Video-Specific SVMs for Colonoscopy Image Classification

Siyamalan Manivannan[1(✉)], Ruixuan Wang[1], Maria P. Trujillo[2], Jesus Arbey Hoyos[3], and Emanuele Trucco[1]

[1] CVIP, School of Computing, University of Dundee, Dundee, UK
msiyamalan@computing.dundee.ac.uk
[2] Escuela de Ingenieria de Sistemas y Computacion, Universidad Del Valle, Cali, Colombia
[3] Hospital Universitario del Valle Evaristo Garcia ESE, Cali, Colombia

Abstract. We propose a novel classification framework called the video-specific SVM (V-SVM) for normal-vs-abnormal white-light colonoscopy image classification. V-SVM is an ensemble of linear SVMs, with each trained to separate the abnormal images in a particular video from all the normal images in all the videos. Since V-SVM is designed to capture lesion-specific properties as well as intra-class variations it is expected to perform better than SVM. Experiments on a colonoscopy image dataset with about 10,000 images show that V-SVM significantly improves the performance over SVM and other baseline classifiers.

1 Introduction

Colorectal cancer is the second most common cause of cancer mortality among men and women [1]. Colonoscopy remains the gold standard for colorectal cancer screening because of its high sensitivity and specificity for detecting polyps and cancer [1]. Adenoma detection rate (ADR), in terms of lesion detection, is a surrogate marker of quality of colonoscopy [2]. An automated system detecting abnormalities (including polyps, cancer, ulcers, etc.) in colonoscopy videos would be a useful tool in improving ADR. Here, we concentrate on normal-abnormal white-light colonoscopy image classification, a challenging task as abnormalities in colon vary in size, type, color, and shape (Fig. 1).

While most colonoscopy image classification systems [3–9] focus on designing various image features, this paper focuses on designing a new classifier. The most popular classifier adopted in these classification systems is the support vector machine (SVM). In general, a binary SVM is trained to classify any colonoscopy image into one of two classes, e.g., normal versus abnormal [3,4,7,8], or normal versus a specific lesion, e.g., polyp [9]. In order to train a binary SVM for normal-vs-abnormal classification, a training dataset consisting of labeled normal and abnormal images need to be obtained in advance. Although each class of images are highly variable in appearance and textures (e.g., due to different colon segments, different patients, and different types of lesions in colons), such

X. Luo et al. (Eds.): CARE 2014, LNCS 8899, pp. 11–21, 2014.
DOI: 10.1007/978-3-319-13410-9_2

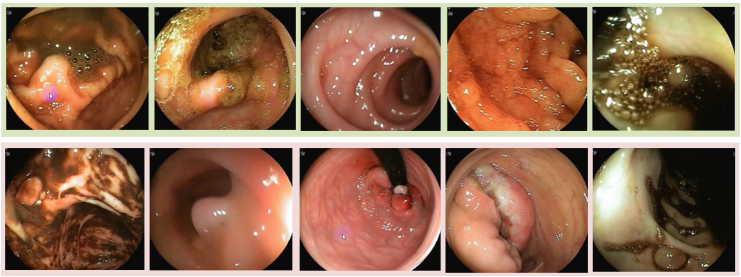


Fig. 1. Example images from the database. Normal (top) and Abnormal (bottom)

intra-class variations were not explored in the previous colonoscopy image classification systems [3–7,9]. In computer vision, it has been shown that considering the intra-class variations by learning a set of sub-class classifiers greatly improves the classification performance over a single classifier for natural images [10–13], because each sub-class often corresponds to a specific viewpoint or pose of the same class of objects and may therefore capture more detailed viewpoint-specific visual properties within the class. An extreme instance is the recently developed Exemplar SVMs (E-SVM) [14]. E-SVM is an ensemble of linear SVMs, with each SVM trained on a single positive example and a million of negative examples (Fig. 2c). However, E-SVM may not be appropriate for colonoscopy image classification: each SVM in the E-SVM may become highly over-fitting because the number of similar images for each exemplar positive (i.e., abnormal) image is often very limited due to the highly intra-class variations (Fig. 3).

In this paper we propose a new classifier, called video-specific SVMs (V-SVM), which can be considered as a trade-off between the traditional single SVM and the extreme E-SVM. V-SVM is an ensemble of linear SVMs, with each trained based on the set of positive images from a particular colonoscopy video and all the negative images from all the videos (Fig. 2b). Since each video, if containing lesions, often captures a particular type of lesion under different viewpoints and appearance variations, a video-specific SVM may capture that lesion-specific properties and becomes an expert to classify similar kinds of lesions under different viewpoints or appearance. Also, since the number of positive videos (which contain lesions) in the training data is very small compared to the number of positive images, V-SVM dramatically reduces the computational complexity compared to E-SVM. Our main contributions is the new classification framework called V-SVM, and experimental evidence suggesting that V-SVM outperforms SVM, E-SVM and some other baseline classifiers in colonoscopy image classification.

In the following, this paper first introduces the V-SVM (Sect. 2), and then empirically evaluate of the V-SVM (Sect. 3), followed by conclusions and future work (Sect. 4).

2 Video-Specific SVMs

Our objective is to learn a classifier from a set of videos, with each video frame labeled as abnormal (i.e., positive) or normal (i.e., negative). Assume that we have a number of $V = V^+ + V^-$ videos, consisting of V^+ abnormal and V^- normal videos. For each abnormal video, some video frames (i.e., positive images) contain a certain type of lesion, while the other video frames are normal (i.e., negative images). In comparison, all the images in each normal video are normal. For the v-th video ($v = 1, \ldots, V$), suppose there are $|\mathcal{N}_v^+|$ positive images and $|\mathcal{N}_v^-|$ negative images, with the index set for positive images denoted by $\mathcal{N}_v^+$, and the index set for negative images denoted by $\mathcal{N}_v^-$. Note that $|\mathcal{N}_v^+| = 0$ for any normal videos.

Give the training dataset $\{(\mathbf{x}_{vi}, y_{vi})|i = 1, \ldots, |\mathcal{N}_v^+| + |\mathcal{N}_v^-|; v = 1, \ldots, V\}$, where $\mathbf{x}_{vi} \in \mathbb{R}^d$ is the feature representation for the i-th image in the v-th video and $y_{vi} \in \{-1, +1\}$ represents the label for the image, the traditional (single) SVM classifier can be trained without considering any video-level information, e.g., which video does each image come from. Alternatively, an ensemble of exemplar SVMs (E-SVM) can be trained as proposed by [14]. Different from both the single SVM and the E-SVM, we propose a new SVM-based classifier, called video-specific SVM (V-SVM), which can be considered as a trade-off between the single SVM and the E-SVM.

2.1 The Optimization Function for Video-Specific SVMs

In the V-SVM, an ensemble of V^+ linear SVM classifiers $\{f_v(\mathbf{x})|v = 1, \ldots, V^+\}$ was learned, with each linear classifier $f_v(\mathbf{x}) = \mathbf{w}_v^T\mathbf{x} + \mathbf{b}_v$ corresponding to a specific abnormal video with index v, trying to discriminate all the positive images $\{\mathbf{x}_{vi}|\forall i \in \mathcal{N}_v^+\}$ in the abnormal video v from all the negative images $\{\mathbf{x}_{kj}|\forall j \in \mathcal{N}_k^-; k = 1, \ldots, V\}$ in all the videos including v. Learning the weight vector $\mathbf{w}_v$ and the bias $\mathbf{b}_v$ for a particular video-specific SVM classifier $f_v(\mathbf{x})$ can be achieved by solving the following SVM-like optimization problem, i.e.,

$$\min_{\mathbf{w}_v, \mathbf{b}_v} \|\mathbf{w}_v\|^2 + C^+ \sum_{i \in \mathcal{N}_v^+} h\left(\mathbf{w}_v^T\mathbf{x}_{vi} + \mathbf{b}_v\right) + C^- \sum_{k=1}^{V} \sum_{j \in \mathcal{N}_k^-} h\left(-\mathbf{w}_v^T\mathbf{x}_{kj} - \mathbf{b}_v\right) \quad (1)$$

where h is the hinge loss function $h(z) = \max(0, 1 - z)$, and C^+ and C^- are the regularization parameters for the imbalanced positive and negative classes.

With the objective function in Eq. 1, V^+ linear video-specific SVM classifiers will be independently trained, each trying to discriminate the positive images in a particular abnormal video from the negative images in all the videos (Fig. 2b).

2.2 Platt Calibration

The V^+ independently learned SVM classifiers need to be assembled to generate a final classifier. Different individual classifiers may have different ranges of

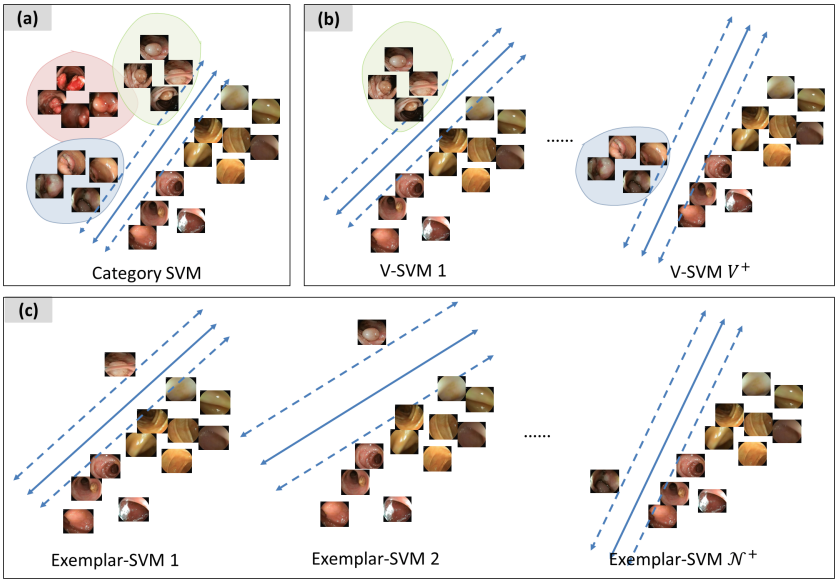


Fig. 2. Category SVM (a) vs. Exemplar SVMs (b) vs. V-SVMs (c). The video information is discarded in the category and exemplar SVMs. In V-SVM we train a set of SVMs; each separates the positive images of a particular video from all the negative images of all the videos. (Coloured ellipses group the positive images from each positive video)

output, making the outputs of video-specific SVM classifiers not directly comparable. As in [14], we use the Platt calibration method [15] to normalize individual classifiers such that their outputs are more directly comparable. The Platt calibration method [15] maps any SVM output $f_v(\mathbf{x})$ with the range $[-\infty, +\infty]$ to a posterior probability P_v with the range $[0, 1]$ by a sigmoid function, i.e.,

$$P_v(y = 1|f_v(\mathbf{x})) = \frac{1}{1 + \exp^{(a_v f_v(\mathbf{x}) + b_v)}} \tag{2}$$

where $P_v(\cdot)$ represents the probability of the image (represented by its feature $\mathbf{x}$) being positive.

To learn the parameters a_v and b_v for each sigmoid function $P_v(\cdot)$, a training set $T_v = \{f_v(\mathbf{x}_i), t_i\}$ is required, where $f_v(\mathbf{x}_i)$ is the individual SVM classifier's output for $\mathbf{x}_i$ and used as the input to the sigmoid function, and t_i is the expected output of the sigmoid function for the input $f_v(\mathbf{x}_i)$. Although t_i could be simply set by $t_i = (y_i + 1)/2$ where $y_i = -1$ for a negative image and $y_i = +1$ for a positive image, Platt [15] suggested using the regularized expected output to handle possible imbalance between the number of positive and negative training images, i.e.,

$$t_i = \frac{M_v^+ + 1}{M_v^+ + 2} \tag{3}$$

when the image $\mathbf{x}_i$ is positive, and

$$t_i = \frac{1}{M^- + 2} \tag{4}$$

when the image $\mathbf{x}_i$ is negative. M_v^+ and M^- are respectively the number of positive and negative images in the training set T_v.

To generate the training set T_v, Platt [15] suggested a cross-validation method. More specifically, the available training dataset for one video-specific classifier is randomly partitioned into L subsets, and then $L-1$ subsets are used to train the SVM classifier f_v. f_v is then used to obtain the predicted output scores $f_v(\mathbf{x})$ for all the images $\{\mathbf{x}\}$ in the remaining subset. This process is repeated L times, each time with a different remaining subset. The union of the predicted SVM scores and the corresponding t_i's are used to learn the sigmoid function. Such a process makes full use of the available training data to learn the sigmoid function, therefore reducing the possibility of over-fitting during sigmoid learning.

2.3 Ensemble of Video-Specific SVM Posteriors

Once all video-specific SVMs have been calibrated, they can be easily assembled to generate the final ensemble classifier $g(\mathbf{x})$. Since each calibrated SVM classifier P_v is only responsible for a specific video and therefore only valid to recognize a small part of positive images, the appropriate assembling choice is the maximum operation over all the video-specific classifiers when predicting the class of any new image [14], i.e.,

$$g(\mathbf{x}) = \mathbb{1}\{\max_v(P_v(\mathbf{x})) > \tau\}, \tag{5}$$

where $\mathbb{1}\{\cdot\}$ is the indicator function. As proposed by Platt [15] the optimal threshold τ is set to $\tau = 0.5$. This means, a new image $\mathbf{x}$ is predicted positive (i.e., $g(\mathbf{x}) = 1$) if at least one video-specific SVM classifier predicts that the image is positive. Otherwise, the image is predicted as negative.

3 Experiments

The proposed V-SVM was evaluated on a colonoscopy image dataset by comparing with the baseline methods including SVM, E-SVM, bagging-based ensemble of SVMs, and clustering-based ensemble of SVM.

3.1 Experimental Setup

Nine abnormal and ten normal videos (with the length of 8−15 min for each) were originally obtained from Hospital Universitario del Valle Evaristo Garcia ESE, Cali, Colombia. Each video was manually divided into non-overlapping normal and abnormal segments by clinical annotators. Due to high redundancy in visual information within each video (e.g., neighboring video frames are often very similar), each video was uniformly sampled at the rate of 3 images per

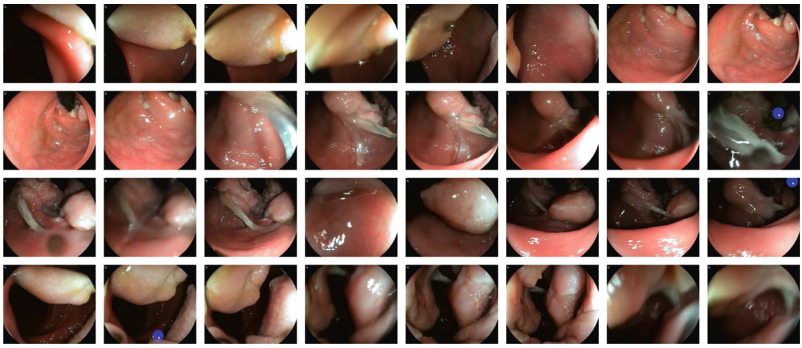


Fig. 3. The abnormal images from a particular video shows the variations in appearance.

second, and then a subset of representative images were further selected from the initially sampled images using K-means clustering. More specifically, if the initially sampled images from a video include N_1 positive images and N_2 negative images, K-means was applied to form $\frac{N_1}{2}$ clusters for positive images and $\frac{N_2}{4}$ clusters for the negative images. One frame per cluster is selected for the final dataset. In total, 10,658 images were selected from the nineteen videos to represent the final dataset, with 1856 images being positive and the rest being negative. All images were rescaled by preserving their row to column aspect ratio to make their maximum size (row or column) is 300 pixels. Some example images from the final dataset are shown in Figs. 1 and 3.

Each image in the dataset was represented based on sparse coding of two types of features, root-SIFT (rSIFT) [16] and multi-resolution local patterns (mLP) [17,18]. To learn a dictionary of visual words for each type of feature, 300,000 local features were randomly sampled from the training images, and then clustered into 2000 clusters using *K-means*, with each cluster center representing a visual word in the dictionary. To represent an image, patches with size 16×16 pixels were densely sampled from the image, with the sampling step being 4 pixels along both horizontal and vertical directions. Then, both rSIFT and mLP were extracted from each color channel for each image patch. Finally, for each of the two feature types, Locality Constrained Linear coding (LLC) [19] together with max-pooling was applied to all the local features (of the same type) to generate a 2000-dimensional feature vector.

When comparing the proposed V-SVM with other baseline classifiers, P percent of both positive and negative images were randomly selected from each video (but note that there is no positive image in normal videos) for training and the rest of the images for testing, where $P \in \{10, 20, \ldots 90\}$. Liblinear [20] was used to train the SVM classifiers. In all the experiments, the parameters C^+ and C^- (Eq. 1) was empirically set $C^+ = 50$ and $C^- = \frac{|\mathcal{N}_v^+|}{\sum_{k=1}^{V} |\mathcal{N}_k^-|} C^+$ for the v-th video-specific SVM, where $\frac{|\mathcal{N}_v^+|}{\sum_{k=1}^{V} |\mathcal{N}_k^-|}$ was used to deal with the imbalanced

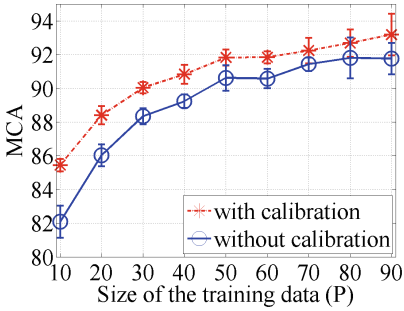


Fig. 4. V-SVM with and without Platt calibration. Vertical bars on each curve represent the standard devision of MCA over 10 runs.

dataset for two classes. For all other classifiers, SVM parameters were learned based on a 3-fold cross validation on the training set. For each exemplar SVM in E-SVM, 10 images with the highest exemplar SVM scores are considered as positive and used to learn the Platt function. Due to the imbalanced dataset, the average over true positive rate (or sensitivity) and true negative rate (or specificity), namely mean class accuracy (MCA), was used to evaluate each classifier's performance. All the experiments were repeated 10 times and the MCA results were averaged over all the 10 runs.

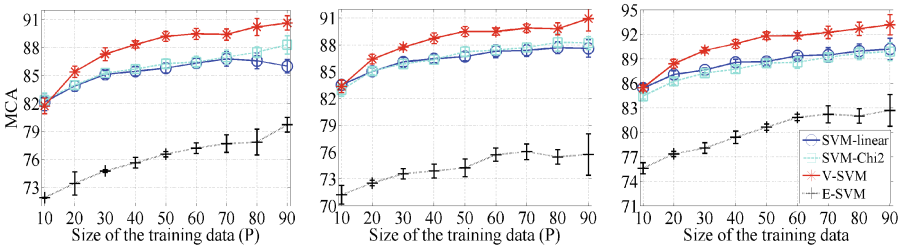
3.2 Effect of the Platt Calibration

To evaluate the effectiveness of the Platt calibration, the proposed V-SVM is compared with its variant version without the Platt calibration. In the variant V-SVM version, the ensemble classifier will classify a test image as positive if at least one video-specific SVM gives a positive output score. Figure 4 shows that inclusion of Platt calibration in V-SVM performs better than without Platt calibration for different sizes of training data. This is probably because Platt calibration can reduce the over-fitting issue which happened in individual video-specific SVMs (Sect. 2.2).

3.3 Performance of V-SVM

To evaluate the performance of V-SVM, the two most relevant classifiers, SVM and E-SVM, were used to compare with V-SVM with different sizes of training data. In all the following experiments Wilcoxon rank sum test at the significance level 0.01 was used to compare the difference in classification performance between the proposed V-SVM classifier and any other baseline classifier.

Figure 5 shows that V-SVM performs significantly better than SVM and E-SVM regardless of feature types. For example, the p-value is 1.8×10^{-4} when comparing the V-SVM with the linear SVM for $P = 30$. Similar significance results were obtained for other conditions (when $P \geq 20$) as demonstrated in



(a) SIFT (b) mLP (c) SIFT + mLP

Fig. 5. Comparison of SVM, E-SVM and V-SVM (MCA ± std).

Fig. 5. Note that V-SVM performs better even than Chi-square kernel SVM, probably due to the capability of capturing intra-class variations and lesion-specific properties of V-SVM. Linear SVM and Chi-square kernel SVM showed similar performance, which has also been observed in natural image classification when LLC encoding was used to represent images [21]. Another observation is that E-SVM performs significantly worse than SVM and V-SVM. In E-SVM, similar positive images of an exemplar are necessary to learn the Platt function [14]. The high variations in visual properties with each video (Fig. 3) make it difficult to find enough number of similar images for each exemplar positive image, which probably makes Platt calibration difficult and therefore leads to a worse performance of E-SVM. In addition, Fig. 5(c) also shows that combining the two features improves the performance of all the classifiers. Therefore in the following experiments only the combined features are considered.

3.4 Effect of Video-Specific Classifier Training

V-SVM is basically an ensemble classifier. To demonstrate that the better performance of V-SVM is not solely from the assembling of multiple classifiers, two other ensembles of classifiers were used to compare with V-SVM. One is the Bagging classifier [22], where a set of SVM classifiers are trained independently, with each trained based on a randomly chosen subset (here 80 %) of training images. The *majority voting* from all the individual SVM classifiers are used to predict the class of any new image [22]. We call this classifer 'bagging-SVM'. In our test, different number of SVM classifiers $\{5, 10, 15, 20\}$ in bagging are tried and the best performance is reported. Figure 6 shows that V-SVM performs significantly better than bagging-SVM when $P \geq 20$ (p-value $= 1.8 \times 10^{-4}$), suggesting that soley assembling of mutliple classifiers cannot explain the better performance of V-SVM.

Another baseline ensemble classifier is 'clustering-SVM'. For the proposed V-SVM, its better performance might come from (1) clustering the positive images into a set of clusters and (2) then learning a classifier to separate each cluster of positive images from all the negative images. To investigate this possibility, the positive images in the training dataset were clustered into V^+ clusters using

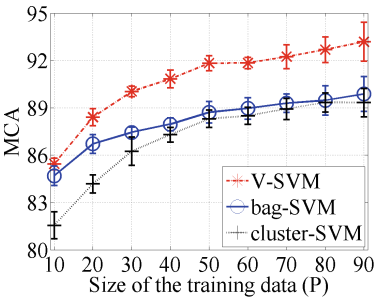


Fig. 6. Comparison of V-SVM with bagging-SVM and clustering-SVM (MCA $\pm$ std).

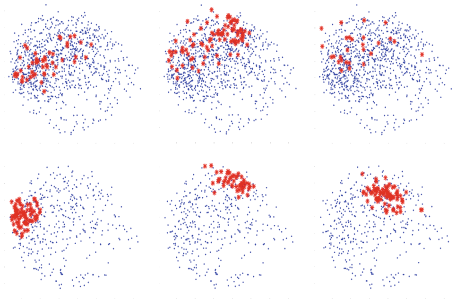


Fig. 7. Visualizations of positive (red) and negative (blue) images in 2D feature space. Positive images from three different videos usedfor V-SVM (first row) and three different positive clusters used for clustering-SVM (second row) (Color figure online).

k-means, and then V^+ linear SVM classifiers are trained as in V-SVM. The only difference between the 'clustering-SVM' and the proposed V-SVM is in the ways to cluster positive images for each individual SVM classifier. Again, Fig. 6 shows that V-SVM performs significantly better than the clustering-SVM for all different P (p-value < 0.01), suggesting that the better performance of V-SVM is not solely from the clustering of positive images into multiple subsets for subsequent classifier learning. Actually, by reducing feature vectors of all images into a 2D feature space via PCA and then visualizing the distribution of each set of positive images together with the distribution of all negative images, we observed that although positive images used for each individual SVM in the clustering-SVM are locally clustered (Fig. 7, second row), the positive images used for each video-specific SVM in the V-SVM are not clustered in local feature space (Fig. 7, first row). Such un-clustered property in the feature space may somehow help V-SVM identify more detailed lesion properties during the training, leading to better performance during testing.

4 Conclusions

This paper proposed a new ensemble classifier called V-SVM, which can be considered as a trade-off between single SVM and the E-SVM. Evaluations on a colonoscopy dataset shows that V-SVM performs significantly better than SVM, E-SVM, and other relevant ensemble classifiers. Future work will explore the possible empirical and theoretical reasons which cause better performance of V-SVM.

Acknowledgement. This work is funded by 2011–2016 EU FP7 ERC project "CODIR: colonic disease investigation by robotic hydrocolonoscopy", collaborative between the Universities of Dundee (PI Prof Sir A Cuschieri) and Leeds (PI Prof A Neville).

References

1. Winawer, S.J.: Colorectal cancer screening. Best Pract. Res. Clin Gastroenterol. **21**(6), 1031–1048 (2007)
2. Wallace, M.B.: Improving colorectal adenoma detection: technology or technique? Gastroenterology **132**, 1221–1223 (2007)
3. Manivannan, S., Wang, R., Trucco, E., Hood, A.: Automatic normal-abnormal video frame classification for colonoscopy. In: IEEE International Symposium on Biomedical Imaging (2013)
4. Manivannan., S., Wang, R., Trucco, E.: Extended gaussian-filtered local binary patterns for colonoscopy image classification. In: IEEE International Conference on Computer Vision Workshops (2013)
5. Kumar, R., Zhao, Q., Seshamani, S., Mullin, G., Hanger, G., Dassopoulos, T.: Assessment of crohn's disease lesions in wireless capsule endoscopy images. Biomed. Eng. Online **11**, 59 (2012)
6. Bejakovic, S., Kumar, R., Dassopoulos, T., Gerard Mullin, G.H.: Analysis of crohn's disease lesions in capsule endoscopy images. In: IEEE International Conference on Robotics and Automation (2009)
7. Li, P., Chan, K.L., Krishnan, S.: Learning a multi-size patch-based hybrid kernel machine ensemble for abnormal region detection in colonoscopic images. In: IEEE Conference on Computer Vision and Pattern Recognition (2005)
8. Li, P., Chan, K.L., Krishnan, S., Gao, Y.: Detecting abnormal regions in colonoscopic images by patch-based classifier ensemble. In: International Conference on Pattern Recognition (2004)
9. Zhao, Q., Meng, M.H.: Polyp detection in wireless capsule endoscopy images using novel color texture features. In: World Congress on Intelligent Control and Automation (2011)
10. Shan, Y., Han, F., Sawhney, H., Kumar, R.: Learning exemplar-based categorization for the detection of multi-view multi-pose objects. In: IEEE Computer Vision and Pattern Recognition (2006)
11. Li, S.Z., Zhu, L., Zhang, Z., Blake, A., Zhang, H., Shum, H.: Statistical learning of multi-view face detection. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002, Part IV. LNCS, vol. 2353, pp. 67–81. Springer, Heidelberg (2002)
12. Viola, M., Jones, M.J., Viola, P.: Fast multi-view face detection. In: Computer Vision and Pattern Recognition (2003)
13. Mohan, A., Papageorgiou, C., Poggio, T.: Example-based object detection in images by components. IEEE Trans. Pattern Anal. Mach. Intell. **23**, 349–361 (2001)
14. Malisiewicz, T., Gupta, A., Efros, A.A.: Ensemble of exemplar-svms for object detection and beyond. In: IEEE International Conference on Computer Vision (2011)
15. Platt, J.C.: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In: Smola, A.J., Bartlett, P., Scholkopf, B., Schuurmans, D. (eds.) Advances in Large Margin Classifiers. MIT Press, Cambridge (1999)

16. Arandjelović, R., Zisserman, A.: Three things everyone should know to improve object retrieval. In: IEEE Computer Vision and Pattern Recognition (2012)
17. Manivannan, S., Li, W., Akbar, S., Wang, R., Zhang, J., McKenna, S.J.: Hep-2 cell classification using multi-resolution local patterns and ensemble SVMs. In: ICPR I3A Workshop on Pattern Recognition Techniques for IIF Images (2014)
18. Manivannan, S., Li, W., Akbar, S., Wang, R., Zhang, J., McKenna, S.J.: Hep-2 specimen classification using multi-resolution local patterns and SVM. In: ICPR I3A Workshop on Pattern Recognition Techniques for IIF Images (2014)
19. Wang, J., Yang, J., Yu, K., Lv, F., Huang, T., Gong, Y.: Locality-constrained linear coding for image classification. In: IEEE Computer Vision and Pattern Recognition (2010)
20. Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: LIBLINEAR: a library for large linear classification. J. Mach. Learn. Res. **9**, 1871–1874 (2008)
21. Chatfield, K., Lempitsky, V., Vedaldi, A., Zisserman, A.: The devil is in the details: an evaluation of recent feature encoding methods. In: British Machine Vision Conference (2011)
22. Kim, H.-C., Pang, S., Je, H.-M., Kim, D., Bang, S.-Y.: Support vector machine ensemble with bagging. In: Lee, S.-W., Verri, A. (eds.) SVM 2002. LNCS, vol. 2388, pp. 397–408. Springer, Heidelberg (2002)

Springer

http://www.springer.com/978-3-319-13409-3

Computer-Assisted and Robotic Endoscopy
First International Workshop, CARE 2014, Held in Conjunction with MICCAI 2014, Boston, MA, USA, September 18, 2014. Revised Selected Papers
Luo, X.; Reichl, T.; Mirota, D.; Soper, T. (Eds.)

2014, X, 130 p., Softcover
ISBN: 978-3-319-13409-3