# Instrument Classification in Laparoscopic Videos

Manfred Jürgen Primus and Klaus Schoeffmann and Laszlo Böszörmenyi
Institute of Information Technology, Universität Klagenfurt,
Klagenfurt, Austria
(mprimus | ks | laszlo)@itec.aau.at

***Abstract*—In medical endoscopy more and more surgeons record videos of their interventions in a long-term storage archive for later retrieval. In order to allow content-based search in such endoscopic video archives, the video data needs to be indexed first. However, even the very basic step of content-based indexing, namely content segmentation, is already very challenging due to the special characteristics of such video data. Therefore, we propose to use instrument classification to enable semantic segmentation of laparoscopic videos. In this paper, we evaluate the performance of such an instrument classification approach. Our results show satisfying performance for all instruments used in our evaluation.**

## I. Introduction

Medical endoscopy, also known as keyhole surgery, is a minimally invasive method for therapeutic and diagnostic interventions (procedures and surgeries) in hollow body regions of a human. In the special field of laparoscopic endoscopy the intervention is performed in the abdomen through small incisions. During the intervention the surgeon typically uses several instruments, such as graspers, scissors, clip appliers, and an endoscope. The endoscope contains a tiny camera that generates a high-resolution video signal, which is projected to a large screen in the operation room. This is the only source of information (i.e., vision) for the operating surgeons.

Nowadays more and more surgeons record and archive the video content of endoscopic interventions due to several reasons. First, these videos contain important information for medical doctors in case of follow-up interventions, because the video content shows exactly the same images as seen by the operation team during the intervention. Additionally, the videos can be used as information material for explanations to the patients as well as for training young surgeons. Moreover, some countries (e.g., The Netherlands) enforce the recording and archival of endoscopic videos by law, for reasons of evidence in case of lawsuits. For these reasons, it is important to provide content-based search in endoscopic video archives.

Typically, endoscopic videos are only sparsely annotated or not annotated at all [1], [2]. In order to enable content-based search, the content in an endoscopic video archive needs to be indexed first. The very first step of video indexing is segmentation of the video content into basic units, such as shots [3]. However, recordings from endoscopic interventions contain only a single shot. Moreover, the images in a recorded video, which can be several hours long, are typically highly similar (i.e., contain very similar colors and textures) and have other special characteristics (see [2], [4] for details). Due to these special characteristics common shot detection methods (e.g., color- or texture-based methods, such of the majority of those used in TRECVid over the years) do not work for endoscopic video content. Also motion-based methods quickly reach their limits with this very special kind of video content [4]. More importantly, medical doctors expect the videos to be segmented by semantic boundaries, instead of low-level boundaries at positions of visual changes that are considered by typical shot boundary detection methods.

The basic step for scene segmentation in an endoscopic video is the reliable detection of particular phases of the intervention. For the special field of laparoscopy [5], this can be achieved through recognition of surgery instruments. The appearance of certain instruments (or combinations of them) signals the beginning of a new phase. Therefore, in this paper, we investigate the performance of concept classifiers for the recognition of surgery instruments in laparoscopy. We use Support Vector Machines (SVM) with Bag-of-visual-Words (BoW) learned from densely sampled keypoints of training images. We evaluate three different keypoint descriptors for the purpose of instrument classification in endoscopic video, namely ORB [6], SIFT [7], and SURF [8]. Our results show that for the commonly used set of instruments in laparoscopy, the classifiers achieve quite good results especially with ORB, which significantly outperforms SIFT and SURF in terms of precision. ORB is designed to reliably extract keypoints also for noisy data, which is relevant for endoscopic videos.

## II. Related Work

Object detection and recognition is a widely researched task in the computer vision community. In the field of medicine a large area of research puts great emphasis on the detection and localization of instruments. A major part of the work is dedicated to the detection and localization of instruments — mostly the tips of the instruments — for robot assisted surgeries. One example is the approach described by Allan et al. [9]. They select scale invariant feature transform (SIFT), color-based SIFT, and Histograms of oriented Gradients (HOG) as features. The features are classified with random forests, which provides better accuracy than a normal Bayesian classifier. The classification of images showing the three different areas of a esophagogastroduodenoscopy or polyp images is proposed by Surangsrirat et al. [10]. They use only the gray-level representation of images to train a SVM and achieve a reasonable accuracy. Speidel et al. classifies instruments recorded with a stereo endoscope based on the recognition of the tip of the instruments [11]. First the shaft of the instrument is segmented using a Bayes classifier that calculates the probability that a pixel belongs to a tissue or an instrument. Additionally, they use heuristics like the size of an instrument, the length of a line, and the position of the centroid of a region to calculate the position of the tip. The contour of the instrument is calculated with a chain code algorithm and transformed to a normalized representation. For learning a three-dimensional CAD model an Eigenspace representation is generated, which is used for the recognition of instruments. The instrument detection approach proposed by Marcano-Gamero [12] is somehow similar to the one described in this paper. However, as their paper lacks of details we were not able to perform a comparison nor to reproduce their results. More precisely, in their paper only the performance of the training data is given by $\xi\alpha$-estimates that are used to conservatively bias the parameters of the SVM. Secondly, the choice of a vocabulary with 15 entries lacks the discriminative power of the BoW approach. As stated already by Jiang et al. [13], the typical number of entries in the BoW-dictionary is between 200 and 10,000. Moreover, in a similarly large dataset 5,000 entries show the best performance.

The common part of these works is that they use single image representations but not video sequences, where all potential inconveniences like blur because of an incorrect focus or a dirty lens, motion blur, partial occlusion of instruments, permanent changing position and angle of instruments etc. occur on a regular basis. In this paper, special attention is paid on the effectiveness and efficiency of the BoW and SVM classification toolchain to the recognition of endoscopic instruments in laparoscopic videos.

## III. Experimental Setup

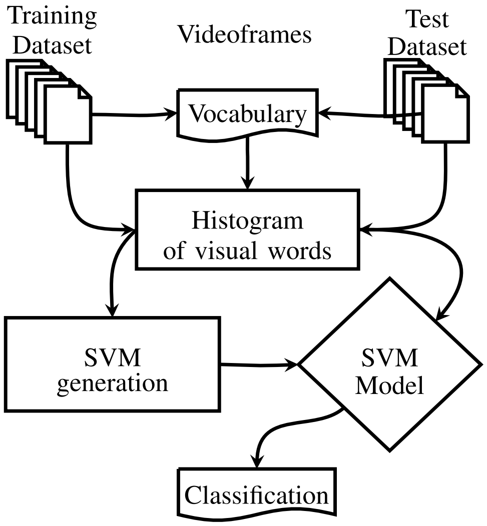


Fig. 1: Object classification tool-chain.

Our implementation follows the classical BoW-SVM-toolchain process described in Figure 1. This toolchain consists of a dataset of videoframes, which is divided into a training and a testing set. Visual features of frames are clustered and the centroids of the clusters build a vocabulary. Features of an image are mapped to the vocabulary and a histogram of visual words, with the vocabulary as basis, is used as representation for this image. The histograms of visual words from the training dataset are used to train an SVM. The resulting SVM-models are used to classify the frames from the test set in order to predict to which class a given frame, i.e., histogram of visual words, belongs to. More recent approaches, like the Bag-of-Regions approach [14] in combination with texture and shape descriptors, remains subject of future work.

### A. Dataset

The dataset consists of videos recorded at laparoscopic cholecystectomies. A cholecystectomy is the highly standardized surgical removal of the gallbladder. We have been provided by our medical partner with an almost day-long recording of this procedure with an overall length of seven and a half hour (more than 680 000 frames). Contained are twelve surgeries with a length between 20 and 60 minutes, which is the typical length of a cholecystectomy without major complications, counted from the insertion of the first trocar to the removal of the trocar used for the insertion of the endoscopic camera. The videos are recorded in full HD resolution (1920x1080) with a framerate of 25 fps and MPEG-2 as video codec. The video bitrate is set to a constant rate of 35 MBit/s. Due to technical issues of the recording system the videos are split into chunks of 1.81 GB, which corresponds to a duration of 7 minutes and 25 seconds.

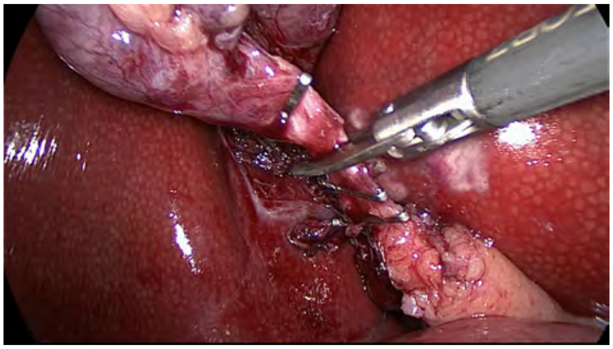
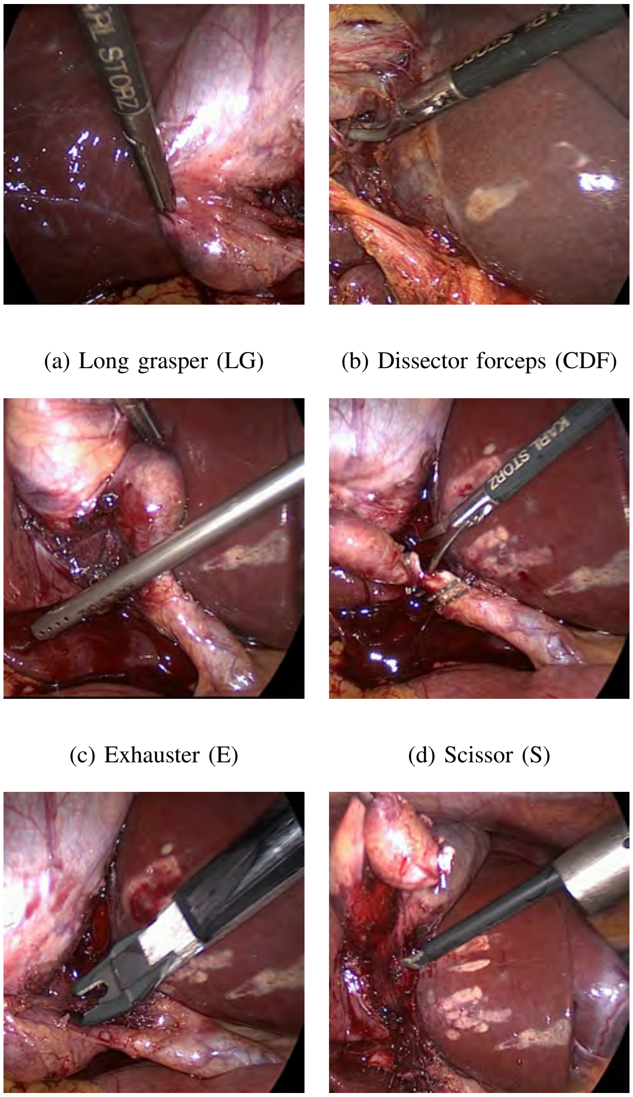


Fig. 2: The critical view of safety is achieved when the cystic artery and duct are separated, clamped and cut. The use of the clipping device and the scissor are explicit markers of this phase.

A key stage in the cholecystectomy is the part where the cystic artery and the cystic duct are clipped and cut as shown in Figure 2. This phase is also called *critical view of safety*. The scene before, during, and after this critical view of safety phase show all of the six instruments, which are used for a cholecystectomy, in an almost unchanged sequence (see the instruments in Figure 3. To limit the computation time for the evaluation reasonable and proportionate we select the 14 video chunks that show this phase. Therefore, the dataset for evaluation consists of about 160 000 frames where more than 225 000 instrument occurrences have to be classified.

Visual content classification in endoscopic video is very challenging due to several problems. The problems can be divided in two groups. One concerns the video-capturing-system, the other group concerns the appearance of the objects. Most of the problems of the first group are caused by the artificial light source fixed next to the lens and the image sensor. These problems are a high amount of noise, reflections on tissues and instruments, interplay of light and shadow. The visible area is heavily magnified. Thus, the instruments are differently scaled when they are moved away from or closer to the lens. Coding issues as block artifacts have also negative impact on the classification quality. The second group concerns the appearance of the instruments. Most of the instruments have an identically looking shaft that does not provide information for the classification task as can be seen in Figure 3. Only the tips of the instruments show discriminative elements. The view of these elements change also when the instruments are rotated along the longitudinal axis. Some of these elements change when the instrument is used, i.e. when the scissor or the grasper is opened or closed. Occlusion of parts of the tip of an instrument leads to classification problems especially when the hidden parts provide specific characteristics compared to similar instruments.



(a) Long grasper (LG)

(b) Dissector forceps (CDF)

(c) Exhauster (E)

(d) Scissor (S)

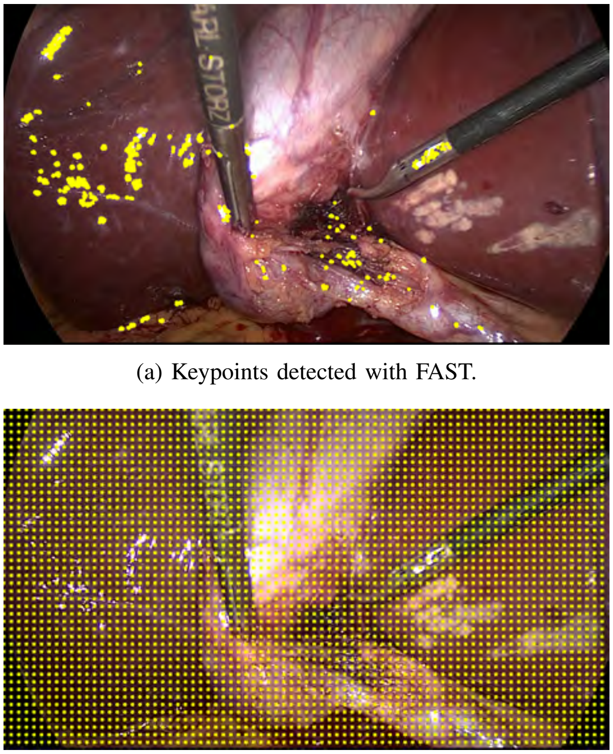
(e) Clipping device (CD)

(f) Monopolar spatula (MS)

Fig. 3: The six instruments used in a cholecystectomy.

### B. *Feature extraction*

A crucial part of the classification tool-chain is the selection of content features. It is important to choose the right feature representation for frames, dependent on the content. After experimenting with a number of different features (e.g. HOG), we chose ORB [6] as a starting point for our experiments. The ORB keypoint descriptor uses “oriented FAST” as keypoint detector. It is a high-performance keypoint detector, locating feature points at reasonable corner points. This creates a serious problem in our case. The illumination unit of the endoscopic camera sits beside the optic of the camera. This leads to large quantities of reflections distributed over the frames throughout the entire video. Keypoints tend to be located at such areas far more likely than on the remaining image as shown in Figure 4. In several parts of the videos the number of keypoints found on instruments (which should be classified) is very low with the default keypoint detector of ORB. To overcome this issue we use dense sampling on a regular grid at distances of 8, 12, 16 and 24 pixels.



(a) Keypoints detected with FAST.

(b) Dense sampled keypoints.

Fig. 4: Almost no keypoint is located on the instruments (a). Dense sampling assure that a sufficient amount of keypoints is positioned on the instruments (b).

Additionally to the classification of instruments based on ORB features we further evaluate scale-invariant feature transform (SIFT) [7] and speeded-up robust feature (SURF) [8]. Both keypoint descriptors are well known state-of-the-art methods and locate feature points on blob like structures. For the same reasons as mentioned above we use dense sampling of the keypoints. The grid size for dense sampling is chosen to be the grid size used for the top result achieved by the use of ORB.

### C. *Vocabulary*

The objective of the BoW approach is the generalization of variations within image descriptors without the loss of accuracy, but with an increase of computational efficiency [15]. The generalization can be done by clustering the features with vector quantization algorithms or clustering algorithms like the improved version of k-means introduced by Arthur et al. in [16]. The evaluation of the experiments in Section IV will show the influence of various sized codebooks to the accuracy of our approach. The sizes of the codebook are chosen to be 512, 1024, 2048 and 4096 entries for the experiments with ORB features. For the experiments with SIFT and SURF we choose the codebook size that is used for the top result using ORB features.

### *D. Support Vector Machine (SVM)*

SVM is a machine learning approach to solve binary classification problems and is primary based on the work of Cortes et al. [17]. In our implementation we use the radial basis function (RBF) as kernel, which is a commonly used initial choice. A typical strategy to use a two-class classifier like the SVM for multiclass classification is the *one-against-all* approach. If there are $c$ classes the result of this approach are $c$ SVM classifiers, which decide if an object belongs to a class or not. To train the $c$-th SVM, labeled examples of instruments are needed. These examples are divided into positive ($p$) and negative ($n$) ones, where $p$ denotes examples showing the objects of the $c$-th class and $n$ include examples, which do not show objects from the $c$-th class.

### *E. Training and testing*

The annotation of the videos, which are used for training and testing, was performed manually for all six tested instruments (see below). For training a classifier for an instrument, we used only frames showing one specific instrument without any other instrument.

The evaluation of this approach has been done using $k$-fold cross-validation. As 14 videos are used it is obvious to choose $k = 7$. Each of the seven partitions is used for testing the SVMs that are trained with the videos of the remaining partitions. The resulting precision, recall and accuracy scores are averaged for each pair of feature density and vocabulary size parameters. Precision is defined as the ratio of the correctly classified occurrences of an instrument (true positive [TP]) to the total number of classifications as this instrument (TP + false positive [FP]), with

$$\text{Precision} = \text{True Positive Accuracy} = \frac{TP}{TP + FP}. \tag{1}$$

Recall defines the ratio of the correctly classified occurrences of an instrument (TP) to the total number of occurrences of this instrument (TP + false negative [FN]), with

$$\text{Recall} = \text{True Positive Rate} = \frac{TP}{TP + FN}. \tag{2}$$

Accuracy is the degree of correct classifications (TP + true negative [TN]) to the total number of classifications of an instrument (TP + FP + TN + FN), with

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}. \tag{3}$$

## IV. Evaluation Results

We have evaluated instrument classifiers for six types of instruments, which are used in laparoscopy: *long grasper* (LG), *curved dissector forceps* (CDF), *exhauster* (E), *scissor* (S), *clipping device* (CD), *monopolar spatula* (MS).

### *A. Performance of ORB*

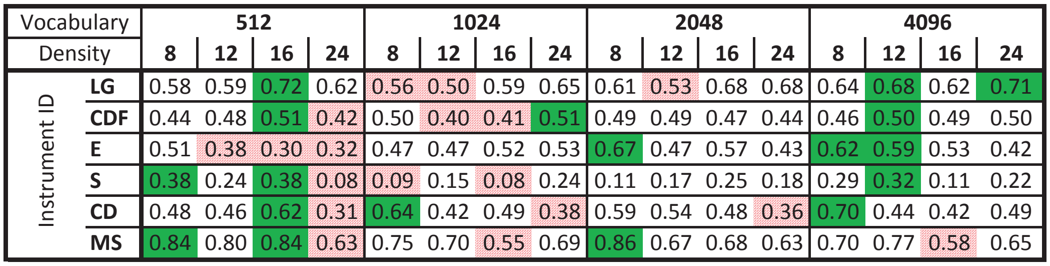
The overall results of the experiments with four different codebooks, four different sample densities of ORB features and an SVM with an RBF kernel do not show a common parameter set that would provide the best performance for all instruments. Table Ia shows the averaged precision values, Table Ib shows the averaged recall values and Table Ic shows the averaged accuracy values from the k-fold cross-validation of the SVMs per instrument by the use of ORB features.

As shown in the tables, three instruments (LG, CD, MS) show pretty good precision values. The MS has the best average precision with 86%, followed by the LG with 72% and the CD with 70% average precision. The reason, why the LG and the MS achieve a better performance could be the shape and surface of the tips of the instruments, which are more unique than the other tips. A major cause for false prediction is the similarity of the surfaces and the shaft of the instruments. If the tips of instruments are covered, it is impossible to distinguish the instruments.

Another problem is the partial occlusion of instrument parts, which often removes some unique features. The exhauster, for example, has four unique holes to absorb and to spout fluids. If the holes are not visible to the camera or occluded by tissue or fluids, the instrument has the form of a shaft only. The rather poor performance of the scissor can be attributed, beside the already mentioned problems, to the smaller number of available training samples in the video data, compared to the number of training samples of the other instruments.

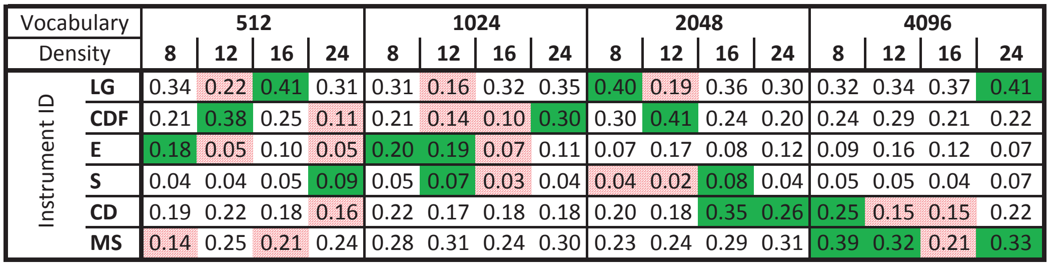
The recall values in Table Ib show values from less than 10% up to values of 41%. The most serious problems are caused by fumes due to cautery, incorrect focus, blur and coverage. The poor performance of the classifier of instrument S can be explained again by the relatively small proportion of occurrence in the videos. The classifier for E shows slightly better results but its perfomance is poor due to occlusion. We suppose that for semantic partition of endoscopic videos based on instrument detection low recall values can be absorbed by considering the temporal component of videos. If, e.g., an instrument is detected for a certain period of time in every third frame then we may assume that a scene with the instrument has been found.

The accuracy of this approach is between 70% and 95% as shown in Table Ic. The expressiveness of the classification of E and S has to be seen in combination with the results in Table Ia (precision) and in Table Ib (recall). As already said before, E and S appear unproportionally less often in the videos than other instruments. Thus, the high accuracy values show that the amount of correctly classified frames, where the instruments do not appear (true negatives), is high. The main reason for not performing better in some cases can be identified by the high amount of falsely negative classified instruments.



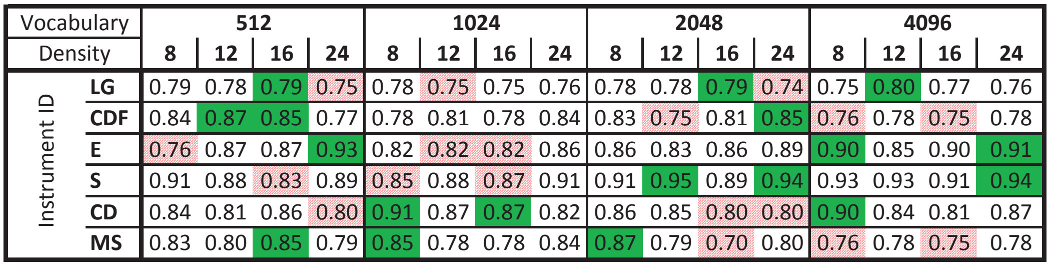
| Vocabulary | | 512 | | | | 1024 | | | | 2048 | | | | 4096 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Density | | 8 | 12 | 16 | 24 | 8 | 12 | 16 | 24 | 8 | 12 | 16 | 24 | 8 | 12 | 16 | 24 |
| Instrument ID | LG | 0.58 | 0.59 | 0.72 | 0.62 | 0.56 | 0.50 | 0.59 | 0.65 | 0.61 | 0.53 | 0.68 | 0.68 | 0.64 | 0.68 | 0.62 | 0.71 |
| | CDF | 0.44 | 0.48 | 0.51 | 0.42 | 0.50 | 0.40 | 0.41 | 0.51 | 0.49 | 0.49 | 0.47 | 0.44 | 0.46 | 0.50 | 0.49 | 0.50 |
| | E | 0.51 | 0.38 | 0.30 | 0.32 | 0.47 | 0.47 | 0.52 | 0.53 | 0.67 | 0.47 | 0.57 | 0.43 | 0.62 | 0.59 | 0.53 | 0.42 |
| | S | 0.38 | 0.24 | 0.38 | 0.08 | 0.09 | 0.15 | 0.08 | 0.24 | 0.11 | 0.17 | 0.25 | 0.18 | 0.29 | 0.32 | 0.11 | 0.22 |
| | CD | 0.48 | 0.46 | 0.62 | 0.31 | 0.64 | 0.42 | 0.49 | 0.38 | 0.59 | 0.54 | 0.48 | 0.36 | 0.70 | 0.44 | 0.42 | 0.49 |
| | MS | 0.84 | 0.80 | 0.84 | 0.63 | 0.75 | 0.70 | 0.55 | 0.69 | 0.86 | 0.67 | 0.68 | 0.63 | 0.70 | 0.77 | 0.58 | 0.65 |

(a) Averaged precision values.



| Vocabulary | | 512 | | | | 1024 | | | | 2048 | | | | 4096 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Density | | 8 | 12 | 16 | 24 | 8 | 12 | 16 | 24 | 8 | 12 | 16 | 24 | 8 | 12 | 16 | 24 |
| Instrument ID | LG | 0.34 | 0.22 | 0.41 | 0.31 | 0.31 | 0.16 | 0.32 | 0.35 | 0.40 | 0.19 | 0.36 | 0.30 | 0.32 | 0.34 | 0.37 | 0.41 |
| | CDF | 0.21 | 0.38 | 0.25 | 0.11 | 0.21 | 0.14 | 0.10 | 0.30 | 0.30 | 0.41 | 0.24 | 0.20 | 0.24 | 0.29 | 0.21 | 0.22 |
| | E | 0.18 | 0.05 | 0.10 | 0.05 | 0.20 | 0.19 | 0.07 | 0.11 | 0.07 | 0.17 | 0.08 | 0.12 | 0.09 | 0.16 | 0.12 | 0.07 |
| | S | 0.04 | 0.04 | 0.05 | 0.09 | 0.05 | 0.07 | 0.03 | 0.04 | 0.04 | 0.02 | 0.08 | 0.04 | 0.05 | 0.05 | 0.04 | 0.07 |
| | CD | 0.19 | 0.22 | 0.18 | 0.16 | 0.22 | 0.17 | 0.18 | 0.18 | 0.20 | 0.18 | 0.35 | 0.26 | 0.25 | 0.15 | 0.15 | 0.22 |
| | MS | 0.14 | 0.25 | 0.21 | 0.24 | 0.28 | 0.31 | 0.24 | 0.30 | 0.23 | 0.24 | 0.29 | 0.31 | 0.39 | 0.32 | 0.21 | 0.33 |

(b) Averaged recall values.



| Vocabulary | | 512 | | | | 1024 | | | | 2048 | | | | 4096 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Density | | 8 | 12 | 16 | 24 | 8 | 12 | 16 | 24 | 8 | 12 | 16 | 24 | 8 | 12 | 16 | 24 |
| Instrument ID | LG | 0.79 | 0.78 | 0.79 | 0.75 | 0.78 | 0.75 | 0.75 | 0.76 | 0.78 | 0.78 | 0.79 | 0.74 | 0.75 | 0.80 | 0.77 | 0.76 |
| | CDF | 0.84 | 0.87 | 0.85 | 0.77 | 0.78 | 0.81 | 0.78 | 0.84 | 0.83 | 0.75 | 0.81 | 0.85 | 0.76 | 0.78 | 0.75 | 0.78 |
| | E | 0.76 | 0.87 | 0.87 | 0.93 | 0.82 | 0.82 | 0.82 | 0.86 | 0.86 | 0.83 | 0.86 | 0.89 | 0.90 | 0.85 | 0.90 | 0.91 |
| | S | 0.91 | 0.88 | 0.83 | 0.89 | 0.85 | 0.88 | 0.87 | 0.91 | 0.91 | 0.95 | 0.89 | 0.94 | 0.93 | 0.93 | 0.91 | 0.94 |
| | CD | 0.84 | 0.81 | 0.86 | 0.80 | 0.91 | 0.87 | 0.87 | 0.82 | 0.86 | 0.85 | 0.80 | 0.80 | 0.90 | 0.84 | 0.81 | 0.87 |
| | MS | 0.83 | 0.80 | 0.85 | 0.79 | 0.85 | 0.78 | 0.78 | 0.84 | 0.87 | 0.79 | 0.70 | 0.80 | 0.76 | 0.78 | 0.75 | 0.78 |

(c) Averaged accuracy values.

TABLE I: The tables shows the precision, recall and accuracy values of the 16 classifiers. Each table is grouped into the four vocabulary sizes (512, 1024, 2048 and 4096 entries) and in the four different grid sizes (8, 12, 16, 24 pixel distance of dense sampling). The rows represent different instruments. The best three values per instrument are with green, the three poorest performing classifiers per instrument are with red background.

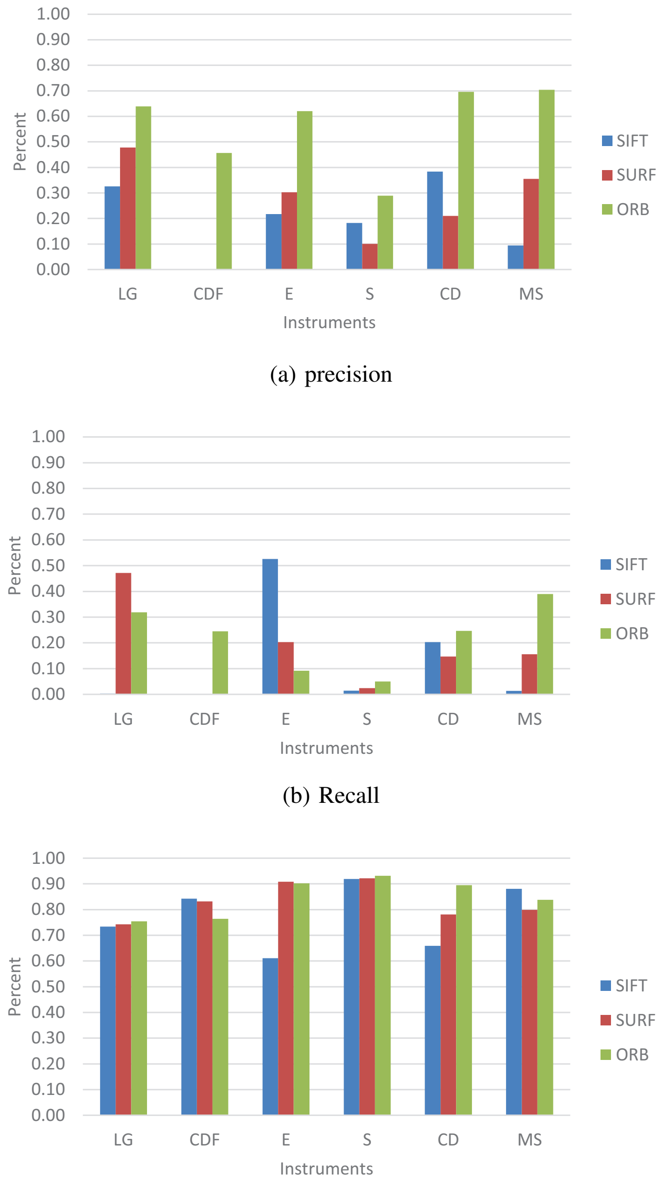
### B. *Performance of SIFT and SURF*

An important goal in our work was to find the best performing configuration of vocabulary size and keypoint sampling density. A vocabulary size of 4096 entries and a dense sampling rate with a distance of eight pixels shows the best performance in terms of precision, recall and accuracy. Hence, we use the same configuration for evaluation with other keypoint descriptors. The results for SIFT and SURF are shown in Figure 5a, 5b and 5c. As apparent in the figures, ORB features show the best precision for all instruments. Using this configuration precision is about 70% for CD and MS, more than 60% for LG and E. CDFs precision is a bit less than 50%. Instrument S is beyond 30%. The best result with SIFT is achieved for the CD with 38%. The best result in combination with SURF is obtained with 48% classifying the LG.

Figure 5a does not show any precision value for CDF with SIFT and SURF. The reason can be found in Figure 5b. The SIFT classifiers have problems assigning instruments with a similar shaft correctly. Less than 1% of the LG instrument is found correctly and none of CDF. But also the recall values for instrument S and MS are very low. These four instruments have been labeled as instrument E very often. These misclassifications lead to a low precision value of instrument E, although E show the best recall values with 53% over all classifiers. CD has another kind of shaft than the other instruments. Therefore, it shows better recall results.

A similar result is shown for SURF classifying CDF. SURF does not classify any appearance of instrument CDF correctly. Therefore, the precision values and the recall values are zero. In this case the classifier based on SURF tend to classify CDF as LG but also as E. On an overall basis SURF shows better recall values than SIFT but worse than ORB.

The accuracy values of Figure 5c show that SIFT classifiers for E and CD produces many false negative results. SIFT shows a lower average accuracy value than SURF or ORB. In terms of accuracy SURF and ORB are comparable. But taking the precision and recall values into account, ORB outperforms SIFT and SURF clearly.



(a) precision

(b) Recall

(c) Accuracy

Fig. 5: Performance comparison between SIFT, SURF and ORB.

## V. Conclusions

In this work we have evaluated SVM-based instrument classification with BoW, from densely sampled ORB keypoints, for the domain of laparoscopic videos. The evaluation results show that the classifiers for all six instruments achieve pretty high accuracy and work reliably in this very special kind of video content. The comparison with the classifiers using SIFT and SURF features show that ORB is the better choice for classification tasks in laparoscopic videos. In future work we will use the instrument classifiers as a basis for a semantic video segmentation approach of recorded laparoscopy videos.

## Acknowledgment

This work was supported by Lakeside Labs GmbH, Klagenfurt, Austria and funding from the European Regional Development Fund and the Carinthian Economic Promotion Fund (KWF) under grant KWF - 20214 22573 33955.

## References

[1] M. Lux, O. Marques, K. Schöffmann, L. Böszörmenyi, and G. Lajtai, "A novel tool for summarization of arthroscopic videos," *Multimedia Tools and Applications*, vol. 46, no. 2-3, pp. 521–544, 2010.

[2] B. Münzer, K. Schoeffmann, and L. Böszörmenyi, "Relevance segmentation of laparoscopic videos," in *Proceedings of 2013 IEEE International Symposium on Multimedia (ISM 2013)*, Anaheim, California, USA, December 2013, pp. 84–91.

[3] A. F. Smeaton, P. Over, and A. R. Doherty, "Video shot boundary detection: Seven years of trecvid activity," *Computer Vision and Image Understanding*, vol. 114, no. 4, pp. 411–418, 2010.

[4] M. J. Primus, K. Schoeffmann, and L. Böszörmenyi, "Segmentation of recorded endoscopic videos by detecting significant motion changes," in *Proceedings of the International Workshop on Content-Based Multimedia Indexing (CBMI 2013)*, 2013, pp. 223–228.

[5] B. Münzer, K. Schoeffmann, L. Böszörmenyi, J. Smulders, and J. J. Jakimowicz, "Investigation of the impact of compression on the perceptional quality of laparoscopic videos," in *Proc. of the 27th IEEE Int. Symp. on Computer-Based Medical Systems*, 2014, p. 6.

[6] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "Orb: An efficient alternative to sift or surf," in *Computer Vision (ICCV), 2011 IEEE International Conference on*, 2011, pp. 2564–2571.

[7] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *Int. J. Comput. Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[8] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool, "Surf: Speeded up robust features," *Computer Vision and Image Understanding (CVIU)*, vol. 110, pp. 346–359, 2008.

[9] M. Allan, S. Ourselin, S. Thompson, D. J. Hawkes, J. Kelly, and D. Stoyanov, "Toward detection and localization of instruments in minimally invasive surgery," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 4, pp. 1050–1058, Apr. 2013.

[10] D. Surangsrirat, M. A. Tapia, and W. Zhao, "Classification of endoscopie images using support vector machines," in *IEEE SoutheastCon 2010, Proc. of the*. IEEE, 2010, pp. 436–439.

[11] 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*, Feb. 2009, pp. 72 610A–72 610A–8.

[12] C. R. Marcano-Gamero, "Identification of surgical instruments contained in laparoscopic images," *International Journal of Robotics and Automation (IJRA)*, vol. 1, no. 3, pp. 57–65, Dezember 2010.

[13] Y.-G. Jiang, C.-W. Ngo, and J. Yang, "Towards optimal bag-of-features for object categorization and semantic video retrieval," in *Proceedings of the 6th ACM international conference on Image and video retrieval*. ACM, 2007, pp. 49–501.

[14] R. Vieux, J. Benois-Pineau, and J.-P. Domenger, "Content based image retrieval using bag-of-regions," in *Advances in Multimedia Modeling*, ser. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2012, vol. 7131, pp. 507–517.

[15] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *Workshop on statistical learning in computer vision, ECCV*, vol. 1, 2004, pp. 1–22.

[16] D. Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding," in *Proc. of the 18th ACM-SIAM Symp. on Discrete algorithms*. Society for Industrial and Applied Mathematics, 2007, pp. 102–1035.

[17] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 27–297, 1995.