Instruments Localisation and Identification for Laparoscopic Surgeries

Antoine Letouzey[1], Marion Decrouez[2], Anthony Agustinos[1], and Sandrine Voros[1,3]

[1] Univ. Grenoble Alpes, TIMC-IMAG, CNRS, F-38000 Grenoble, France
[2] SurgiQual Institute, Grenoble, France
[3] INSERM, TIMC-IMAG, F-38000 Grenoble, France

Abstract. In this paper we present a method for combined detection and identification of surgical instruments during laparoscopic surgeries. We use our previous work on the automatic detection of laparoscopic instruments to automatically construct thumbnails describing all the instruments visible in the laparoscopic images. These thumbnails are then used to train a SVM classifier for tool identification. Our results show that we are able to localise and identify various surgical instruments in previously unseen video sequences.

Keywords: Laparoscopy, surgical instruments localisation and identification

1 Introduction

In this paper we present a novel method that achieves both precise localization and identification of instruments for laparoscopic surgeries. Such minimally invasive surgeries are more beneficial to the patients but significantly increases the complexity of the surgical gestures. The constraints for surgeons are mostly ergonomic with the manipulation of surgical instruments and the visualization of the surgical scene. Automatic localization and identification of instruments can be helpful to respond to several limitations of laparoscopy and to assist surgeons. During a medical intervention it could provide the surgeon with important informations relative to the instrument's position and motion and help prevent unsafe situations or even injuries. It could also be used after the surgery, to automaticaly generate reports on the procedure, allowing for a better understanding of the interventions and eventually saving a lot of time for the medical staff. Being able to infer reliable informations using only the video stream from the endoscope is important because it does not interfer in any way with the execution of the surgeries. No additional device and no additional action from the staff is required.

Early efforts focusing on instruments detection made sometimes use of passive[12] or even active[11] markers on the tools. But these are not convenient for use in real surgeries and are mostly designed for training and skills evaluation on test-bench. Recent approaches, such as [9] based on color segmentation

or [8] based on a supervised classification method, do not requier to alter the instruments and are more suited for real procedures. Beyond localisation, tool description is the next important step in video understanding. In [10], additional informations are given on the tools, such as operational state (open/closed), blood-stained state.

In this work we focus on localisation and identification. We decided to divide the problem of surgical instruments labeling in video sequences in two parts. Firstly we try to locate them in the images, and secondly we assign a label the detected tools. This allows us to aim for much smaller computation time in the classification step since we are not working on full images anymore but only on small patches that represent the instruments, which is important for real time applications during surgeries. We first use a variation of the algorithm presented in [4] for tool detection and localisation. This algorithm is both fast and accurate. It relies on a succession of rather simple image processing techniques. We then train a linear Support Vector Machine classifier to distinguish between the seven different instrument present in the videos.

After presenting the method in details, results are exposed and prediction accuracy is discussed.

2 Precise Instruments Detection

This work is an extension of the algorithm presented in [4]. Detection and localisation of each instrument is done through a combination of rather simple image processing procedures. Figure 1 shows the complete detection pipeline.

First the input image is converted from the RGB color space to the CIELab color space. This allows to make our approach robust to illumination changes, which can affect dramatically the global appearance of the images. We then convert the two remaning channels a and b into a grayscale image corresponding to the chromaticity $C_{ab} = \sqrt{a^2 + b^2}$. This new image is then binarized with a threshold computed automatically using Otsu's method [2]. Otsu's method relies on the assumption that there are two classes of pixels in the image, foreground and background pixels. This assumption holds in the case of laparoscopic surgeries where the instruments often have a color very distinct from the background. Small artefacts in the binary image are removed through an erosion step using a cross shaped kernel. We then detect edges using a regular Canny filter. Some of these detected edges represent the instruments borders. Since all instruments roughly share the same "long and thin" shape, we use a Hough transform [3] to identify lines among these edges. We first assume that every pair of lines that are closeby indicates the presence of an instrument, then these assumptions are pruned by checking in the binary image that they actually indicate a region labeled as foreground. By computing the main axis for each of the instrument we can infer the position of its tip and border.

This method allows for fast and reliable instrument localisation. But it is not always perfect and sometimes fat tissues or other anatomical structures are wrongly identified, see 2.

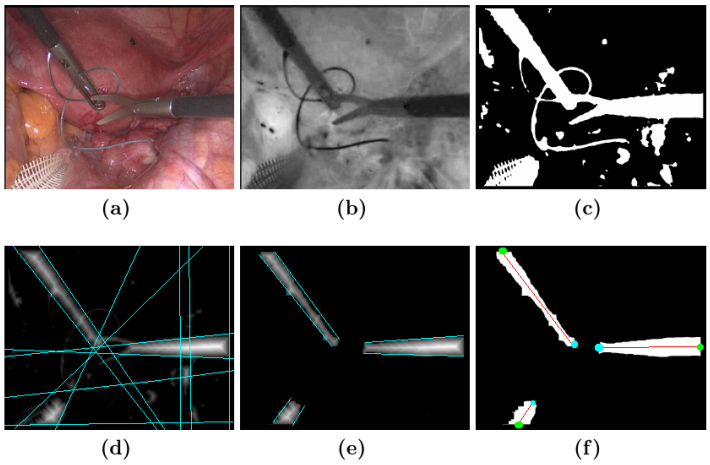


Fig. 1. Instrument detection. (a) Original image, (b) Grayscale image after conversion to CIELab colorspace, (c) Binary image with Otsu's method, (d) Hough lines detection, (e) Pruning and (f) Final image mask with main axis (red), tip (blue) and border (green) for each instrument.

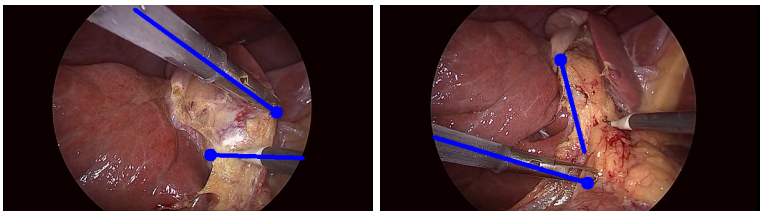


Fig. 2. (a) Two good detections and (b) one good and one erroneous detections.

3 SVM Classifier Training

In this section we describe how the input ground truth data is processed to train our classifier and then how we use it for instruments identification. It is worth noting that these two steps are fully automated. No manual input is need neither for training nor identification.

3.1 Building Training Input Data

The precise tool detection step gives us various information on every tool visible in each frame, such as the 2D position of the tip and the entry point in the image. With this information at hand we can extract form every frame the bounding box of each visible tool. In order to efficiently train our classifier we need to make sure the input labeled images are consistent with each other. For that purpose we ensure that all the extracted images have the same orientation, size and illumination.

The orientation and size are given by the 2D vector between the entry point and the tip of a detected tool. From this information we can easily rotate the image so that the tool's main axis becomes horizontal and the tip lies on the right side of the image. Having rotated the image we can crop a rectangular bounding box around the tool. Obviously, detected instruments might have a different size in length. To overcome this issue, we just rescale the images so that all our training images will have the same size. We chose a size of 256×30 pixels, these values where selected because they match the average size (lenght and radius) of the tools seen in the video sequences. Finally we perform a simple histogram equalization to minimize the inpact of ilumination variation on the images. Figure 3 illustrate these steps.

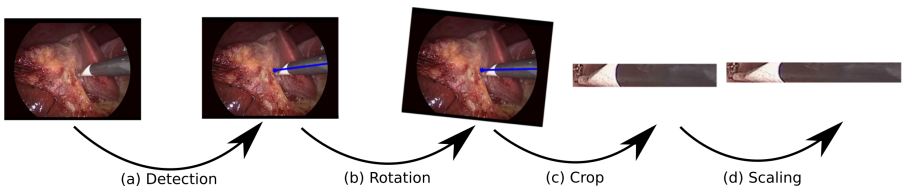


Fig. 3. Tool thumbnail extraction process. (a) Position and orientation detection, (b) Alignment with the tool, (c) Bounding box cropping and (d) Final scaling.

Using this fully automated procedure we can extract tools from the background and hence have access to the meaningful portion of each frame. But our detection algorithm only performs tool localisation and not identification. To train our classifier we need to label all these detected instruments. From the ground truth data provided within the challenge we know which tool are present on every frame. We use this information and select the frames where only one

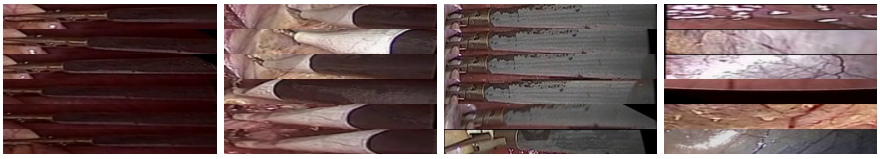
tool is visible. This allows us to be certain which tool was detected using our algorithm and prevents mis-labeling.

We thus automatically create reliable sets of reference images for each instrument based on the input ground truth data. Before using these images to train our classifier we create an artificial garbage class (called "not_a_tool") to help us handle misdetection. We build an additional set of images by taking samples at random position and orientation in frames in which no instrument appears. This set of images can be seen as a model of the background.

Figure 4 shows a subset of images built with our method and used to train our classifier.

3.2 Feature Vector and Classifier Training

We chose to use a linear SVM classifier[5] for its simplicity and robustness. In the training phase it takes as input a feature vector associated with a label for each element of the training dataset. We decided to use HOG descriptor[1] to convert our input images into feature vectors. This descriptor has initially been designed to detect pedestrian in images. It is well designed for describing full image patches, fast and has been widely used by the community for image classification. It caracterize appearance and shape within an image by computing distribution of gradient directions in uniformly divided image cell. Hence it should be robust to color and illumination variation which can be problematic for laparoscopic videos. Other descriptors, such as SIFT[6] or SURF[7], could have been used but they are designed mostly to describe interest points and their neighbourhood but not full images. Finally, we train our classifier by feeding it these feature vectors along with their associated labels.



(a) Clipper (b) Hook (c) Grasper (d) not_a_tool

Fig. 4. Random samples taken from 3 instrument classes along with a few example of our garbage class.

3.3 Tool Identification

Having trained our classifier we can now use it to predict the label for any given image. We just need to make sure these images go through the same process as those from our training dataset (i.e. rotation, scaling and illumination equalization). When asked to predict the label of an input image, the SVM

classifier returns a confidence vector. Each of its values reflect the likelihood of the input image to belong to each of the class. A simple solution is then to look for the maximum value and use it to assign a label to the image. But we chose not to label at all images whose maximum response is below a threshold, treating them as part of the garbage class instead of assigning them a label with low confidence.

4 Results

Table 1. Prediction accuracy

	grasper	bipolar	hook	scissors	clipper	irrigator	specimenbag	not_a_tool	**Total**
Train/Test size	155	14	188	7	20	22	7	210	**623**
Accuracy	68.39%	26.67%	89.30%	12.5%	15.0%	36.36%	33.33%	73.93%	**71.63%**

To evaluate the quality of our SVM classifier, we divided the labeled data in two groups of same size. One group is used to train the classifier and the other one as a set of test images. The same evaluation procedure has been conducted multiple times, with the two groups being selected randomly each time. Table 1 shows the number of images to train and test each classe along with the correct identification ratio. As we can see the number of images available to train each class can vary dramatically, over 150 for "hook" and "grasper" but only 7 for "scissors" or "specimenbag". This is directly related to the number of appearance of each instrument in the input ground truth data. Scissors do appear more than 7 times in the video but as our algorithm is fully automatic we did not want to add manually labeled images. Specimen bags also appear more often but their shape breakes our "long and thin" assumption, hence they are less likely to be correctly localised.

We obtained an average score of 71% for correct labeling. As we can see on figures 5 and 6, our classifier has trouble identifing some classes, such as "scissors" or "irrigator". On the other hand "hook", "grasper", and "not_a_tool" yeld great results. This is in direct correlation with the number of images available for each classe during the training phase. With almost 90% accuracy "hook" is by far the more reliable class. This instrument is very distinct from the others because of its white tip (see figure 4(b)), yealding a strong image gradient in the thumbnail compared to other instruments. This kind of feature is exactly what our classifer uses to distinguish tools. The confusion matrix shows that the specimen bag is mostly labelled as "not_a_tool". Its color and shape makes it sometimes hard to differentiate it from anatomical structures.

5 Discussion

During our evaluations we found that most of the error came from the detection process. Most of our errors are false negative. With a average accuracy of 71%,

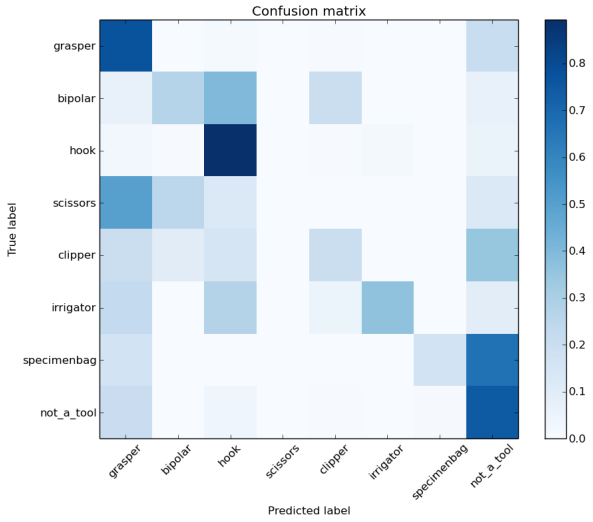


Fig. 5. Confusion matrix for our SVM classifier.

the classification step yelds relatively few errors (false positives). But it cannot identify a tool that has not been detected. The method we inspired ours from is designed to detect multiple tools in video sequences and rely on accurate frame to frame tracking. Whereas, in our implementation we dissabled the tracking component, mostly to speed up computation. This led to a less stable algorithm and less precise detections. Our experimentations also highlighted a limitation of our detection algorithm related to instruments which break the "long and thin" shape assumption we made early in its development.

Our SVM classifier offers good results. But with 71% accuracy there is room for improvement. A first direction for improvment would be to replace the fully automated learning phase with a semi-automated one. This would allow us to gather much more labeled input data and enhance the accuracy for classes for which very few reference images are currently available. We use a rather simple and generic linear approach for the identification step. We could try to improve these results by using a more efficient classification process. In that aspect there has been recently a great enthusiasm for methods based on Convolutional neural networks (CNN) for image labeling.

References

1. Dalal, N., Triggs, B.: Histograms of Oriented Gradients for Human Detection. International Conference on Computer Vision & Pattern Recognition (CVPR 2005)
2. Otsu, N.: A threshold selection method from gray-level histograms. Automatica, (1975)

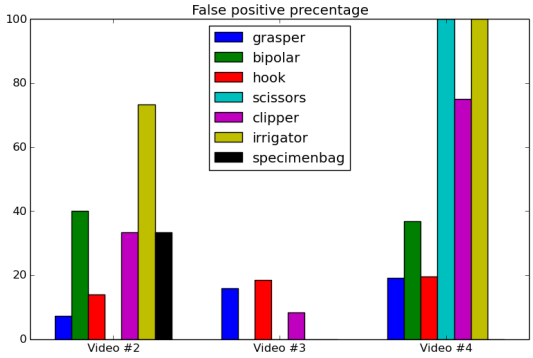


Fig. 6. Ratio of false positive during the labeling process (in %).

3. Duda, R. O., Hart, P. E.: Use of the Hough Transformation to Detect Lines and Curves in Pictures. Comm. ACM, Vol. 15, pp. 1115 (1972)
4. Agustinos, A., Voros, S.: 2D/3D Real-Time Tracking of Surgical Instruments Based on Endoscopic Image Processing. Computer-Assisted and Robotic Endoscopy: Second International Workshop, (CARE 2015)
5. Chih-Chung, C., Chih-Jen, L.: LIBSVM: A Library for Support Vector Machines. National Taiwan University, Taipei, Taiwan (2001)
6. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International journal of computer vision 60 (2), 91-110, (2004)
7. Baya, H., Essa, A., Tuytelaarsb, T. Van Gool, L.: Speeded-Up Robust Features (SURF). Computer Vision and Image Understanding, Volume 110, Issue 3, (2008)
8. Allan, M., Ourselin, S., Thompson, S., Hawkes, D.J., Kelly, J., Stoyanov, D.: Toward Detection and Localization of Instruments in Minimally Invasive Surgery. IEEE Transactions on Biomedical Engineering Volume: 60, Issue: 4, (2013)
9. Doignon, C., Graebling, P., de Mathelin, M.: Real-time segmentation of surgical instruments inside the abdominal cavity using a joint hue saturation color feature. Real-Time Imaging 11(5-6), 429442 (2005)
10. Kumar, S., Narayanan, M.S., Singhal, P., Corso, J.J., Krovi, V.: Surgical Tool Attributes from Monocular Video. IEEE International Conference on Robotics & Automation (ICRA), (2014)
11. Krupa, A., Gangloff, J., Doignon, C.,de Mathelin, M.F., Morel, G., Leroy, J., Soler, L., Marescaux, J.: Autonomous 3-d positioning of surgical instruments in robotized laparoscopic surgery using visual servoing. IEEE Tran. Rob. and Auto. (2003)
12. Groeger, M., Arbter, K., Hirzinger, G.: Motion tracking for minimally invasive robotic surgery. Medical Robotics, I-Tech Education and Publishing, pages 117148, (2008)