs $y_i(t) = 1$; and (ii) all SVM outputs $y_i(t) = 0$.

The second step is to make use of the temporal structure of surgical phases (monotonically increasing, mutually exclusive, and collectively exhaustive) to correct SVM predictions, resolve the ambiguous cases stated above, and compute a single time-series of phase predictions. We achieve this using an observation function $\phi(V, s, \alpha, \beta)$ that takes a sequence of SVM outputs V, the current phase hypothesis s, a certainty parameter $\alpha \in [0,1]$, and lookback parameter $\beta \in \mathbb{Z}^+$, and returns the next phase prediction. We start with an initial phase hypothesis $s = 0$. Then, given the current phase hypothesis and a matrix V where V_i is the column vector of SVM outputs at time i, the current phase estimate p is determined by

$$\phi(V,s,\alpha,\beta) = \begin{cases} \arg\max_i \sum V_{i,1,\dots,\beta}, & \text{if } \sum V_i / \sum V > \alpha \\ s, & \text{otherwise} \end{cases} \tag{1}$$

The intuition for this function is as follows. We have a matrix consisting of several independent SVM output vectors (with a memory trail of the last β-1 such vectors). The observation function then updates the current phase, if and only if the vector sum for another phase exceeds the current one by a certainty threshold α, in which case we update our phase hypothesis to the next phase – otherwise the current phase persists.

The observation function combines many independent SVM outputs into one single set of phase observations. The values of α and β are essentially high- and low-pass filter parameters, and it is trivial to show correctness by considering that both $\alpha, \beta \in [0,1]$.

Phase transitions are modeled using an HMM with the left-right restriction as in [13]. This function is non-restrictive in terms of skipping phases and going backwards (thus violating our assumptions of the phases' temporal structure), but this is not necessary to enforce in the classification layer, and it is resolve by the HMM. The final observation sequence $Q = p_1, \dots, p_N$ is the emission sequence. Finally, we run the Viterbi algorithm [1] on the emission sequence to find the most likely sequence of hidden states (the phases).

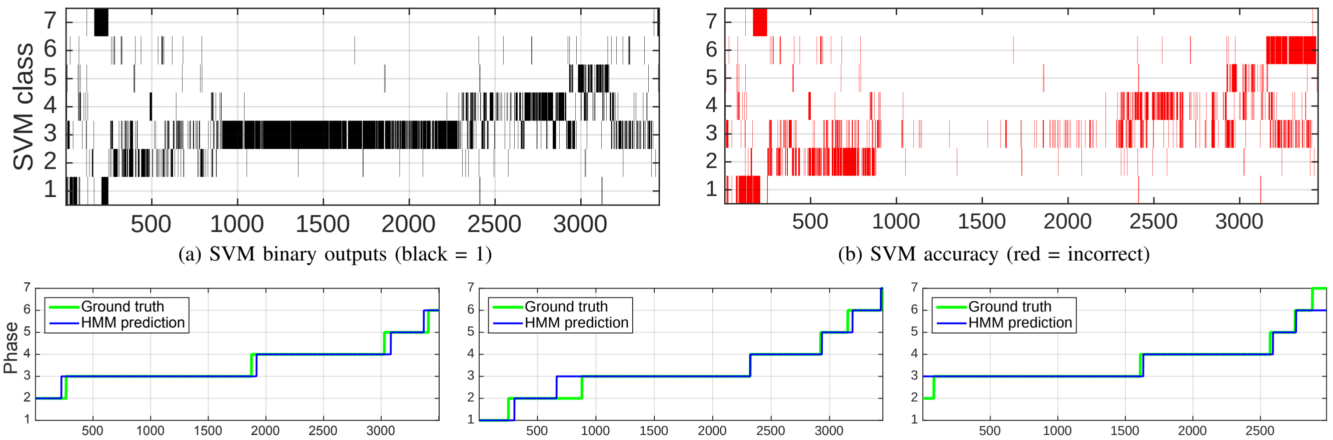
III. Analysis

While coresets have been used before for video summarization [26] and loop-closure [31], there is the question of what kind of guarantees are provided by coresets under the SVM/HMM model that we are using. The following theorem shows that under specific conditions, k-segment coresets can be used in order to efficiently compute the data log-likelihood of solutions, including the optimal solution, subject to constraints on the number of location of the transition boundaries between labels. This includes linear classifiers, or classifiers whose training hinges on linear classification, such as SVMs. We give a brief motivation and outline for the proof, and refer the reader to [26] for more details on the specific coreset used.

For a given coreset segment C in the k-segment coreset, it can be shown [26] that for k-segment with segments $\{x(t) = a_j t + b_j\}, x(t), a_j, b_j \in \mathbb{R}^d$, the coreset provides a good approximation for the fitting cost:

$$(1-\varepsilon) \leqslant \frac{\sum_j d^2\left(C_j, a_j t_i + b_j\right)}{\sum_i d^2\left(x\left(t_i\right), a_j t_i + b_j\right)} \leqslant (1+\varepsilon), \tag{2}$$

where C_j denotes the sufficient statistic matrix saved for each coreset segment j, i denotes the data points indices inside the



(a) SVM binary outputs (black = 1)

(b) SVM accuracy (red = incorrect)

(1) Example 1

(2) Example 2

(3) Example 3

Fig. 5: Typical experimental results: x = frame number, y = SVM class, i.e. phase number. 5(a–b) highlight the shortfalls of an SVM-only system with no temporal component; red lines indicate incorrect phase predictions from (possibly conflicting) SVM outputs. 5(1–3) shows HMM-corrected predictions.

segment or over the whole time axis, and d denotes the least-squares distance. ε is a small constant that depends on the coreset construction parameters and can be adapted. (here we assume the k-segments to align with the coreset segments, although this is not required in [26].)

Our main motivation is linear classifiers — if we assume a linear observation / classification model, of the form:

$$\log P\left(x\left(t_i\right) \mid H_j\right) = w_j^T x\left(t_i\right) + d_j, \tag{3}$$

Where H_j is a hypothesis on which phase we are in at time t_i, and w_j, d_j are the parameters of the appropriate linear classifier. Assuming a state doesn't change within a coreset segment, the log-probability of the observations for all times during the coreset segment, $t_i \in T$, is also well-approximated,

$$(1-\varepsilon) \le \frac{\sum_j f(d(C_j, w_j, d_j)}{\sum_{t_i \in T} \log P\left(x\left(t_i\right) \mid H_j\right)} \le (1+\varepsilon). \tag{4}$$

f defines here the computation of the log probability using the linear classifier parameters and the coreset for the data in segment j. This can be seen by substituting

$$w_j^T x + d_j = d^2(x, -w_j/2) - \|w_j/2\|^2 - \|x\|^2 + d_j. \tag{5}$$

into the guarantees in [26] and the definition of the coreset there, as well as adding a correcting term for the norm of this vector and the bias vector. Assuming the signal is separable into a single class per segment with a non-zero margin gives a straightforward result.

The same justification can be used for other methods based on linear classification followed by non-linear operators, such as the one shown in Subsection II.3. This allows us to estimate the label per segment with bounded increase in the classification error, $\sum_i \frac{1}{2}\left(1 - y_i \,\mathrm{sgn}\left(w_j^T x_i + d_j\right)\right)$. For a video whose frames are classified using the SVM, it can be shown that by assuming a single label for each coreset segment and using the coreset segment C to obtain the sign leads to a bounded increase in the error. We again assume here the number of different phases is bounded by k and that the phase is separable.

IV. Experimental Results

For this study we used 10 videos of the laparoscopic vertical sleeve gastrectomy (LSG) procedure performed by expert surgeons at the MGH. This allowed us to test the system with enough variability between data examples, while training the features and low-level classifier channels under the assumption of limited training data.

For this procedure, the surgeons identified 7 basic phases: (1) port, (2) biopsy, liver retraction, (3) omentum removal, dissection, hiatus inspection, (4) stapling, (5) bagging, (6) irrigation, (7) final examination, withdrawal (Fig. 4). We note that some phases have multiple stages, and a much finer granularity is generally possible. In-fact, some very complicated procedures such as The Whipple Procedure (pancreaticoduodenectomy) can have more than a hundred identifiable phases, wherein a single misstep can result in morbidity and mortality [17].

In the surgeries obtained in this study, phases always occur in the specified order. We also note that not all videos contain all the phases, which presents an additional challenge to segment videos with missing phases. We then interviewed the surgeons who performed the procedures, and collected two kinds of information: (1) Qualitative annotations describing how they identified the phase from the video; (2) Specific timestamps of phase transitions that serve as ground truth.

The qualitative annotations are (in principle) innumerable, consisting of natural language descriptions of the surgical process. Conversely, we note that the timestamp annotations are very sparse. We only have $k-1$ indices as the entire annotation for an k-phase procedure.

We assess our system with cross-validation experiments, using both the entire video and the coreset representation, and evaluate accuracy against ground truth segmentation. We perform tests by training the system on each subset of $N-1=9$ videos in the dataset, using a standard 80/20 training/validation split. The system is then tested on each remaining unseen video, and the results aggregated over the N subsets. Fig. 5.1–5.3 shows typical results. We demonstrate a 90.4% SVM prediction accuracy, and improve to 92.8%

when combined with HMM. These results are on par with similar work in the surgical video domain [15], while achieving a 90+% coreset compression over the original video.

A. Discussion and Conclusions

In this work we showed how a carefully calibrated feature space can facilitate a very high classification accuracy for phase detection in video streams

The main focus of ongoing work is to improve the phase prediction accuracy. We extend our system to consider continuous likelihood models, allowing us use temporal regularity to handle ambivalent phase predictions more effectively. We are looking into other temporal models for non-monotonic phase sequences. With more video data we aim to evaluate our predictive model across different surgical procedures. Lastly, our goal is to extend the use of coresets in this work beyond segmentation, applying the framework we presented in [31] to generate an interactive visual summary of laparoscopic and robot-assisted surgeries.

References

[1] M. Allan, P.-L. Chang, S. Ourselin, D. J. Hawkes, A. Sridhar, J. Kelly, and D. Stoyanov. Image based surgical instrument pose estimation with multi-class labelling and optical flow. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 331–338. Springer, 2015.

[2] T. Blum, H. Feußner, and N. Navab. Modeling and segmentation of surgical workflow from laparoscopic video. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 400–407. Springer, 2010.

[3] E. M. Bonrath, N. J. Dedy, L. E. Gordon, and T. P. Grantcharov. Comprehensive surgical coaching enhances surgical skill in the operating room: a randomized controlled trial. *Annals of surgery*, 262(2):205–212, 2015.

[4] L. Bouarfa, P. Jonker, and J. Dankelman. Surgical context discovery by monitoring low-level activities in the or. In *MICCAI workshop on modeling and monitoring of computer assisted interventions (M2CAI). London, UK*, 2009.

[5] L. Bouarfa, P. P. Jonker, and J. Dankelman. Discovery of high-level tasks in the operating room. *Journal of biomedical informatics*, 44(3):455–462, 2011.

[6] D. Bouget, R. Benenson, M. Omran, L. Riffaud, B. Schiele, and P. Jannin. Detecting surgical tools by modelling local appearance and global shape. *IEEE transactions on medical imaging*, 34(12):2603–2617, 2015.

[7] A. F. Costa, G. Humpire-Mamani, and A. J. M. Traina. An efficient algorithm for fractal analysis of textures. In *Graphics, Patterns and Images (SIBGRAPI), 2012 25th SIBGRAPI Conference on*, pages 39–46. IEEE, 2012.

[8] O. Dergachyova, D. Bouget, A. Huaulmé, X. Morandi, and P. Jannin. Automatic data-driven real-time segmentation and recognition of surgical workflow. *International journal of computer assisted radiology and surgery*, pages 1–9, 2016.

[9] B. L. Ecker, R. Maduka, A. Ramdon, D. T. Dempsey, K. R. Dumon, and N. N. Williams. Resident education in robotic-assisted vertical sleeve gastrectomy: outcomes and cost-analysis of 411 consecutive cases. *Surgery for Obesity and Related Diseases*, 2015.

[10] K. Kahol, N. C. Krishnan, V. N. Balasubramanian, S. Panchanathan, M. Smith, and J. Ferrara. Measuring movement expertise in surgical tasks. In *Proceedings of the 14th ACM international conference on Multimedia*, pages 719–722. ACM, 2006.

[11] U. Kannan, B. L. Ecker, R. Choudhury, D. T. Dempsey, N. N. Williams, and K. R. Dumon. Laparoscopic hand-assisted versus robotic-assisted laparoscopic sleeve gastrectomy: experience of 103 consecutive cases. *Surgery for Obesity and Related Diseases*, 2015.

[12] M. Kranzfelder, A. Schneider, A. Fiolka, E. Schwan, S. Gillen, D. Wilhelm, R. Schirren, S. Reiser, B. Jensen, and H. Feussner. Real-time instrument detection in minimally invasive surgery using radiofrequency identification technology. *journal of surgical research*, 185(2):704–710, 2013.

[13] F. Lalys, L. Riffaud, D. Bouget, and P. Jannin. A framework for the recognition of high-level surgical tasks from video images for cataract surgeries. *IEEE Transactions on Biomedical Engineering*, 59(4):966–976, 2012.

[14] F. Lalys, L. Riffaud, X. Morandi, and P. Jannin. Automatic phases recognition in pituitary surgeries by microscope images classification. In *International Conference on Information Processing in Computer-Assisted Interventions*, pages 34–44. Springer, 2010.

[15] F. Lalys, L. Riffaud, X. Morandi, and P. Jannin. Surgical phases detection from microscope videos by combining SVM and HMM. In *Medical Computer Vision. Recognition Techniques and Applications in Medical Imaging*, pages 54–62. Springer, 2010.

[16] B. P. Lo, A. Darzi, and G.-Z. Yang. Episode classification for the analysis of tissue/instrument interaction with multiple visual cues. In *MICCAI*, pages 230–237. Springer, 2003.

[17] G. Marangoni, G. Morris-Stiff, S. Deshmukh, A. Hakeem, and A. M. Smith. A modern approach to teaching pancreatic surgery. *Journal of gastrointestinal surgery*, 16(8):1597–1604, 2012.

[18] N. Mehta, R. Haluck, M. Frecker, and A. Snyder. Sequence and task analysis of instrument use in common laparoscopic procedures. *Surgical Endoscopy And Other Interventional Techniques*, 16(2):280–285, 2002.

[19] T. Neumuth and C. Meißner. Online recognition of surgical instruments by information fusion. *International journal of computer assisted radiology and surgery*, 7(2):297–304, 2012.

[20] N. Padoy, T. Blum, S.-A. Ahmadi, H. Feussner, M.-O. Berger, and N. Navab. Statistical modeling and recognition of surgical workflow. *Medical image analysis*, 16(3):632–641, 2012.

[21] N. Padoy, T. Blum, I. Essa, H. Feussner, M.-O. Berger, and N. Navab. A boosted segmentation method for surgical workflow analysis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 102–109. Springer, 2007.

[22] N. Padoy, T. Blum, H. Feussner, M.-O. Berger, and N. Navab. On-line recognition of surgical activity for monitoring in the operating room. In *AAAI*, pages 1718–1724, 2008.

[23] A. Reiter, P. K. Allen, and T. Zhao. Feature classification for tracking articulated surgical tools. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 592–600. Springer, 2012.

[24] N. Rieke, D. J. Tan, M. Alsheakhali, F. Tombari, C. A. di San Filippo, V. Belagiannis, A. Eslami, and N. Navab. Surgical tool tracking and pose estimation in retinal microsurgery. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 266–273. Springer, 2015.

[25] J. Rosen, M. Solazzo, B. Hannaford, and M. Sinanan. Task decomposition of laparoscopic surgery for objective evaluation of surgical residents' learning curve using hidden markov model. *Computer Aided Surgery*, 7(1):49–61, 2002.

[26] G. Rosman, M. Volkov, D. Feldman, J. W. Fisher III, and D. Rus. Coresets for k-segmentation of streaming data. In *NIPS*, pages 559–567. Curran Associates, Inc., 2014.

[27] P. Singh, R. Aggarwal, M. Tahir, P. H. Pucher, and A. Darzi. A randomized controlled study to evaluate the role of video-based coaching in training laparoscopic skills. *Annals of surgery*, 261(5):862–869, 2015.

[28] S. Speidel, J. Benzko, S. Krappe, G. Sudra, P. Azad, B. P. Müller-Stich, C. Gutt, and R. Dillmann. Automatic classification of minimally invasive instruments based on endoscopic image sequences. In *SPIE Medical Imaging*, pages 72610A–72610A. International Society for Optics and Photonics, 2009.

[29] R. Stauder, A. Okur, L. Peter, A. Schneider, M. Kranzfelder, H. Feussner, and N. Navab. Random forests for phase detection in surgical workflow analysis. In *International Conference on Information Processing in Computer-Assisted Interventions*, pages 148–157. Springer, 2014.

[30] R. Sznitman, C. Becker, and P. Fua. Fast part-based classification for instrument detection in minimally invasive surgery. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 692–699. Springer, 2014.

[31] M. Volkov, G. Rosman, D. Feldman, J. W. Fisher III, and D. Rus. Coresets for visual summarization with applications to loop closure. In *ICRA*, Seattle, Washington, USA, May 2015. IEEE.

[32] E. Wild, D. Teber, D. Schmid, T. Simpfendörfer, M. Müller, A.-C. Baranski, H. Kenngott, K. Kopka, and L. Maier-Hein. Robust augmented reality guidance with fluorescent markers in laparoscopic surgery. *International journal of computer assisted radiology and surgery*, pages 1–9, 2016.