2014 IEEE International Conference on Robotics & Automation (ICRA)
Hong Kong Convention and Exhibition Center
May 31 - June 7, 2014. Hong Kong, China

Surgical Tool Attributes from Monocular Video

Suren Kumar[1], Madusudanan Sathia Narayanan[1], Pankaj Singhal[2], Jason J Corso[3] and Venkat Krovi[4]

***Abstract*— HD Video from the (monocular or binocular) endoscopic camera provides a rich real-time sensing channel from surgical site to the surgeon console in various Minimally Invasive Surgery (MIS) procedures. However, a real-time framework for video understanding would be critical for tapping into the rich information-content provided by the non-invasive and well-established digital endoscopic video-streaming modality. While contemporary research focuses on enhancing aspects such as tool-tracking within the challenging visual scenes, we consider the associated problem of using that rich (but often compromised) streaming visual data to discover the underlying semantic attributes of the tools.**

Directly analyzing the surgical videos to extract more realistic attributes online can aid in the decision-making and feedback aspects. We propose a novel probabilistic attribute labelling framework with Bayesian filtering to identify associated semantics (open/closed, stained with blood etc.) to ultimately give semantic feedback to the surgeon. Our robust video-understanding framework overcomes many of the challenges (tissue deformations, image specularities, clutter, tool-occlusion due to blood and/or organs) under realistic in-vivo surgical conditions. Specifically, this manuscript performs rigorous experimental analysis of the resulting method with varying parameters and different visual features on a data-corpus consisting of real surgical procedures performed on patients with da Vinci Surgical System [9].

I. INTRODUCTION

Increasingly, surgical procedures are being performed using MIS techniques which rely on the endoscopic camera to provide a rich real-time sensing channel from surgical site to the surgeon console. However, the rich information content (often already in digital form as real-time HD video) is underutilized in current surgical procedures. The "last-mile" still remains the analog rendering of the digital image-stream back to the eyeballs of the surgical team. A real-time framework for video capture, processing and understanding (building upon the non-invasive and well-established digital endoscopic video modality) is critical to reaching the goal of intelligent intra-operative surgical assistance. Significant research efforts (surveyed later) are already under-way focusing on locating and tracking critical elements (e.g tools) within the visual scene to enhance situational awareness. In contrast, however, in this manuscript we consider the associated problem of using that rich streaming visual data to discover the underlying semantic attributes of the tools.

Semantic attributes (in the domain of minimally-invasive surgery) broadly can now include information about the state of the surgical tools, environment and tool-environment interactions. Knowledge of semantic attributes (such as tool-operational state, blood-stained state, holding tissue or not, cauterization state etc.) would be an essential step in the path towards autonomous robotic surgery. Some example attributes are open/closed, stained with blood or not, holding tissue/without it, state of cauterizing tools etc. Furthermore, this type of attribute labelling can potentially provide an additional layer of safety for the state-of-art human-in-the-loop surgical robotic systems. Semantic attribute feedback in real-time would be beneficial to avoid critical failures and possible surgical errors due to surgeon's lack of experience and/or situation-awareness, inappropriate operation or communication between master-slave type systems [11], [17].

With the commercial success of Intuitive robotic surgery platform [9], a variety of surgical robotic systems, such as RAVEN [20], DLR MiroSurge [10], with widely varying architectures and instrumentation are being developed. However, as teleoperated devices, all examples are fitted with one (or more) camera(s) to provide the surgeon with visual feedback [27]. Hence, our semantic attribute understanding framework (built solely upon sensed visual information) would have wider applicability. Gaining semantic knowledge directly from the actual videos in surgical settings becomes important from multiple perspectives and more specifically for identifying the surgical gestures and providing a context specific surgical feedback.

The principal focus of this work is to estimate two specific (binary) semantic attributes of tools, namely open/closed states and blood-stain condition of tools using only a monocular video stream. Figure 1 summarizes our algorithm. Given a video with bounding boxes for tools, we first extract visual features and adapt the probabilistic Support Vector Machine (SVM) formulation to learn a visual attribute scoring function. We feed output of this function to a novel Bayesian tracking framework to maintain accurate and smooth estimates of the semantic attributes. *To the best of our knowledge, this is the first work that performs an online probabilistic semantic attribute labeling and tracking from visual data alone.*

*This work is in part supported by Kaleida Health Western New York, the UB Foundation Bruce Holm Fund and NSF IIS-1319084

[1]Suren Kumar and Madusudanan Sathia Narayanan are PhD Candidates with Department of Mechanical Engineering, University at Buffalo, SUNY Buffalo NY 14260 USA, email:{surenkum,ms329}@buffalo.edu

[2]Pankaj Singhal is the Director of Robotic Surgery, Kaleida Health Western New York, Buffalo NY 14214 USA, email:psinghal@buffalo.edu

[3]Jason J. Corso is an Associate Professor in Department of Computer Science Engineering, University at Buffalo, SUNY, Buffalo NY 14260 USA, email: jcorso@buffalo.edu

[4]Venkat Krovi is an Associate Professor in Department of Mechanical Engineering and an Adjunct Professor in Department Obstetrics and Gynecology, University at Buffalo, SUNY Buffalo NY 14260 USA, email:vkrovi@buffalo.edu

978-1-4799-3684-7/14/$31.00 ©2014 IEEE

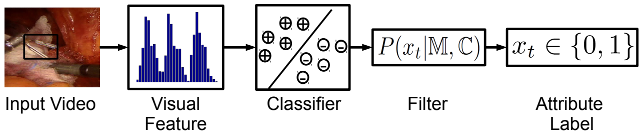


Fig. 1: Flow chart of the attribute classification algorithm

Such probabilistic labelling not only enables trajectory smoothing but on a higher level can prove to be valuable for decision making using risk minimization techniques [2]. The proposed attribute labelling algorithm can be extended for multi-valued semantic labelling across different platforms for robotic surgery and surgical procedures using monocular or stereo videos. Preliminary results and parts of this idea have been presented as part of poster-sessions and workshops [15]–[17].

II. Related Work

With the ubiquitous presence of video streams in surgical operations, there has been considerable research in surgical video understanding [1], [8], [14], [22], [24], [28], [29], [32]. Prior research has however focussed principally on detecting and tracking surgical tools from videos. Holistic surgical video understanding requires finding relationships between tools and their immediate environment. Surgical tool detection and tracking, as it has been considered so far, can be used only for low level feedback tasks such as vision based control [14].

Predominant approaches have employed fiducial markers on the tool to assist in tool detection and tracking. Groeger et. al [8] use color markers and thresholding in HSV space to find the tool. Wei et. al [32] color code the tips of surgical instruments and use a simple color segmentation to detect tools. Krupa et. al [14] attach light emitting diodes to the tip of instruments and then detect these markers in endoscopic images. Other non-interfering approaches include using color space for classification of pixels into instruments and organs, doing shape analysis of these classification labels and then predicting the location of the tool in the next frame using an auto regressive Model [29]. In [22], a similar method is used for surgical tool classification which uses particle filter for tracking instruments in a video. Recently there has been some work on locating specific landmarks on the surgical tools, learning a Random Forest based classifier to classify these landmarks from images and using an Extended Kalman Filter (EKF) to smooth the tool poses [24].

Despite the importance of learning to identify semantic attributes, there has been little work in the surgical vision literature to address this issue. The sub-problem of semantic attribute labelling that has been most studied and used is tool-tissue interaction. Rosen et al. [25] used manually labelled video data of tool-tissue interaction together with force data for evaluating surgical expertise. API data from Intuitive Surgical da Vinci platform has been used by Lin et al. [19] to detect surgical gestures in a suturing task. Voros et. al [31] extract carefully defined features using API and vision data from the da Vinci robot to recognize binary tool-tissue interaction (no interaction/-interaction). Other semantic attributes such as tool open/closed have been used for automated analysis of surgical skills [12]. Jun et. al [12] find the tool open/closed state using the 3D information from stereoscopic video. Semantic attributes for ensuring patient safety such as clipping of the cystic duct in laparoscopic cholecystectomy has been performed using eye-gaze tracking and visual features [11]. There also has been work on using classifiers on visual features to identify 5 different application specific binary cues to discriminate various high-level surgical tasks in cataract surgery [18]. In contrast to the existing state-of-the art, we propose using only the monocular stream with generic visual features to probabilistically extract the semantic attributes.

III. Attribute Labeling

To recognize semantic attributes, we are inspired from work done in the computer vision literature [7] that identifies shape, part and material attributes for different types of objects by learning a SVM classifier [4] from visual features. We extract features from the image that capture color, shape and texture information. We perform a comprehensive evaluation of these features to assess classification performance in surgical settings. After extracting features from the image of a tool, we use a probabilistic SVM (PSVM) classifier [23] to calculate the probability of a tool being in one of the classified states. Furthermore, a different classifier will be learned for a different attribute so that the classification is always a binary problem. Unlike a standard two class SVM classifier, which only gives a classification decision by locating a test point in the feature space with respect to a separating hyperplane, a PSVM gives a probability measure based on the distance from the separating hyperplane. Let the feature vector be given as, $y \in R^n$ and the class variable be represented by x. Standard linear SVM learns a hyperplane ($f(y) = h^T y + b$) in feature space which separates two classes, where h is the normal vector to the hyperplane and b is proportional to the offset of hyperplane from the origin. The sign function is used for getting the classification decision for any new test point in feature space y_i.

$$f(y_i) > 0 \implies x_i = 1 \tag{1}$$

$$f(y_i) < 0 \implies x_i = 0 \tag{2}$$

But indeed $f(y_i)$ is proportional to the perpendicular distance of the test point y_i from the separating hyperplane and thus can be used as a measure of confidence of this point belonging to a particular class. Platt [23] uses this idea to propose directly estimating a posterior probability $P(class|input)$ by fitting a sigmoid after the SVM as in Equation 3.

$$P(x = 1|f) = \frac{1}{1 + \exp(Af + B)} \tag{3}$$

The parameters A and B are learned by minimizing the cross-entropy error function.

Unlike image based attributes research in vision community, in robotic surgery one has access to a real-time stream of video data. This streaming temporal data can be leveraged to ensure the smoothness of attribute labelling during a change in state of the tools attribute. We hence propose a Bayesian tracking framework to ensure the smoothness of attribute labelling by leveraging the posterior probabilities given by the classifier. Let us denote the state of a tool attribute at time t as $x_t \in \{0, 1\}$ which can be 0 (e.g: tool closed) or 1 (e.g: tool open). Probabilistic SVM classifier $\mathbb{C}$ gives $P_{\mathbb{C}}(x = 0|y_t)$ and $P_{\mathbb{C}}(x = 1|y_t)$ for an image observation y_t at time step t. From now on, we simplify notation by representing this probability as $P(x = 1|\mathbb{C})$ where $\mathbb{C}$ is the classifier to explicitly denote the posterior probability as given by the classifier. Probability of the other class is simply $P(x = 0|\mathbb{C}) = 1 - P(x = 1|\mathbb{C})$. But since the decision of the classifier is only based on the current image observation, it does not include any prior information about the state of the tool attribute. Hence this decision is only based on the current image and is inherently noisy as demonstrated empirically in Section V.

IV. Attribute Filtering

However as noted in [31], surgical gestures are usually continuous and seldom change abruptly. The authors perform a smoothing operation after getting a rigid classifier class decision ($x_t \in \{0, 1\}$) by using an simple "opening" mathematical morphology operation. Morphological operation is heuristic and doesn't take any model information or discrete label probability into account. In contrast, we use Bayesian tracking to maintain this smoothness by conditioning the class decision on motion ($\mathbb{M}$) probability ($P(x_t = 1|\mathbb{M})$) by using data from τ previous frames. Here $P(x_t = 1|\mathbb{M})$ is the simplified notation for $P_{\mathbb{M}}(x_t = 1|y_{t-\tau}, ..., y_{t-1})$ which specifies the probability of state in current time step belonging to class 1 given the previous τ image observations. Using Bayes rule at time step t and assuming all observations are independent and the current observation only depends on the current state, one can write,

$$P(x_t = 1|\mathbb{M}, \mathbb{C}) = \frac{P(\mathbb{M}, \mathbb{C}|x_t = 1)P(x_t = 1)}{P(\mathbb{M}, \mathbb{C})} \tag{4}$$

where $P(x_t = 1)$ models the prior probability of a tool having a particular attribute state, $P(\mathbb{M}, \mathbb{C}|x_t = 1)$ is the likelihood of observing a motion and classifier producing a certain decision given that tool has state $x_t = 1$ and $P(x_t = 1|\mathbb{M}, \mathbb{C})$ is the posterior probability by considering both motion and classifier on the current frame. We can write a similar equation for the tool attribute being in an other state $x_t = 0$. Since we are only interested in the classification decision we ignore the denominator since it will be same for $P(x_t = 1|\mathbb{M}, \mathbb{C})$ and $P(x_t = 0|\mathbb{M}, \mathbb{C})$. Motion $\mathbb{M}$ models the transition from previous frames to current frame and classifier $\mathbb{C}$ models only the current frame, so they can be assumed to be independent.

$$\begin{aligned} P(x_t = 1|\mathbb{M}, \mathbb{C}) &\propto P(\mathbb{M}, \mathbb{C}|x_t = 1)P(x_t = 1) \\ &\propto P(\mathbb{M}|x_t = 1)P(\mathbb{C}|x_t = 1)P(x_t = 1) \end{aligned} \tag{5}$$

Frame	$P(x = 0\|\mathbb{C})$	$P(x = 1\|\mathbb{C})$
1	0.6	0.4
2	0.65	0.35
3	0.6	0.4
4	0.33	0.67
5	0.6	0.4

TABLE I: Probabilities for attribute classification as given by classifier without smoothing

Further using Bayes rule,

$$\begin{aligned} P(x_t = 1|\mathbb{M}, \mathbb{C}) \propto &\frac{P(x_t = 1|\mathbb{M})P(\mathbb{M})}{P(x_t = 1)} \times \frac{P(x_t = 1|\mathbb{C})P(\mathbb{C})}{P(x_t = 1)} \\ &\times P(x_t = 1) \\ \propto &\frac{P(x_t = 1|\mathbb{M})P(\mathbb{M})P(x_t = 1|\mathbb{C})P(\mathbb{C})}{P(x_t = 1)} \end{aligned} \tag{6}$$

Again, ignoring the component not useful for classification decision and without making any assumptions about the prior probability $P(x_t = 1)$ in Equation 6, we get

$$P(x_t = 1|\mathbb{M}, \mathbb{C}) \propto P(x_t = 1|\mathbb{M})P(x_t = 1|\mathbb{C}) \tag{7}$$

One can consider multiple approaches to modelling system dynamics ($P(x_t = 1|\mathbb{M})$) using Markov process dynamics [2]. We use information from previous frames as given by classifier which still does not violate the independence assumption in Equation 5 because now we consider frames from $t - \tau$ to $t - 1$. We model motion probability by using Bayes rule and probability product rule, taking τ frames prior to current observation to model smoothness which is inherently present in the motion of a surgical tool.

$$\begin{aligned} P(x_t = 1|\mathbb{M}) &\propto \prod_{i=t-\tau}^{t-1} P(x_i = 1|\mathbb{C}) \\ P(x_t = 1|\mathbb{M}, \mathbb{C}) &\propto \prod_{i=t-\tau}^{t} P(x_i = 1|\mathbb{C}) \end{aligned} \tag{8}$$

In Equation 8, the probabilities at each time step $P(x_i = 1|\mathbb{C})$ are directly given by probabilistic SVM classifier. This in effect weights the classification decision by not only the current observation but also by the decision in previous frames. The final derived rule for classification in Equation 8 is similar to the widely known, *"the product rule"* of combining a set of base classifiers [6] with the exception that we are using same classifiers but over different time frames which effects the results in current frame because of motion smoothness.

This rule also resembles Dempster-Shafer theory to combine evidence from different sources by combining probability masses from multiple independent knowledge bases [26]. Another important factor to consider is the nature of probability in Equation 3 which uses sigmoid function. This function ensures that $P(x = 1|f)$ is not zero for any finite value in feature space and hence outliers will not affect the results obtained using Equation 8. Table I - II shows the effect of this smoothing.

Frame	$P(x=0\vert\mathbb{M},\mathbb{C})$	$P(x=1\vert\mathbb{M},\mathbb{C})$
1	0.6	0.4
2	0.65	0.35
3	0.8069	0.0.1931
4	0.5784	0.4216
5	0.5257	0.4743

TABLE II: Probabilities for attribute classification as given by classifier with motion-smoothed. The classifier gives wrong probability for Frame 4 as in Table I which is corrected in motion-smoothed probability by taking $\tau = 2$

V. Experiments

A new dataset consisting of short sequences of real human surgeries performed using da Vinci Surgical System (dVSS) platform is herein proposed to conduct our evaluation studies, as there is no publicly available dataset to the best of our knowledge that is suitable for testing such attribute labelling. The proposed dataset has 23 short sequences with each sequence consisting of 150-300 frames (30 frames per second) for a total of 4500 frames. These sequences were ensured to have various artefacts including tool articulations, occlusions, rapid appearance changes, smoke and specular reflections. We annotate the bounding box of the tools and the corresponding attributes in every frame. In the proposed dataset there are a total of $16,956$ tool attribute annotations with 8028 annotations of blood stained and 8928 instances of open/close attribute. The bounding box size varies frame to frame because the proposed dataset has tool articulations. Entire dataset with code is available to encourage further research on this problem at `http://mechatronics.eng.buffalo.edu/research/rMIS_SkillAssessment/index.html`.

Since the focus of this work is not on detection and tracking of tools, we assume the input for a classifier to be tool bounding box. We propose learning for attribute recognition using SVM classifier based on different features. We test the baseline SVM classification by getting the class using the probabilistic measure $x_t = i$ if $P(x_t = i|y_t) > P(x_t = j|y_t), i, j \in \{0, 1\}$. Both attributes are evaluated using various different visual features. We normalize all the features to have 0 mean and 1 standard deviation before learning the classifier to reduce the computation time for learning the classifier. The overall accuracy of our method is then evaluated using standard performance measures [21] by calculating True Positive (TP), False Positive (FP), True Negatives (TN) and False Negatives (FN). We compare just the filtering part of our algorithm with the smoothing method proposed by Voros et. al [31]. We use 10-fold cross validation [13] to evaluate our method by dividing the entire dataset into 10 parts, and use 9 parts for training and one part for testing. Overall accuracy is the evaluated as average of the testing results for all 10 possible combinations of training and testing.

A. Visual Features

We extract different types of features [30] that capture various visual aspects of an image patch. We test our attribute labelling algorithm using Histograms of Oriented Gradients (HOG) [5], Pyramidal Histograms of Oriented Gradients (PHOG) [3], Local Binary Patterns (LBP) and Red-Green-Blue (RGB) intensity histograms. HOG captures appearance and shape within an image by representing distribution of edge directions in uniformly divided image cell. For experiments in the current paper, we resize the image to 64×64, choose a image cell size of 16 pixels and the number of undirected orientations in orientation histogram as 9, yielding a 496 dimensional vector. PHOG captures the histogram of orientation gradients at various image pyramids. For PHOG computation, we choose 8 bins for angles to specify angles from 0^0 to 360^0 and 2 pyramid levels to get an image feature of vector dimension 168. LBP captures the texture in an image patch by comparing each pixel to its 8 surrounding neighbours and producing a binary code based on the comparison results. For our experiments we use 32 pixels as image cell size for a 64×64 resized image to get 232 dimensional vector sized feature. RGB histograms are extracted as 10 dimensional histogram in each color channel. Dimensions of these features were chosen heuristically to balance between computational complexity in learning the classifier and obtaining good attribute classification.

Feature	Baseline	Ours	[31]
RGB	94.05%	**94.36**%	94.03%
HOG	97.01%	**97.19**%	97.02%
PHOG	93.84%	**94.89**%	93.96%
LBP	96.33%	**97.21**%	96.33%

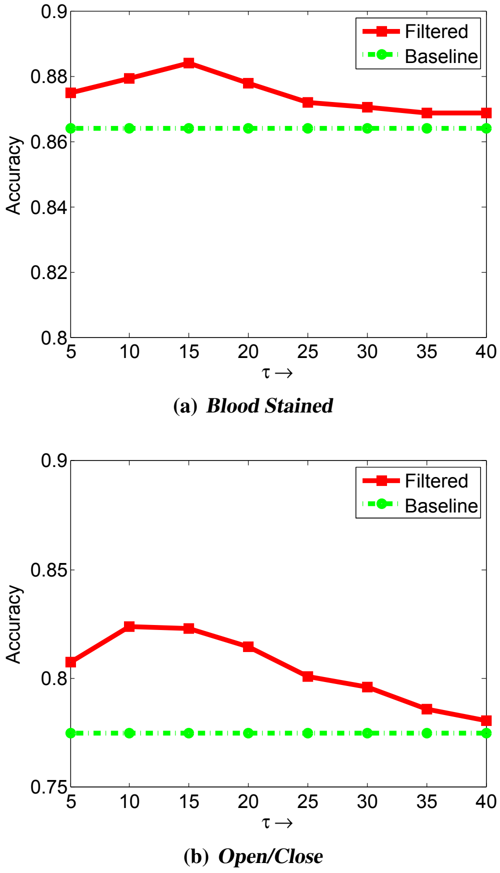
TABLE III: Accuracy Scores for Blood Stained attribute using Baseline Classifier and Tracking Classifier with $\tau = 15$

B. Parameter Selection

The only significant parameter used in current paper is τ. τ captures degree of smoothness expected in a attribute. For understanding the effect of this parameters for both the attributes, we choose a subset of our dataset to choose the value of τ. Figure 2 shows the improved accuracy of filtered results compared to the baseline results for different values of τ, evaluated for blood stain condition attribute using only RGB Histograms and for tool open/close attribute using PHOG features. Clearly there is an optimal value of τ that can be selected empirically for various attributes. Figure 3 shows the improvement using the selected parameter on a *single video* to illustrate the performance. Baseline classifier using RGB histograms on this video for blood stain attribute yields an accuracy of 79.26% with 38 TP, 69 TN, 19 FP, 9 FN, while smoothed classifier with $\tau = 15$ generate accuracy of 90.37% with 45 TP, 77 TN, 11 FP, 2 FN. As can be seen from Figure 3, smoothed results eliminate the jumps present in classifier labels.

C. Feature Evaluation

All the visual features are used to learn a baseline probabilistic classifier for the attributes as described in Section III. Furthermore, all the learned classifier models are filtered to evaluate the performance for both the attributes. To compare



(a) ***Blood Stained***

(b) ***Open/Close***

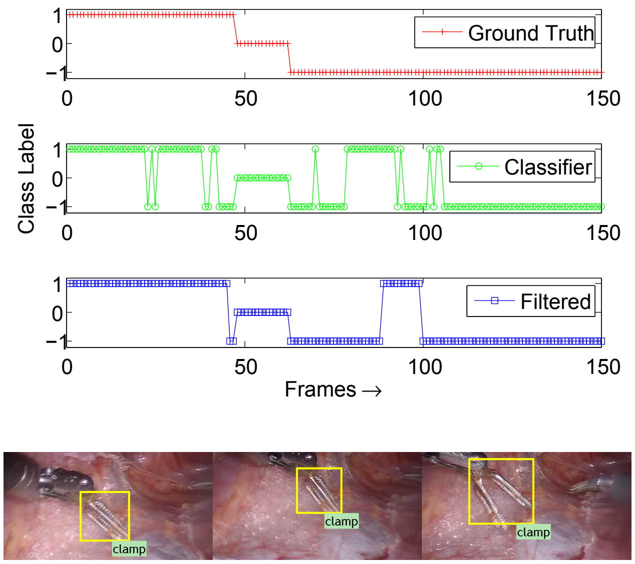
Fig. 2: Accuracy of baseline classifier and Filtered result. For blood stained attribute τ is chosen as 15 while for Open/Close problem τ is chosen as 10

Feature	Baseline	Ours	[31]
RGB	82.64%	**84.05%**	82.84%
HOG	92.51%	**93.43%**	92.51%
PHOG	86.68%	**88.47%**	86.74%
LBP	89.00%	**91.35%**	88.94%

TABLE IV: Accuracy Scores for Open/Close attribute using Baseline Classifier and filtered Classifier compared with with $\tau = 10$

just the filtering part of our framework, we also implemented the "opening" operation with SVM classification as input as proposed in [31]. The method in [31] uses information from future frames while the method proposed in this paper only uses information from past and current frames.

For both the attributes (open/close, blood stained), the HOG feature yields the best baseline semantic attribute classification performance. However after filtering the results of base classifier, LBP marginally outperforms the HOG feature for the blood stained attribute. These performance results are expected as HOG captures the intensity gradients of an image patch while LBP represents the texture in an image patch. As seen in tables III and IV, all the accuracy scores show an improvement upon filtering the baseline results. Sometimes the improvement is marginal because the baseline classifier itself obtains close to 100% accuracy on many videos due to which the increase in average accuracy score is marginal. For example consider the case of HOG feature based attribute results for Blood stained attribute in Table III. Even though the best improvement in accuracy for a video is from 68.66% to 74.00%, the overall increase in accuracy upon averaging is less because most videos have close to 100% accuracy.



(b) ***Frame 1*** **(c)** ***Frame 10*** **(d)** ***Frame 53***

Fig. 3: Ground Truth, Baseline Classifier and Filtered Results with few frames for a Video

D. Computation Time

The proposed method primarily coded in MATLAB^{TM} is faster than real time on a video stream with 30 frames per second using a 1.6 GHz desktop machine. Most of the computation is involved in learning a baseline probabilistic classifier which is done offline. All the visual features except PHOG are coded in C++ [30] for faster evaluation speed.

VI. Conclusion

We presented an algorithm for probabilistic attribute labelling and tracking in real human surgical videos. The proposed framework was successfully tested for two different attributes using various features on a challenging dataset. The proposed attribute filtering algorithm improves results of a probabilistic SVM classifier by smoothing which is expected as a natural outcome in streaming surgical video. The filtering method is highly extensible and can use different types of features, used on different types of attributes and use any classifier that generates posterior class probability.

References

[1] Max Allan, Sébastien Ourselin, Steve Thompson, David J Hawkes, John Kelly, and Danail Stoyanov. Toward detection and localization of instruments in minimally invasive surgery. *IEEE Transactions on Biomedical Engineering*, 60(4):1050–1058, 2013.

[2] Christopher M Bishop. *Pattern recognition and machine learning*, volume 4. springer New York, 2006.

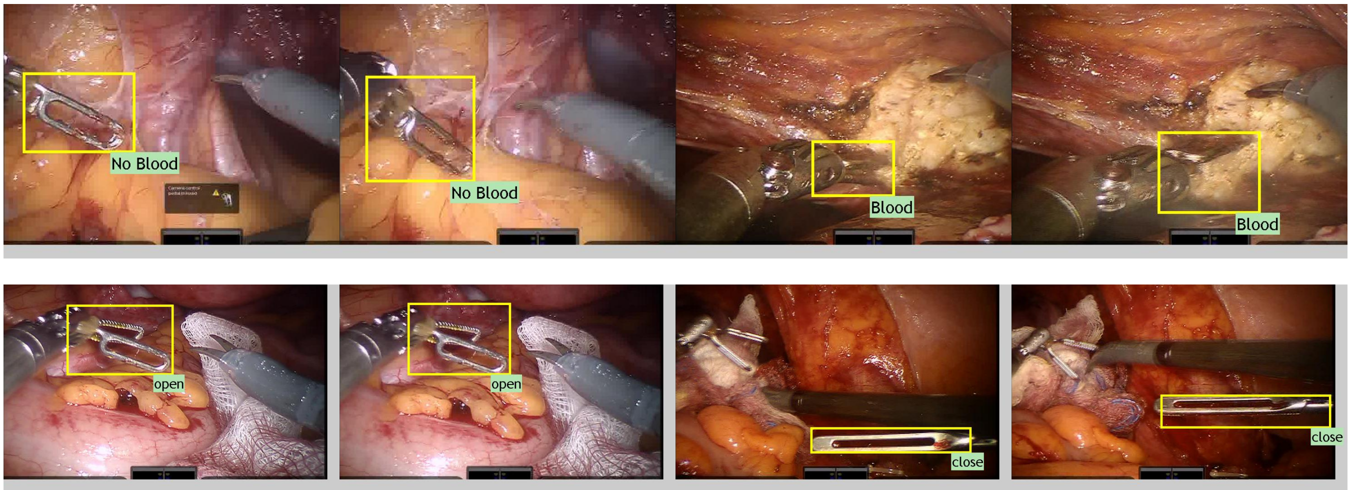


Fig. 4: Representative results of open/close and blood stained attribute for videos in the proposed dataset obtained using HOG based probabilistic attribute classifier

[3] A. Bosch and A. Zisserman. Pyramid histogram of oriented gradients (phog). *University of Oxford Visual Geometry Group, http://www. robots. ox. ac. uk/vgg/research/caltech/phog. html*.

[4] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[5] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *CVPR*, volume 1, pages 886–893. IEEE, 2005.

[6] Robert PW Duin. The combining classifier: to train or not to train? In *ICPR*, volume 2, pages 765–770. IEEE, 2002.

[7] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *CVPR*, pages 1778–1785. IEEE, 2009.

[8] M. Groeger, K. Arbter, and G. Hirzinger. Motion tracking for minimally invasive robotic surgery. *Medical Robotics, I-Tech Education and Publishing*, pages 117–148, 2008.

[9] Gary S Guthart and John Kenneth Salisbury Jr. The intuitive tm telesurgery system: overview and application. In *ICRA*, volume 1, pages 618–621. IEEE, 2000.

[10] Ulrich Hagn, Rainer Konietschke, Andreas Tobergte, Mathias Nickl, Stefan Jörg, Bernhard Kübler, Georg Passig, Martin Gröger, Florian Fröhlich, Ulrich Seibold, et al. Dlr mirosurge: a versatile system for research in endoscopic telesurgery. *International journal of computer assisted radiology and surgery*, 5(2):183–193, 2010.

[11] Adam James, D Vieira, Benny Lo, Ara Darzi, and G Yang. Eye-gaze driven surgical workflow segmentation. *MICCAI*, 2007.

[12] S.K. Jun, M.S. Narayanan, P. Agarwal, A. Eddib, P. Singhal, S. Garimella, and V. Krovi. Robotic minimally invasive surgical skill assessment based on automated video-analysis motion studies. In *BioRob*, pages 25–31. IEEE, 2012.

[13] Ron Kohavi et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *IJCAI*, volume 14, pages 1137–1145, 1995.

[14] A. Krupa, J. Gangloff, C. Doignon, M.F. de Mathelin, G. Morel, J. Leroy, L. Soler, and J. Marescaux. Autonomous 3-d positioning of surgical instruments in robotized laparoscopic surgery using visual servoing. *IEEE Tran. Rob. and Auto.*, 19(5):842–853, 2003.

[15] Suren Kumar, Madusudanan Sathia, Sukumar Misra, Sudha Garimella, Pankaj Singhal, Jason J Corso, and Venkat Krovi. Video-analytics for enhancing safety & decision-support in surgical workflows. In *Frontiers in Medical Devices*. ASME, 2013.

[16] Suren Kumar, Madusudanan Sathia, Sukumar Misra, Sudha Garimella, Pankaj Singhal, Jason J Corso, and Venkat Krovi. *Vision based Decision-Support and Safety Systems for Robotic Surgery*. Medical Cyber Physical Systems Workshop (MEDCPS), 2013.

[17] Suren Kumar, Madusudanan Sathia Narayanan, Sukumar Misra, Sudha Garimella, Pankaj Singhal, Jason Corso, and Venkat Krovi. Video-based framework for safer and smarter computer aided surgery. In *The Hamlyn Symposium on Medical Robotics*, pages 107–108, 2013.

[18] Florent Lalys, Laurent Riffaud, David Bouget, and Pierre Jannin. An application-dependent framework for the recognition of high-level surgical tasks in the or. *MICCAI 2011*, pages 331–338, 2011.

[19] Henry Lin, Izhak Shafran, Todd Murphy, Allison Okamura, David Yuh, and Gregory Hager. Automatic detection and segmentation of robot-assisted surgical motions. *MICCAI 2005*, pages 802–810, 2005.

[20] Mitchell JH Lum, Diana CW Friedman, Ganesh Sankaranarayanan, Hawkeye King, Kenneth Fodero, Rainer Leuschke, Blake Hannaford, Jacob Rosen, and Mika N Sinanan. The raven: Design and validation of a telesurgery system. *The International Journal of Robotics Research*, 28(9):1183–1197, 2009.

[21] John Makhoul, Francis Kubala, Richard Schwartz, and Ralph Weischedel. Performance measures for information extraction. In *Proc. DARPA Broadcast News Workshop*, pages 249–252, 1999.

[22] S.J. McKenna, H.N. Charif, and T. Frank. Towards video understanding of laparoscopic surgery: Instrument tracking. In *Proc. of Image and Vision Computing, New Zealand*, 2005.

[23] J. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

[24] A. Reiter, P. Allen, and T. Zhao. Feature classification for tracking articulated surgical tools. *MICCAI 2012*, pages 592–600, 2012.

[25] Jacob Rosen, Massimiliano Solazzo, Blake Hannaford, and Mika 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] Kari Sentz and Scott Ferson. *Combination of evidence in Dempster-Shafer theory*. Citeseer, 2002.

[27] Farzad Soleimani, Fred Moll, Dan Wallace, Jean Bismuth, and Borut Geršak. Robots and medicine–shaping and defining the future of surgery, endovascular surgery, electrophysiology and interventional radiology. *Slovenian Medical Journal*, 80(7-8), 2011.

[28] Raphael Sznitman, Rogerio Richa, Russell H Taylor, Bruno Jedynak, and Gregory D Hager. Unified detection and tracking of instruments during retinal microsurgery. *PAMI*, 35(5):1263–1273, 2013.

[29] D.R. Uecker, YF Wang, C. Lee, and Y. Wang. Laboratory investigation: Automated instrument tracking in robotically assisted laparoscopic surgery. *Computer Aided Surgery*, 1(6):308–325, 1995.

[30] Andrea Vedaldi and Brian Fulkerson. Vlfeat: An open and portable library of computer vision algorithms. In *Proceedings of the international conference on Multimedia*, pages 1469–1472. ACM, 2010.

[31] Sandrine Voros and Gregory D Hager. Towards real-time tool-tissue interaction detection in robotically assisted laparoscopy. In *BioRob*, pages 562–567. IEEE, 2008.

[32] G.Q. Wei, K. Arbter, and G. Hirzinger. Automatic tracking of laparoscopic instruments by color coding. In *CVRMed-MRCAS'97*, pages 357–366. Springer, 1997.