Boosted Wireless Capsule Endoscopy Frames Classification

Giovanni Gallo, Alessandro Torrisi
Dipartimento di Matematica e Informatica
Università di Catania
Catania, Italy
Email: {gallo, atorrisi}@dmi.unict.it

***Abstract*—Intestinal lumen detection in endoscopic images is clinically relevant to assist the medical expert to study intestinal motility. Wireless Capsule Endoscopy (WCE) produces a high number of frames and automatic classification, indexation and annotation of WCE videos is crucial to a more widespread use of this diagnostic tool. In this paper we propose a novel intestinal lumen detection method based on boosting. In particular, we use a customized set of Haar-like features combined with a variant of Adaboost to select discriminative features and to combine them into a cascade of strong classifiers. Experimental results show the efficacy of boosted classifiers to quickly recognize the presence of intestinal lumen frames in a video.**

***Keywords*-Classification; Pattern Recognition; Boosting; Wireless Capsule Endoscopy; Video Automatic Annotation.**

I. Introduction

Wireless Capsule Endoscopy [1], [2] is a technique to explore small intestine regions that traditional endoscopy does not reach. A video-capsule, that integrates wireless transmission with image technology, is swallowed by the patient and it is propelled through the gut by intestinal peristalsis. Once activated, the capsule captures two frames per second and transmits images to an external receiver. The exam is concluded after about eight hours, that corresponds to the lifetime of the battery of the capsule. Images taken during the entire route of the capsule through the intestine are successively analyzed by an expert. She may spend up to one or more hours to gather the relevant information for a proper diagnosis. This greatly limits the use of the capsule as a diagnotic routine tool.

This state of the things may be greatly improved if the WCE video is automatically segmented into shorter videos, each one relative to a different trait of the bowels, and if reliable automatic annotation tools are available. The goal of automatically produce a summary of the whole WCE video is still unaccomplished. Tools to extract semantic information from such videos are hence relevant.

Within this general framework, in this paper we present a novel method to automatically discriminate a relevant subclass of frames. In particular, our classifier sorts the frames into two categories: "with lumen" (images depicting the stages of a intestinal contraction where the lumen of the bowel is well visible) and "without lumen" (Figure 1). Lumen detection is clinically relevant because it announces the presence of a contraction and helps the physician to study intestinal motility. Alteration of physiological intestinal motility is an indicator of disorders in which the gut has lost its ability because of endogenous or exogenous causes. Lumen detection in our approach is obtained as a special case of object detection, and uses to this aim the Viola and Jones paradigm introduced in 2001 [3].

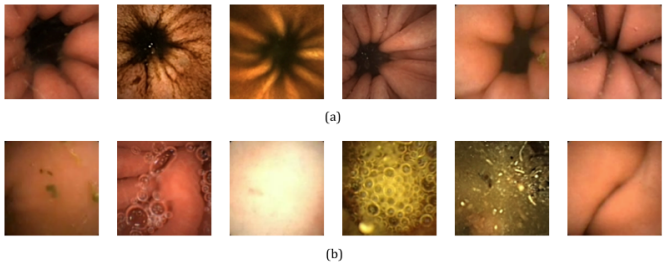

Figure 1. Examples of *lumen* (a) and *not lumen* (b) images extracted from a WCE video.

This paper is organized as follows: Section II lists related works and reports examples of object detection based on an approach similar to ours. Section III describes in detail how Viola-Jones technique is applied to the present problem. Section IV reports the experiments conducted on real WCE video. Finally, Section V draws conclusions and discusses some future works.

II. Related Works

Most of the systems reported in literature to recognize intestinal lumen images refer to classic endoscopy. The motivation behind these methods is to aid the physician to individuate lumen region to avoid the collision of the endoscope instrument with the intestinal mucosa. In this context, Asari [4] proposes a Region Growing Segmentation to extract lumen from gray level endoscopic images.

Recently, the original WCE is evolving into a novel capsule whose movement may be remotely controlled. In this context, the recognition of lumen could help the capsule to go through the intestine minimizing collisions and avoiding to record meaningless frames. To this aim, Zabulis et al. [5] propose a system based on a Mean Shift Segmentation algorithm variant to locate lumen regions in WCE frames. The problem of the detection of images with lumen in WCE

 ISBN: 978-1-61208-158-8

videos to clinical use is not much investigated. Some works study the general problem to find contractions to examine intestinal motility [6]–[8].

The main idea exploited in this work is the Viola-Jones method for object detection [3], [9]. Initially proposed for face detection, this technique is based on the use of simple features calculated in a new representation of the image. Based on the concept of integral image [10], a huge set of features is tested and the boosting algorithm Adaboost is used to reduce this set [11]–[13]. The introduction of a tree of boosted classifiers provides a robust and fast detection and minimizes the false positive rate. This strategy has been proven effective to recognize various kind of objects. Several systems have been proposed for different recognition problems, like face, hands and pedestrian [14]–[17]. The possibility to define a specific set of features and the more recent release of an open source implementation [18] have permitted to use extensively this method in Computer Vision.

III. Proposed Method

We propose a system that automatically learns and classifies frames where intestinal lumen is well visible. The process of automatic annotation of images containing intestinal lumen can be summarized in the following three steps:

- Evaluation of a customized set of Haar features applied to the integral images of the training samples.
- Selection of best discriminative features through AdaBoost algorithm.
- Construction of a final boosted classifier based on a cascade of classifiers whose complexity is gradually increasing.

To be reliable, such a system must satisfy two requirements: it needs a comprehensive set of examples where the object of interest may occur; a suitable selection of descriptors required to describe each possible pattern must be guaranteed. To this aim, Haar-like features, a set derived from Haar wavelets [19], recognize objects using intensity contrast between adjacent regions in an image.

Basic Haar features proposed by Viola-Jones for face detection do not have sufficient discriminative power for lumen investigation: it necessary to define customized variations of this kind of features. In particular, the features proposed in this work should provide a strong positive response on a rectangular region with low intensity called generically "lumen" and a brighter surrounding area in which appears the gut wall. By combining a learned evaluation threshold to each feature, it is possible to assign an image to the appropriate category. Figure 2 shows an example of this first kind of proposed features that we call "center-surround" features.

The typical appearance of a frame that shows an intestinal contractions consists in lumen surrounded by typical rays that muscular tone produces due to the folding of the intestinal wall. We hence define two additional "cross-like" kind

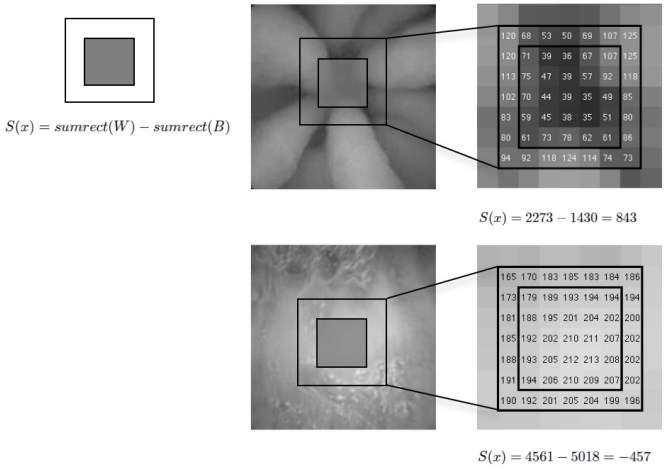

Figure 2. Evaluation of a "center-surround" feature in two images, with and without lumen respectively. $sumrect$ indicates the total value of pixels intensity within a rectangular region. W and B are related to light and dark regions of the rectangle.

of features that enhance the discriminative power produced by the simpler "center-surround" feature (Figure 3). The calculation of this second kind of features may be efficiently obtained as for the simpler "center-surround" features.

Using integral image representation, feature evaluation is accomplished by few memory accesses. It is straightforward to verify that to compute "center-surround" features, at any position or scale, only eight lookups are needed. The remaining two feature typologies require more accesses due to greater number of rectangular areas. "Cross-features" require respectively 16 and 24 references from the integral image.

Once a feature shape has been assigned, it is necessary to specify the position and the scale within the region of interest. Specifically, the features are scanned across the image top left to bottom right using a sliding offset of two pixels both in the horizontal and in the vertical direction. The process is iteratively repeated with different feature scales at each round. To keep the computation of the proposed features within the same number of lookups into the integral image, we choose not to change the scale of the image but to variate instead the size of the features.

The exact representation for the three proposed types of features is as follows:

$$f = [x_w, y_w, s_{wx}, s_{wy}, x_b, y_b, s_{bx}, s_{by}, type, \theta, \rho] \qquad (1)$$

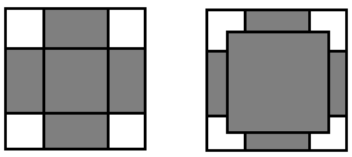

Figure 3. The two "cross-like" patterns used in the proposed method.

 ISBN: 978-1-61208-158-8

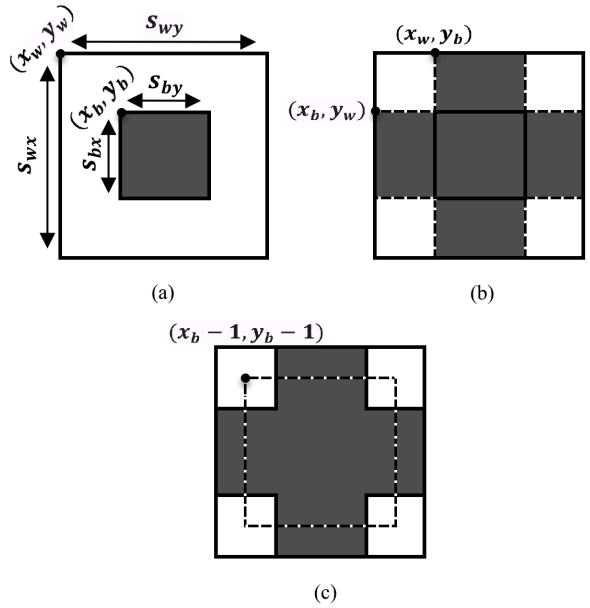

Figure 4. Schematic features representation. (a) Center-surround feature. (b) First cross feature obtained by center-surround feature considering the cross with width s_{by} and height s_{bx}. (c) Second cross feature obtained by the first taking into account a inner square of width and height greater than 1 pixel respect to the previous version.

The first four elements of f refer to the larger square of the feature. Similarly, the following four elements relate to the inner square. The $type$ parameter is an integer that indicates which type of feature is considered. The last two parameters are the optimal learned threshold and the polarity to register the category of images discriminated by that feature.

As discussed before, the "center-surround" features are evaluated considering difference between the sum of the pixels within two rectangular regions (Figure 4a). The second type of features considers a cross-shaped region to enhance lumen area. Location and size of this region are constrained by the size of correlated "center-surround" feature (Figure 4b). The third type of features is calculated in a similar way considering a larger internal rectangular area (Figure 4c). We consider the same total number of features for each type.

We consider only squared features, i.e., those with equal horizontal and vertical scale s_w. The internal region relative to lumen varies from a minimum size 2×2 up to $(s_w - 2) \times (s_w - 2)$ pixels. Once fixed the size of the external section, the descriptor associated with the lumen is shifted across the external descriptor resizing at each step (Figure 5).

The resolution of a WCE frame is reduced to 24×24 pixels in this phase of processing. Hence the total number of features per scale is equal to the total amount of different features in the image multiplied by the allowed variations of magnitude. For example, a 7×7 feature contains sixteen regions of size 2×2, nine of size 3×3, four of size 4×4 and one of size 5×5. Hence, the total number of features

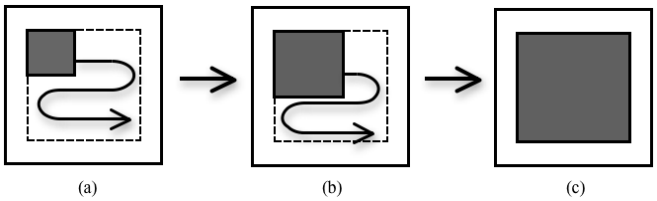

Figure 5. An example of features variations used in our algorithm. Given feature size, all 2×2 regions are considered in each location (a). This cycle is reiterated by increasing the size of the inner square (b) until maximum amplitude is achieved (c).

of size 7×7 is 2430, equal to the number of windows in the image (assuming a horizontal and vertical offset of two pixels) for the total number of variations. Table I summarizes the feature counting for the chosen scales.

A. Training

During the training phase, the dataset is rescaled to 24×24 pixels. The integral image representation of gray tone training samples is used to compute feature scores. Application of AdaBoost provides a list of best discriminative features. In particular, we build a binary classifier for each feature (weak classifier). Initially all the examples have the same weight. For each boosting step, the determination of a new weak classifier involves the evaluation of each feature on training data. The crucial feature is selected according to the weighted error that each feature shows on the training data. In the successive round, the samples are reweighted to emphasize the misclassified ones. This is the most expensive section of the training module.

The result of the training module is a strong classifier computed as weighted linear combination of weak classifiers built during each round of boosting. The whole boosting process is iterated, varying at each step the number of weak classifiers. The result is the realization of a cascade of strong classifiers with a gradually increasing number of features.

Learning requires that each strong classifier shows a prescribed detection rate, while maintaining a definite rate

Table I
FEATURES NUMBER PER SCALE. THE FIRST COLUMN REFERS TO THE SIZE OF THE FEATURE WHILE THE SECOND IS RELATED TO MAXIMUM SCALE ALLOWED FOR THE LUMEN AREA.

$Feature\ size$	$Max\ Internal\ scale$	$\#Features$	$\#Variations$	$Total$
5×5	3×3	100	5	500
7×7	5×5	81	30	2430
9×9	7×7	64	91	5824
11×11	9×9	49	204	9996
13×13	11×11	36	385	13860
15×15	13×13	25	650	16250
17×17	15×15	16	1015	16240
19×19	17×17	9	1496	13464
21×21	19×19	4	2109	8436

 ISBN: 978-1-61208-158-8

of false positives. In particular, it is required a minimum detection rate and a maximum false positive rate to be satisfied at every level of the cascade. For each strong classifier, a weak classifier is added until it reaches the required parameters for the current level of the cascade. Similarly, a new strong classifier is associate to the cascade until total false positive rate is above a certain threshold.

One of the advantages of the proposed system is that the user only needs to define the features set to be used and the false positives and detection rates for each level of the cascade. All the internal parameters are automatically selected during the training phase.

B. Test

In the proposed system, each test image is scaled to 24×24 pixels and it is labelled as "lumen" or "not lumen". This monoscale procedure combined with selection of best features during training allows real time application of our system. Please notice that, differently than in the case where the object to recognize may appear at different scales, in the present case a "mono-scale" choice has been shown adequate.

IV. Experimental Results

In this section, we report of the experiments carried out to verify the efficacy of the proposed method. To train the classifier, we have considered the integral images of a dataset made of 5000 images, 1500 positive and 3500 negative, rescaled to 24×24 pixels. The positive images have been manually selected from WCE videos previously labelled by the expert. The selected images represent a comprehensive set of scenes where the intestinal lumen can be present, including location and scale changes within the image. Differently, the negative examples have been randomly selected from videos that not contain any lumen. Both typical smooth images and images containing other judged negative events, like the presence of bubbles, bleedings, residuals, share this set.

To train the cascade of classifiers, we need to establish a maximum false positive rate and a minimum detection rate to satisfy to each layer of cascade. In particular, we require that 98% of positive images must be recognized at each level while maintaining a maximum amount of false positives equivalent to 80%. At the next levels of the cascade these two values are computed relatively to the new dataset whose positives set is composed by every lumen recognized as such by the previous classifier and the negatives set includes the remaining false positives. A strong classifier will be added to the cascade until the total false positive rate drops to zero. With these data, we have obtained a six level cascade for a total of 325 features (Figure 6). The total detection rate of the cascade, D, and the final false positive rate F, are obtained as a combination of intermediate outcomes on the cascade:

dataset	1500 L 3500 NL
d_1	98,33%
f_1	56,14%
features	3

1° *Strong classifier*

dataset	1475 L 1965 NL
d_2	98,37%
f_2	43,81%
features	19

2° *Strong classifier*

dataset	1451 L 861 NL
d_3	99,65%
f_3	34,95%
features	75

3° *Strong classifier*

dataset	1446 L 301 NL
d_4	99,30%
f_4	27,57%
features	76

4° *Strong classifier*

dataset	1436 L 83 NL
d_5	100%
f_5	1,2%
features	76

5 *Strong classifier*

dataset	1436 L 1 NL
d_6	100%
f_6	0%
features	76

6° *Strong classifier*

Figure 6. Cascade of classifiers. d_i and f_i represent detection and false positive rate at the ith level of cascade. L and NL indicate "lumen" and "not lumen" images, respectively.

$$D = \prod_{i=1}^{N} d_i = 95,7\% \qquad F = \prod_{i=1}^{N} f_i = 0\% \qquad (2)$$

where N is the total number of layers of the cascade.

To test the effectiveness of trained cascade, we have considered a collection of ~ 5000 images randomly extracted from a test set of frames disjoined from the training set. During testing phase, we consider the integral images of test set rescaled to 24×24 pixels with the respective labels, the cascade of boosted classifiers as it has been obtained during training and, finally, a threshold that determines the rigorousness of the classifier. Each test sample gets through each single node of the cascade; a positive outcome is sent by the classifier i to the more complex classifier $i + 1$. An image is labeled as lumen if positively overcomes each node of the cascade. If at any point the test image is judged negative, it is rejected immediately without further test. The classification performance has been evaluated in terms of precision and recall by comparing our results with the annotations provided by the specialist. Table II shows the results.

All experiments have been conducted on a consumer level PC with Intel®Core™2 Duo processor and 4 GB of RAM. Calculations have been performed in MATLAB environment.

By varying the rigidity threshold from a minimum to a maximum value, we can construct a ROC curve comparing the detection rate versus the number of false positives.

Table II
Classification results

Recall	*Precision*	*Accuracy*	*Time*
$90,5\%$	71%	$92,4\%(4642/5025)$	$600 frames/sec$

Higher threshold values minimizes both detection and false positive rates. Similarly, a low threshold will lead to acceptance of a greater number of lumen images while increasing the probability of detecting false positives. Figure 7 reveals that is possible to reach a detection rate above 90%, keeping the amount of false positives at about ~ 500 instances, i.e., 10% of the test dataset.

Figure 8 shows some examples of the false positives for the proposed method. In many circumstances, the intensity contrast between adjacent regions does not correspond to the presence of a lumen. A common problem is the fact that Haar features are sensitive to illumination changes. Variations on the lighting conditions may cause the cascade to detect lumen that was not predicted during the training stage. Likewise, in some images, folds of the intestinal wall may produce contrasted regions that confuse the Haar features. If new kind of images are presented to the classifier, detection is difficult and the amount of false positives increases. To deal with this problem, training data must include as many examples as possible to predict only true lumen.

V. Conclusion

In this paper we introduced an automatic lumen detection algorithm for endoscopic images. Inspired by Viola-Jones object detection system, we show that using AdaBoost learning-based algorithm combined with a cascade of strong classifiers leads to a good rate of detection minimizing running time. Experimental results show that the proposed system detects positive images using exclusively Haar-like proposed features. A cascade of six strong classifiers reaches a recall of 90,5% with a precision of 71%. Our detector is flexible and easily extensible to other semantic objects in endoscopic applications.

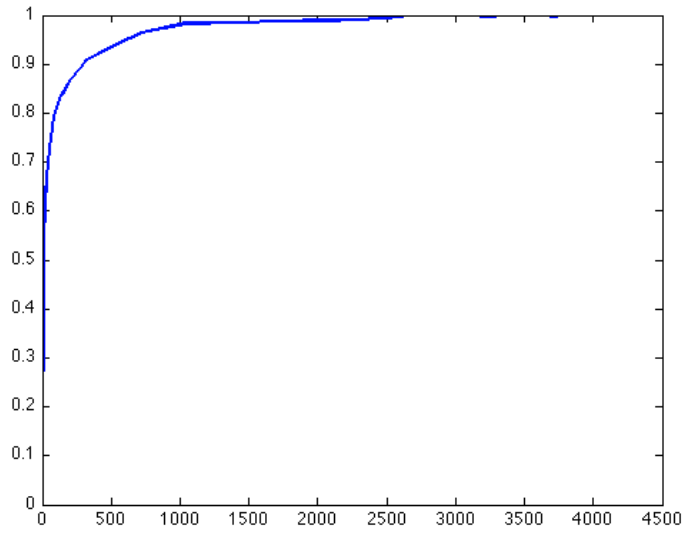

Figure 7. ROC curve for the detector with 325 weak classifiers.

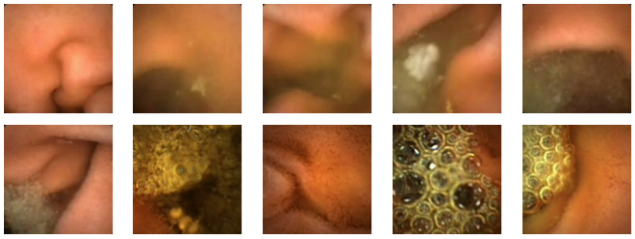

Figure 8. Example of some false positives detected by the system.

References

[1] G. Imaging, "Expanding the scope of gi," Last accessed: July 2011. [Online]. Available: http://www.givenimaging.com

[2] G. Iddan, A. Glukhovsky, and P. Swain, "Wireless capsule endoscopy," *Nature*, vol. 405, pp. 725–729, 2000.

[3] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, vol. 1, 2001, pp. 511–518.

[4] K. Asari, "A fast and accurate segmentation technique for the extraction of gastrointestinal lumen from endoscopic images," in *Medical Engineering and Physics*, vol. 22, 2000, pp. 89–96.

[5] X. Zabulis, A. Argyros, and D. Tsakiris, "Lumen detection for capsule endoscopy," in *Intelligent Robots and Systems, 2008. IROS 2008. IEEE/RSJ International Conference on*, sept. 2008, pp. 3921–3926.

[6] P. Spyridonos, F. Vilarino, J. Vitria, F. Azpiroz, and P. Radeva, "Anisotropic feature extraction from endoluminal images for detection of intestinal contractions," in *Medical Image Computing and Computer-Assisted Intervention*, 2006, pp. 161–168.

[7] G. Gallo and E. Granata, "Lbp based detection of intestinal motility in wce images," vol. 7961, no. 1. SPIE, 2011, p. 79614T. [Online]. Available: http://link.aip.org/link/?PSI/7961/79614T/1

[8] J. Lee, J. Oh, S. K. Shah, X. Yuan, and S. J. Tang, "Automatic classification of digestive organs in wireless capsule endoscopy videos," in *Proceedings of the 2007 ACM symposium on Applied computing*, ser. SAC '07. New York, NY, USA: ACM, 2007, pp. 1041–1045. [Online]. Available: http://doi.acm.org/10.1145/1244002.1244230

[9] P. Viola and M. Jones, "Robust real-time object detection," in *International Journal of Computer Vision*, 2001.

[10] F. C. Crow, "Summed-area tables for texture mapping," in *Proceedings of the 11th annual conference on Computer graphics and interactive techniques*, ser. SIGGRAPH '84. New York, NY, USA: ACM, 1984, pp. 207–212. [Online]. Available: http://doi.acm.org/10.1145/800031.808600

 ISBN: 978-1-61208-158-8

[11] R. E. Schapire, Y. Freund, P. Bartlett, and W. S. Lee, "Boosting the Margin: A New Explanation for the Effectiveness of Voting Methods," *The Annals of Statistics*, vol. 26, no. 5, pp. 1651–1686, 1998. [Online]. Available: http://dx.doi.org/10.2307/120016

[12] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," in *Proceedings of the Second European Conference on Computational Learning Theory*. London, UK: Springer-Verlag, 1995, pp. 23–37. [Online]. Available: http://portal.acm.org/citation.cfm?id=646943.712093

[13] Y. Freund and R. Schapire, "A short introduction to boosting," *J. Japan. Soc. for Artif. Intel.*, vol. 14, no. 5, pp. 771–780, 1999. [Online]. Available: citeseer.ist.psu.edu/freund99short.html

[14] P. Viola, M. Jones, and D. Snow, "Detecting pedestrians using patterns of motion and appearance," in *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, vol. 2, oct. 2003, pp. 734–741.

[15] L. Yun and Z. Peng, "An automatic hand gesture recognition system based on viola-jones method and svms," in *Computer Science and Engineering, 2009. WCSE '09. Second International Workshop on*, vol. 2, oct. 2009, pp. 72–76.

[16] M. Kolsch and M. Turk, "Analysis of rotational robustness of hand detection with a viola-jones detector," in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, vol. 3, aug. 2004, pp. 107 – 110 Vol.3.

[17] M. Castrillon-santana, O. Deniz-suarez, L. Anton-canals, and J. Lorenzo-navarro, "Face and facial feature detection evaluation performance evaluation of public domain haar detectors for face and facial feature detection," 2008.

[18] OpenCV, "Open computer vision library," Last accessed: July 2011. [Online]. Available: http://opencv.willowgarage.com

[19] C. Papageorgiou, M. Oren, and T. Poggio, "A general framework for object detection," in *Computer Vision, 1998. Sixth International Conference on*, jan 1998, pp. 555 –562.

 ISBN: 978-1-61208-158-8