Identification of Surgical Instruments Contained in Laparoscopic Images

Cosme Rafael Marcano-Gamero cosmemarcano@gmail.com
Department of Electronics
National Experimental Polytechnic University
"Antonio José de Sucre" (UNEXPO AJS).
Puerto Ordaz,:8040, Venezuela.

Abstract

In this work, an approach for surgical instruments recognition within images taken from an endoscope video camera is presented. This approach is based on the analysis of the images using the SIFT transform and a clustering method called *k-means*, jointly with the use of Support Vector Machines. The instrument identification might be used for recognizing the action the surgeon is performing during the intervention, which may be useful for monitoring or training purposes. By correlating the action which is being performed with the intervention protocol, a robotic assistant might warns about the correct order in which these actions should be done, or the time they should take, according to average measurements normally accepted by medical organizations. Other approaches, based on the analysis of the instruments trajectories and forces/torques exerted by the surgeon have been proponed by Rosen *et al*. That approach implies the need for attaching sensors to the handles used by the surgeon to manipulate the laparoscopic tools, which not always is possible or might be not admitted by the surgeon for his own comfort. The original aspect of this work is to take the images directly from the Camera Control Unit (CCU) of a laparoscopic system, which provides video signal output of the embedded camera, instead of wiring additional sensors, which have to be connected and calibrated each time a new intervention is started.

Keywords: Instrument recognition Support Vector Machines, Image analysis, SIFT transform.

1. INTRODUCTION

In the recent decades, a huge effort has been put forth in the recognition of objects contained in images. To do that, many image analysis techniques have been developed. The *corner detector* and the *Scale Invariant Features Transform* (SIFT), due to C. Harris [1] and D. Lowe [2], respectively, are two of the most popular of those techniques. The SIFT has shown remarkable properties which make it a powerful tool, even in situations where illumination, object rotation and/or partial occlusion make difficult the recognition task.

In the present work, the transform SIFT [2] is used to represent images taken from a DVD recording of a laparoscopic hysterectomy, which was captured from an endoscope video camera. The camera head is connected to a Camera Control Unit (CCU) [3], which provides video signal output through a conventional S-Video connector. This connector allows us to attach the CCU to a DVD recorder and other equipments.

The images are then captured through an application in C/C++ that uses the Intel OpenCV library [4]. Captured images are then transformed by SIFT [2]. This transform firstly detects interest points which are further processed to select some of them and compute the so-called descriptors points. After transformation, the numeric representation (in descriptors points) of each image is clustered, by using some clustering method as k-means [5], to obtain a shorter representation consistent of a number of terms equals to the number of clusters, k, used by k-means. These short representations may be analyzed in form of histograms, and are stored in a text file, one per row, to form a vocabulary, which may be understood as a collection of words that characterize the images we expect to classify for objects identification purposes. Some researchers usually refer to each text file row as a bag-of-visterms (visterms stands for visual terms) or BOV [6]. All this preparation procedure tries to resemble another interesting task as the text categorization [6][7][8], and makes it easier the image classification.

The text file is then inputted to an application which trains a Support Vector Machine (SVM) [9][10][11]. Training process generates a model which will be used by another application along with another text file which contains the BOV of the images to be classified. There is an SVM for each instrument we want to identify in the images.

Projects oriented to the study and analysis of object classification methods are being developed by many universities and institutes around the world, like the Classification of visual scenes using Affine invariant Regions and TExt Retrieval methods (the project CARTER), initiated by the Katholieke Universiteit Leuven, (KUL), Belgium, through its institute IDIAP, [12].

2. SOME COMMENTS ON SUPPORT VECTOR MACHINES

Support Vector Machines (SVM) refers to a set of functions which define *separating hyperplanes* (H1 and H2 in Figure 1), which allow the classification of items in two different categories.

The easier classification case using SVM consists in a linear classifier defined on a data set with separable elements. Nevertheless, SVM can be used in a variety of cases, including the use of non linear kernels on datasets not easily separable, which constitutes its heaviest application.

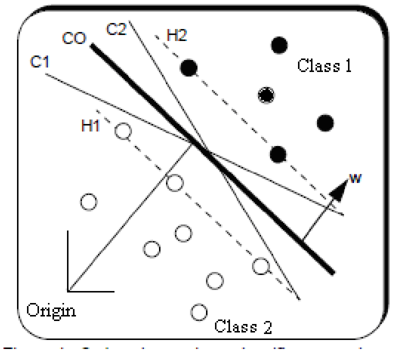

FIGURE 1: A 2-classes classifier.

In this work, we use linear 2-classes classifiers due to the separable nature of the dataset involved in experiments. For training/classification tasks, a software named SVM^{light}, [9] [10], is used. This software is based on the calculation of the $\xi\alpha$-estimators, which have shown a fast and stable behavior. Also, it utilizes the statistical strategy known as *leave-one-out*, proposed by Lunts and Brailovskiy, [13].

According to Vapnik [14], SVM are founded on the optimization principle of minimizing the so-called *structural risk*, from the statistical learning theory.

In its simpler form, the SVM learn from linear decision rules of the form $h(\vec{x}) = sign(\vec{w} \cdot \vec{x} + b)$, which are described by a weight vector $\overline{w}$ and a threshold *b*. For a given training dataset S_n, the SVM finds the soft margin hiperplane that is a maximum. The calculation of this hiperplane is equivalent to solve the following optimization problem [14], (PO1):

Minimize:

$$V(\vec{w},b,\vec{\xi}) = \frac{1}{2}\vec{w}\cdot\vec{w} + C\sum_{i=1}^{n}\xi_i \qquad (1)$$

Subject to:

$$\begin{aligned} &\forall i = 1,2,...,n : y_i[\vec{w}\cdot\vec{x}_i + b] > 1 - \xi_i \\ &\forall i = 1,2,..,n : \xi_i > 0 \end{aligned} \qquad (2)$$

The restrictions imposed by (2) require all training samples be correctly classified within a strip of amplitude equal to ξ_i. The ξ_i values constitute the elements of the *loss vector* $\vec{\xi}$. If one of the training samples falls into the wrong side of the hiperplane, the corresponding ξ_i is greater or equal to 1. Therefore, $\sum_{i=1}^{n}\xi_i$ is an upper bound to the number of training errors. The factor El factor *C* en (1) es un parámetro que permite hacer un compromiso entre el error de entrenamiento y la complejidad del modelo. Instead of solving the linear problem (primal) PO1, it is easier to solve the dual one, arose by Wolfe [14], (PO2):

Minimize:

$$W(\vec{\alpha}) = -\sum_{i=1}^{n}\alpha_i + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} y_i y_j \alpha_i \alpha_j (\vec{x}_i \cdot \vec{x}_j) \qquad (3)$$

subject to:

$$\sum_{i=1}^{n} y_i \alpha_i = 0 \wedge \forall i = 1,...,n : 0 \le \alpha_i \le C \qquad (4)$$

However, (3) and (4) can not be directly calculated, but after training the model on S_n. The estimators input are the vector $\vec{\alpha}$, which solves the training problem PO2, and the vector $\vec{\xi}$, from the primal problem solution PO1. Due to this dependence, these estimators are referred to as estimators-$\xi\alpha$. Both $\vec{\alpha}$ and $\vec{\xi}$ are available without no further effort immediately after training the SVM.

The application of Quadratic Programming (QP) to the SVM optimization reduces to minimizing the following expression (5) and (6):

$$W(\vec{\alpha}) = -\sum_{k=1}^{n}\alpha_k - \frac{1}{2}\sum_{k=1}^{n}\sum_{l=1}^{n}\alpha_k \alpha_l Q_{kl} \qquad (5)$$

where:

$$Q_{kl} = y_k y_l \Phi(x_k)\Phi(x_l) \qquad (6)$$

Trying to solve the primal problem QP1 in a straight way, we would have to do $R^2/2$ dot product calculations to prepare the matrix Q. On the other hand, each dot product requires to make $m^2/2$ addition and multiplication operations; therefore, the total computational cost would be $m^2R^2/4$, which represents a huge amount of operations. However, it can be noted that:

$$\Phi(a) \cdot \Phi(b) = 1 + 2\sum_{i=1}^{n} a_i b_i + \sum_{i=1}^{n} a_i^2 b_i^2 + \ldots$$
$$\ldots + \sum_{i=1}^{n} \sum_{j=1}^{m} 2 a_i b_i a_j b_j) = (a \cdot b + 1)^2 \qquad (7)$$

It can be observe that the calculation of (7) requires only $nR^2 / 2$. In the bi-dimensional case, like to the 2-class classifier, the SVM hyperplanes [1] are defined as follows:

$$\begin{aligned} H1 &= \{\vec{x} : \vec{x} \cdot \vec{w} + b = +1\} \\ H2 &= \{\vec{x} : \vec{x} \cdot \vec{w} + b = -1\} \end{aligned} \qquad (8)$$

A new sample is classified according to (9):

$$\begin{aligned} y_i &= +1 \quad si \quad \vec{x}_i \cdot \vec{w} + b \geq +1 \\ y_i &= -1 \quad si \quad \vec{x}_i \cdot \vec{w} + b \leq -1 \end{aligned} \qquad (9)$$

which can be rewritten as a unique decision rule as shown in (10):

$$y_i(\vec{x}_i \cdot \vec{w} + b) - 1 \geq 0 \quad \forall i \qquad (10)$$

where $\vec{x}_i$ denotes the *i-th* vector; each vector represents a training sample in the set; y_i denotes a respective label assigned to each of these training samples. In this case, $y_i \in \{-1,+1\}$ due to only two classes are being considered. Finally, $\vec{w}$ denotes the weight vector.

On the other hand, the *leave-one-out* strategy consists in the following. If a sample $(\vec{x}_i, y_i)$ is incorrectly classified by an SVM which was trained on the subset $S_n^{\setminus i}$, (i.e., the set S_n without the *i-th* sample), then the sample $(\vec{x}_i, y_i)$ is completely removed from the training dataset S_n. Furthermore, this sample should satisfy $(\rho\alpha_i R_\Delta^2 + \xi_i) \geq 1$, if we are trying to classify using an SVM trained on the complete dataset S_n. Samples which were leaved out from the dataset are said that caused a *leave-one-out error*.

The total number of *leave-one-out* errors divided by the total number of training samples, *n*, constitutes the estimate of the *leave-one-out* error. On the other hand, let us denote *d* the total number of training samples for which $(\rho\alpha_i R_\Delta^2 + \xi_i) \geq 1$ holds. From the previous analysis it is easily derived that *d* is an upper bound for the test of *leave-one-out*.

This kind of classifiers belongs to the category of *supervised machine learning* because the label assignation to the training samples must be done by the SVM designer.

3. EXPERIMENT: IDENTIFYING AN INSTRUMENT

As an example of application of this approach, we take the task of identifying a scalpel for laparoscopic procedures, like the HARMONIC Ultracision Scalpel 300, [15], by *Ethicon Endo-Surgery, Inc* (henceforth referred to as scalpel), contained in a sequence of frames which were taken from the endoscope video camera, like those shown in Figure 2. These images belong to a radical hysterectomy and pelvic lymphadenectomy, which have become a standard procedure for invasive cervical cancer, [16] [17].

Figure 2 shows some images which contains only tissue, and others containing a scalpel in different poses, with variations in orientation and illumination. SIFT transform has shown a very high level of tolerance to these variations, and also outstandingly tolerates partial occlusion of the objects to be identified, in comparison with other kind of mathematical transforms such as the Rotation Invariant Features Transform, commonly referred to as RIFT.

As mentioned before, the SIFT transform generates two matrices: frames and descriptors. The first is 4xP and the other 128xP, where P is the number of interest points selected during the first phase of the transformation process, referred to as point detector phase, [2]. The number of points P, may vary from one image to other. The number of features, 128, is derived from the number of gradients which are calculated at each descriptor point in eight different directions with respect to its nearest neighbors.

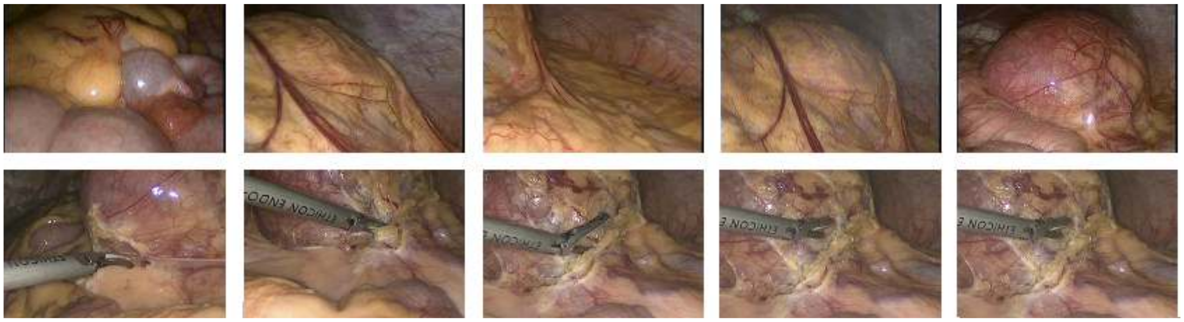

FIGURE 2: Some frames taken from the endoscope video camera.

Interest points are filtered by removing those points which show very similar gradients, in order to improve the SVM discrimination capacity at recognition time.

Taking as an example one frame similar to those shown in the Figure 2, we proceeded to apply the SIFT transform. In this case, the SIFT generated the matrix *frames*, of *4x980* and the matrix *descriptors*, of *128x980* elements, respectively. Figure 3 shows the blob-like areas surrounding the descriptors points.

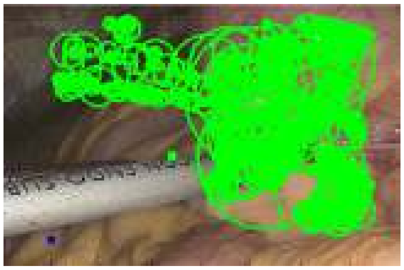

FIGURE 3: Descriptors points of a sample.

In order to reduce the size of the dataset file, and makes it faster the recognition process, we can group each descriptor matrix by using a clustering method such as *k-means*, so the *128x980* features can be reduced to *k* representing *centroids* (the centers of each cluster). In relation to the choice of the clusters count, *k*, to be indicated to the k-means, we made some trial and error tests to get a convenient value for all the images. It is necessary due to the fact that k-means removes those clusters with no assigned points (which are referred to as a *dead class*). It yields BOV of different number of terms, which is inconvenient for the SVM training and classification process.

In our case, essays results gave *k=15* as a good value for all images. So, each BOV resulted in a sequence of 15 terms for each processed image, which eventually showed the following shape:

```
+1 1:27 2:22 3:27 4:21 5:39 6:26 7:39 8:21 9:43 10:39 11:23 12:16 13:19 14:30 15:17
```

Notice that the label *+1* indicates that this BOV corresponds to an image which contains the interest instrument, i.e., the scalpel. This label should be *-1* if the image does not contain it.

All the BOVs must have the form of *clusternumber:members-count*, as required by SVM^{light}. Figure 4 shows the representation of this particular BOV as a histogram.

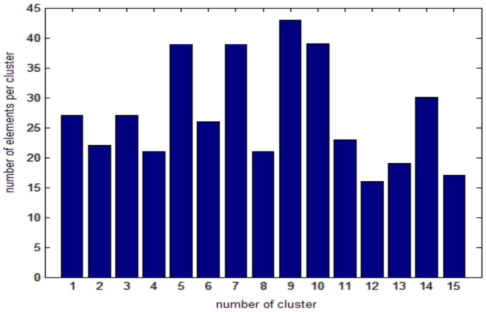

FIGURE 4: Histogram corresponding to a sample.

These BOV are stored in a text file, one BOV per row; each one represents a different image. This file constitutes a *vocabulary*, and will be inputted to the SVM training module of the software SVM^{light}.The vocabulary extension depends on the clusters count *k*, and the number of training samples we used. Some researchers store some additional information in the BOV data file, as the frequency of appearance of each descriptor in each processed image [18]

For this experiment, a training dataset of 2261-samples were processed, half of them corresponding to images which contained the scalpel and the other did not.

By using the software SVM^{light} , [9] [10], an SVM was trained and some data text files, containing non labeled BOV, were tested. The training procedure yields an error estimate equal or less than 9.43% (see Table 1), and a precision equal or greater than 91.74%. One of the test file contained 568 non labeled samples and only 11 of them were wrongly classified by the classification module of the used software. The error was of 1.94%, which is within the calculated error estimation. Other essays shown that the error and, correspondingly, the precision estimates might be improved by reducing the number of misclassified samples at training time. Also, increasing the number of training samples, error, recall and precision indices can be improved. The misclassified samples are those which were classified as scalpel containing only tissue, and vice versa.

Of course, the SVM performance is immune to draw samples from the training dataset which does not constitute support vectors. In fact, notice that of the 2261 training samples used to learn the current SVM, only 268 of them were constituted as support vectors (see Table 1).

To illustrate the results given by the SVM^{light} software, Table 1 shows the values reported after training a SVM using a dataset of 2261 samples. These results were obtained on a desktop computer based on an *Intel Pentium IV*, @3.066 GHz, and 2 GB of RAM. The computation time (without including I/O operations) for this case was of 0.15 seconds, including the $\xi\alpha$-estimates. It is highly important to notice that the training phase is done off-line, and the classification phase, which should be done in real time, took even less computation time.

Samples count	Total of misclassified samples	Total of Support Vectors	Total SV at upper bound	Number of kernel evaluations	VCdim	$\xi\alpha$-estimates		
						Error %	Recall %	Precision %
2261	120	281	268	35883	≤ 76.305	≤ 9.43	≥ 91.68	≥ 91.74

TABLE 1: Typical Values Calculated By SVM^{light} , while processing a dataset of 2261 samples.

It is also important to highlight that the values for the $\xi\alpha$-estimates, shown in Table 1, allow us to assure the quality or goodness of the model which was obtained. On the other hand, a relevant index of performance of the SVM is the so-called Vapnik-Chernovenkis dimension or *VCdim*, [11][14] [19], which is a measurement of the discrimination capacity of the model.

The VCdim may be though about, in a very loosely way, as a SVM performance index having an upper bound in the quotient of the diameter of the smallest sphere that can enclose all the high-dimensional term-vectors derived from the training set, and the smallest margin we will let the SVM use. This can be used in Structural Risk Minimization (SRM) for choosing the polynomial degree, RBF standard deviation σ , or another parameter according to the function we are using as kernel for training the SVM, although many researchers prefer to use Cross-Validation [11][14].

The training process should be repeated for each instrument which may be used during the intervention. Considering that each SVM works as a 2-class classifier, we used a *one-against-all* strategy in order to discriminate each of the instruments in the images under analysis.

After we have all the needed SVM installed on a computer connected to the Camera Control Unit of an IMAGE1 HUB HD Endoscope System, images can be captured in real time and the software of recognition would allow us to identify the different instruments involved in the intervention through its different stages.

Taking into account that the video stream taken from the CCU is conformed by many images (25 frames per second), the software may take various consecutive frames and try to recognize the instruments contained in each one, and select as a “truly valid” identification the instruments that have been recognized a greater number of times in these frames, in order to get more confidence level in the identification process. Of course, in doing so, there would be computation time considerations to be taken into account.

4. CONCLUSION

As we can see in this work, the use of both the Scales Invariant Features Transform (SIFT) [2], and the Support Vectors Machines (SVM), [9] [10] [11], is a very convenient and feasible way to face the surgical instruments identification task in images directly taken from an endoscope video camera, used in laparoscopic surgery, and represents a good alternative to other approaches based on wiring many additional sensors, as proposed in [20]. The approach presented here requires no complicated mounting, adjustment or calibration procedures, and is less susceptible to failure and errors.

Future work foresees the use of Hidden Markov Models (HMM) in order to correlate the object recognition with the intervention protocols, for identifying the action the surgeon is doing in a given instant during the intervention.

5.ACKNOWLEDGMENTS

The author wants to thank Oscar Hurtado-Patiño, M.D., Specialist in Oncology, for facilitating the endoscope video recordings used to prepare the datasets involved in this work. Also, he wants to thank Denisis for her always lovely and helpful support.

6. REFERENCES

1. C. Harris and M.J. Stephens. *“A combined corner and edge detector”*. In Alvey Vision Conference, 147–152, 1988.

2. D. Lowe. *“Distinctive image features from scale-invariant keypoints”*. *MIT Press*, 14(1), 2002.

3. KARL STORZ – *“Endoscope. IMAGE1 HUB HD Camera Syste: Instruction Manual”*. KARL STORZ GmbH & Co. KG Postfach 230. 78503. Tuttlingen, Germany.

4. Intel Corporation. *OpenCV Documentation* May. 2009, USA.

5. D. Fisher. *“On groping for maximum homogeneity*”. Journal of American Statististical Association, 53:789–798, 1958.

6. P. Quelhas, F. Monay, J. Odobez, Daniel Gatica-Perez and Tinne Tuytelaars. *“A Thousand Words in a Scene”.* Pattern Analysis and Machine Intelligence, IEEE Transactions on. 29(9). 1575 – 1589. September 2007.

7. A. Basu, C. Watters, and M. Shepherd. “S*upport Vector Machines for Text Categorization”.* Proceedings of the 36th Hawaii International Conference on System Sciences (HICSS’03) 2002.

8. J. Sivic and A. Zisserman. *“Video google: A text retrieval approach to object matching in videos”.* In Proc. International Conference on Computer Vision (ICCV), 2003.

9. J. Thorsten. *“Advances in kernel methods: support vector learning. ch. 11: Making large-scale support vector machine learning practical”*. MIT Press Cambridge, MA, USA, 69–184, 1999.

10. J. Thorsten. *“Estimating the Generalization Performance of an SVM Efficiently*”. International Conference on Machine Learning. Published by Morgan Kaufman. San Francisco, 431--438, 2000.

11. A.W. Moore (2001) *“Support Vector Machines”* Tutorial. School of Computer Science of the Carnegie Mellon University. Available at http://www.cs.cmu.edu/~awm/tutorials . [Accessed August 16, 2009].

12. D. Gatica- Pérez, T. Tuytelaars, P. Quelhas, F. Monay, J.M. Odobez (2005). *”CARTER: Classification of visual scenes using Affine invariant Regions and TExt Retrieval methods*”. Available at: http://www.idiap.ch/carter/. [Accessed August 12, 2009].

13. A. Lunts, and V. Brailovsky. *“Evaluation of attributes obtained in statistical decision rules*”. Engineering Cybernetics, *3*, 8–109. 1967.

14. V. Vapnik, “*Statistical Learning Theory”*. Wiley- Interscience, 1998, New York.

15. A. Siperstein, E. Berber, and E. Morkoyun. *“The Use of the Harmonic Scalpel vs ConventionalKnot Tying for Vessel Ligation in Thyroid Surgery”.* American Medical Association. ©2002. (REPRINTED) ARCH SURG/Vol. 137, Feb 2002 , pp.137-142.[online] Available at http://www.archsurg.com. [Accessed November14, 2010].

16. S. F Willis, D. Barton and T. EJ Ind. "*Laparoscopic hysterectomy with or without pelvic lymphadenectomy or sampling in a high-risk series of patients with endometrial cancer*". *International Seminars in Surgical Oncology* 2006. September 2006.

17. Y. Chen, H. Xu, Y. Li, D. Wang, J. Li, J. Yuan and Z. Liang. *"The outcome of laparoscopic radical hysterectomy and lymphadenectomy for cervical cancer: a prospective analysis of 295 patients*". Ann Surg Oncol. 2008 Oct. 15(10):2847-55. Epub 2008 Jul 23.

18. J. Sivic, A. Zisserman. *"Video Data Mining Using Configurations of Viewpoint Invariant Regions".* Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on. 1(I). 488 - 495. June 2004.

19. C. Burges. *"A tutorial on support vector machines for pattern recognition*". Data Mining and Knowledge Discovery, 2(1):121–166, 1998.

20. J. Rosen, D. Brown. *"Generalized approach for modeling minimally invasive surgery as a stochastic process*". IEEE Transactions On Biomedical Engineering, 53(3), 2006.