Enhancing Normal-Abnormal Classification Accuracy in Colonoscopy Videos via Temporal Consistency

Gustavo A. Puerto-Souza[1], Siyamalan Manivannan[2], Mariana Trujillo[3], Jesus Hoyos[4], Emanuele Trucco[2], and Gian-Luca Mariottini[1]

[1] Department of Computer Science and Engineering, University of Texas at Arlington, Texas, USA
[2] CVIP, School of Computing, University of Dundee, UK
[3] Escuela de Ingeniería de Sistemas y Computación, Universidad del Valle, Cali, Colombia
[4] Hospital Universitario del Valle Evaristo Garcia ESE, Cali, Colombia

Abstract. This paper proposes a novel hierarchical approach to improve the accuracy of the classification of normal-vs-abnormal frames in white-light colonoscopy videos. The existing approaches label each frame independently, without considering the temporal consistency between adjacent frames. Temporal consistency, however, can improve the classification accuracy in the presence of unclear/uncertain images. We propose to leverage temporal consistency between adjacent frames for colonoscopy video frame classification using a novel hierarchical classifier. Comparative experiments with five challenging full colonoscopy videos show that the proposed approach considerably improves the mean class normal/abnormal classification accuracy compared to the approaches where the frames are classified independently.

1 Introduction

Colorectal cancer is the second leading cause of cancer death in the world and the third most common cancer in the UK [1]. Although colonoscopy remains the gold standard for colorectal cancer screening, its miss rate for colorectal cancer has been reported to be as high as 6% [2], posing the risk of developing colon cancer due to failure to detect treatable lesions in time. This motivates research into automated, repeatable systems detecting abnormalities (including polyps, cancer, ulcers, etc.) in colonoscopy videos, which could provide a second quantitative opinion and ultimately contribute to reduce the miss rate.

In this paper, we concentrate on classifying white-light colonoscopy images into 2 classes, normal and abnormal. Abnormal frames contain one or more lesions (e.g., polyps, adenomas); normal frames contain none and show a healthy colon wall. The majority of the work reported for colonoscopy image classification focuses mainly on designing or identifying appropriate features and classifiers. Texture, color, shape and their combinations, together with different classifiers,

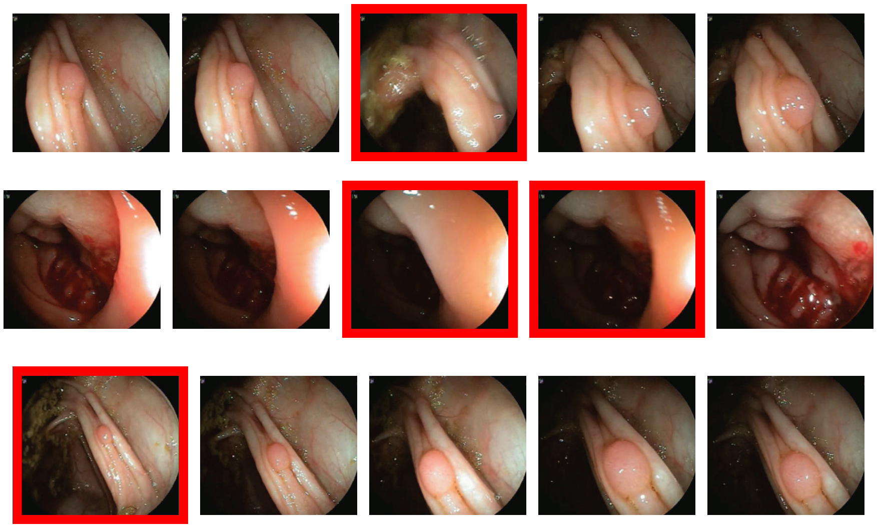


Fig. 1: Examples of three small video segments each contains 5 frames. The images which are difficult to classify due to (1) the lesion is not visible properly, (2) poor illumination, and (3) a very small lesion is highlighted in the 1st, 2nd and 3rd rows, respectively. These images, however, could be correctly classified as abnormal if the temporal information between adjacent frames were considered.

such as SVM and neural nets, have been explored for lesion detection and/or frame classification: texture features for normal/abnormal classification [3–5], lesion detection [6–8]; color histograms and related statistics for bleeding detection [9, 10]; and shape-based features, such as edge orientation histograms for Crohn disease classification [11]. For a complete review of the aforementioned methods, we direct the reader to [12].

Up to our knowledge, the state-of-the-art colonoscopy video frame classification approaches assume frames independent of each other. In reality, if a lesion appears in a particular frame, previous and successive frames are very likely to include it, albeit from different viewpoints as the scope is moved. One expects, therefore, that temporal consistency should improve the accuracy of colonoscopy frame classification compared to single-frame schemes.

There are further reasons to expect that temporal consistency will improve the classification. First, some frames are genuinely ambiguous, and a single view will not be sufficient for reliable classification even for experts, whose decisions are based on multiple observations generated by moving the scope. Second, the colonic wall may not be clearly visible in specific frames due to poor illumination, blur due to fast camera movements, and surgical smoke. Third, the appearance of lesions (e.g., scale, orientation) varies in different frames. Fourth, frame-level representations for classification are often obtained by aggregating the statistics of the local features extracted from that frame (e.g. bag-of-visual-words). Such representations may not capture small lesions sufficiently well, vis-á-vis the volume and appearance of background features (extracted from normal tissue).

Figure 1 shows three example video sequences, each containing a few frames which are difficult to classify. A system trained on individual frames independently is likely to classify these frames erroneously as normal. However, a classifier using temporal consistency information would classify these frames correctly as abnormal.

In this paper, we propose a three-level hierarchical classification approach which makes use of the temporal-context information across adjacent frames to classify any individual frame. In the first level, we assume the frames are independent to each other, hence we learn a classifier based on individual frame-level representations. The second level classifier is trained to leverage the temporal consistency information using the weighted similarities between frames in a temporal window and the classification outputs computed from the first level. We propose a max-margin approach to learn these weights based on the given training set. The third level applies a temporal filtering which refines the output from the second level by majority voting. We experimentally show that the proposed hierarchical approach outperforms the single-level classifier approaches such as SVM and random forests which were trained to classify frames independently. Note that our technique could be used to assess proficiency of gastroenterologist doctors either by analyzing colonoscopy videos both retrospectively or in real time depending on the parameters of the sliding window.

In the following, we first we explain the proposed hierarchical classification approach in detail, and then provide experimental evidence showing that the proposed approach performs better than any single level classification approach.

2 Methodology

In this section, we present an algorithm to classify normal-abnormal frames in colonoscopy videos. Our approach is based on a three-layer hierarchical classifier that leverages the strengths of SVM, in terms of accuracy and robustness, and the temporal consistency between adjacent frames based on a max-margin formulation.

In our proposed approach, we make use of the similarities between adjacent frames, in addition to the frame-level features. The similarities (e.g., number of image correspondences) between adjacent frames play an important role in this classification. Lets consider two consecutive frames I_i and I_j, if I_i has a high similarity with I_j it is most probable that both I_i and I_j are belonging to the same class.

Our approach is illustrated in Fig. 2. In the first level, frames are assumed to be independent to each other, and a SVM is trained to classify frames independently based on the frame-level features. In the second level, we make use of the temporal-context information between adjacent frames; which are measured by weighted similarity between a frame and its temporal neighbors, as well as the outputs obtained by the first level classifier. We propose an approach to learn these weights by maximizing the margin between normal and abnormal classes.

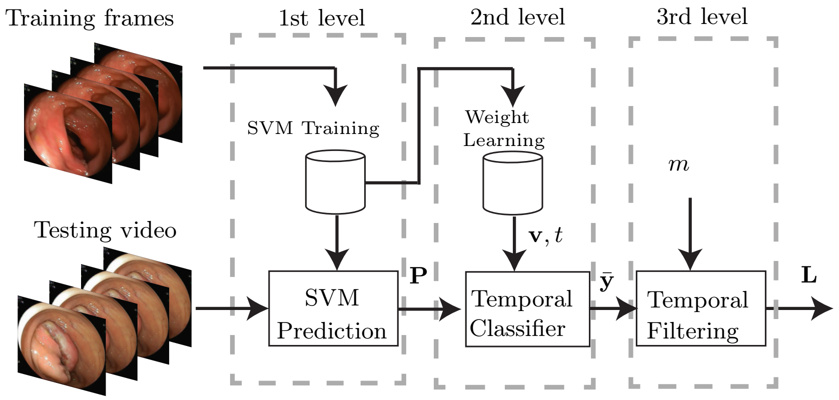


Fig. 2: The proposed hierarchical classifier. The first level outputs the confidence values based on classifying independent frames. The second level uses these confidence values in addition to the similarities between adjacent frames. The final level applies a majority voting on the second-level outputs to obtain the final labels of individual frames.

Finally, the resulting classification is passed to a third level that refines further the output from the second level by using a voting scheme over adjacent frames.

In the following, first we describe the first-level classifier and the Platt scaling which is used to convert the outputs of the first-level classifier to probability values. Then, the max-margin formulation of the second-level classifier is explained in detail. Lastly, the section concludes with the temporal filtering.

2.1 The first-level classifier

This classifier is trained on individual-frame representations to classify each test frame independently, i.e. without considering its temporal context.

Since the number of abnormal and the normal frames are highly unbalanced, we use a SVM with class balancing [13]. Learning the SVM weight vector $\mathbf{w}$ and the bias (b) for the first-level classifier $f(\mathbf{x})$ is achieved by the following formulation,

$$\arg\min_{\mathbf{w},b} \left\{ \|\mathbf{w}\|^2 + \lambda \left[C^+ \sum_{i \in A} h(\mathbf{w}^T \mathbf{x}_i + b, y_i) + C^- \sum_{j \in N} h(\mathbf{w}^T \mathbf{x}_j + b, y_i) \right] \right\} \quad (1)$$

where h is the hinge loss function $h(z,y) = \max(0; 1 - yz)$, with $\mathbf{x}_i$ and $y_i = \{-1, 1\}$ are the feature representation for I_i (the i^{th} frame) and its label, respectively. λ is a regularization parameter controlling the rate of miss-classification, and C^+ and C^- are the class weighting parameters for the unbalanced abnormal (A) and the normal (N) classes, respectively. C^+ and C^- can be selected by setting $\frac{C^+}{C^-} = \frac{n^+}{n^-}$ [13], where n^+ and n^- are the total number of positive (abnormal) and the negative (normal) images in the training set.

Usually SVM outputs decision values represent how far the test feature is from the learned hyper-plane, which is defined by $(\mathbf{w}, b)$. The Platt calibration method [14] maps any SVM output $f(\mathbf{x}_i)$ with the range $[-\infty, \infty]$ to a posterior probability P with the range $[0, 1]$ by a sigmoid function, i.e.,

$$P(y = 1|f(\mathbf{x}_i)) = \frac{1}{1 + \exp(Af(\mathbf{x}_i) + B)} \tag{2}$$

where $P(\mathbf{x}_i)$ represents the probability of the ith image being positive. A and B are two parameters which has to be learned from the training set. As suggested by Platt [14], we use a three-fold cross validation on the training set to learn these parameters.

2.2 The second-level classifier:

This classifier aims to improve the classification accuracy of the first classifier by leveraging temporal consistency. The inputs are the probabilities obtained by the first-level classifier, as well as the similarities, in terms of image correspondences, between a frame and its neighbors.

Similarity between frames: We defined the similarity S_{ij} between two adjacent frames, I_i and I_j, as the number of image correspondences between them. In particular, we extract and match SIFT features because of their stability, distinctiveness, and repeatability, as well as their well known rotation and scale invariance, and robustness to affine distortions, illumination changes, and noise [15]. SIFT detects a sparse set of interest points (keypoints), in the image, obtained as the scale-space extrema of the difference of Gaussians operators. The extracted keypoints are matched according to the nearest neighbor distance ratio of their descriptors, discarding ambiguous matches with ratio greater than 0.8 [15].

The temporal classifier: The proposed temporal classifier assumes that the label of a particular frame I_i not only depends on the classification results of itself, but also on the weighted similarity between that frame and its neighbors as well as on the confidence values of its neighbors. From here and the following we will assume a centered sliding window since our approach targets for maximal performance over retrospective videos. However, our approach can achieve real time performance by using a queue-style sliding window.

Let $P_i = P(y_i = c)$ and $P_j = P(y_j = c)$ represents the probabilities obtained by the first-level classifier for the frames I_i and I_j. We define the label of the frame I_i based on the temporal classifier as follows,

$$\begin{aligned} d_i &= v_i P_i + \sum_{\substack{j=-n \\ j \neq 0}}^{n} v_j S'_{i,j} P_j \\ \bar{y}_i &= \begin{cases} 1 & \text{if} \quad d_i \geq t \\ -1 & \text{otherwise} \end{cases} \end{aligned} \tag{3}$$

where the set $\{v_j\}_{j=-n}^{n}$ are the weights applied to the current frame ($j = 0$) and its neighboring frames in the interval $[-n, n]$. Here t denotes the margin between classes, the size of the considered temporal window is represented by $2n+1$ (i.e., previous n and next n frames are considered around the frame I_i), and $\bar{y}_i$ is the predicted label for the frame I_i. S'_{ij} can be represented by

$$S'_{i,j} = 1 - \exp^{-\beta S_{i,j}} \tag{4}$$

where β is a decay parameter, empirically set to $\beta = 5$ in all the experiments reported in Sect. 3.

Lets define the vectors $\mathbf{v}$ and $\mathbf{u}$ be

$$\mathbf{v} = \begin{pmatrix} v_{i-n} \\ \vdots \\ v_{i-1} \\ v_i \\ v_{i+1} \\ \vdots \\ v_{i+n} \end{pmatrix}, \quad \mathbf{u} = \begin{pmatrix} P_{i-n} S_{i,i-n} \\ \vdots \\ P_{i-1} S_{i,i-1} \\ P_i \\ P_{i+1} S_{i,i+1} \\ \vdots \\ P_{i+n} S_{i,i+n} \end{pmatrix} \tag{5}$$

The classifier defined in Equation (3) can be represented based on vector representations as follows,

$$\bar{y}_i = \begin{cases} 1 & \text{if} \quad \mathbf{v}^T \mathbf{u} - t \geq 0 \\ -1 & \text{otherwise} \end{cases} \tag{6}$$

where $\mathbf{v}$ and t define the temporal classifier, and can be easily learned in a similar manner to the max-margin approach given by Eq. (1).

2.3 Temporal Filtering

This final level refines further the results of the second-level by enforcing, within a sliding window, a temporal constraint based on the classes of the surrounding frames. As a result, the video has smoother transitions between abnormal and normal classes, i.e., the labels of video frames in segments containing lesions are consistently "abnormal", and do not contain noisy "normal" labels surviving the previous classifiers.

We use the second classifier prediction $\bar{y}_i$ to classify the frames, based on a majority-vote scheme over a sliding window. In particular, for each frame I_i, we gather the second-level classifier labels within a window with size $2m + 1$, centered on frame i. Each element within the window yields a vote for either abnormal or normal according to their class $\bar{y}_i$. The frame I_i is classified as the class with the larger number of votes. For example, the frame I_i is classified as abnormal if $C^A_{i,m} > C^N_{i,m}$, where $C^A_{i,m}$ and $C^N_{i,m}$ denote the number of votes for abnormal and normal classes within the window, respectively.

3 Experiments

The aim of these experiments is to compare different classifiers, with and without the hierarchical approach to incorporate temporal consistency, while keeping all the other factors unchanged, e.g. features for computing the frame representations.

In the following dataset, experimental settings and evaluation criteria are first explained. Then experimental validation and analysis of the results are presented.

3.1 Experimental Setup

We define abnormal frames as those that contain various lesions including polyps, cancer and bleeding. Our dataset consists of frames extracted from five colonoscopy videos (1 normal and 4 abnormal) from Hospital Universitario del Valle Evaristo Garcia ESE, Cali, Colombia. Each video has length of 8 - 15 minutes, image resolution of 640×480 and was recorded at 10 frames per second, leading to a total of 41518 extracted frames. For training and evaluation, the entire dataset was annotated at frame-level by an expert colonoscopist. In our two-label scheme and since lesion detection is the clinical target, large blurs and negligible frames were labeled as normal. The number of frames from different classes are given in Table 1; notice that the normal frames (N) constitute 77.5% of the dataset while the 22.5% of the frames are labeled as abnormal (A). All these frames were then rescaled by preserving their row to column aspect ratio to make their maximum size (row or column) equal to 300 pixels.

Frame representation: Each frame in the dataset was represented based on the Locality-constrained Linear Coding (LLC) [16] together with max-pooling on two types of local features: local color histograms and multi-resolution local patterns [17]. These features were extracted from patches of size 16×16 with an overlap of 12 pixels in the horizontal and vertical directions. Since the dimensionality of the local color histogram features are high (equal to 3 colors $\times$ 256 bins), we applied PCA to reduce its dimension to 400. Separate dictionaries of size 500 were learned for each feature type using k-means on a randomly sampled $200,000$ features from the training set. Finally each frame was represented as a feature vector of size 1000, which is a concatenation of the frame representation obtained by each feature type.

video	N	A	%A frames
1	5173	2944	36.3
2	3082	2555	45.3
3	8033	2056	20.4
4	5892	1823	23.6
5	9960	0	0
Total	32140	9378	22.6

Table 1: The number of frames per video in each class (N-normal, A-abnormal)

Evaluation criteria: The classification performance was evaluated based on leave-one-video-out experiments. Due to the highly unbalanced nature of the dataset, the average of the true positive rate (or sensitivity) and true negative rate (or specificity), namely the mean class accuracy (MCA), was used to evaluate the classification performance.

LibLinear [18] was used to train the SVM classifier. The regularization parameter of SVM is learned based on a three-fold cross validation applied on the training set. The `vlfeat` library [19] was used to create the dictionary and to extract the SIFT matches. The code from the authors of [16] was used for LLC.

3.2 Temporal consistency for classification

This section compares a single-layer SVM classifier, which is trained to classify frames independently, with the proposed hierarchical classifier which incorporates the temporal consistency.

Let SVM-TC and SVM-TF represent the second and the third level classifiers proposed in Section 2.2 and Section 2.3 respectively. Table 2 reports the MCA obtained by the single level (first row) and the proposed hierarchical (second and third rows) approaches for different videos.

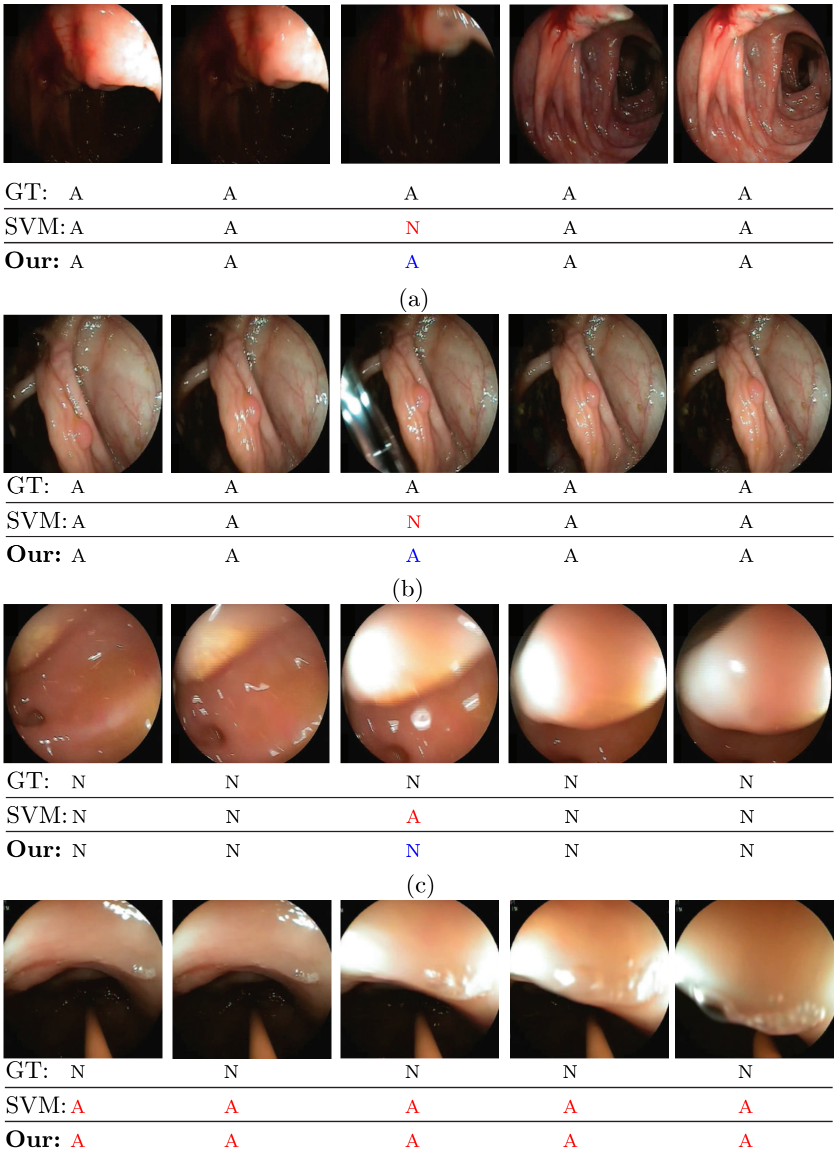
Method	video 1	video 2	video 3	video 4	video 5
SVM	66.8	73.9	89.4	73.7	98.3
SVM-TC	**73.2**	84.7	90.5	74.6	99.1
SVM-TF	72.3	**84.9**	**91.5**	**75.9**	**99.4**
% improvement	**6.4**	**11.0**	**2.1**	**2.1**	**1.1**

Table 2: MCA per video with (2nd and 3rd rows) and without (1st row) the proposed hierarchical approach. SVM was used as the first-level classifier. The fourth row contains the percentage of improvement achieved by our approach (SVM-TF) with respect to the single-layer SVM.

As expected for all the videos adding temporal information considerably improve the MCA. The third level classifier gives modest improvements over the second level one, suggesting that the second level classifier already captures the temporal consistency information.

Figure 3 illustrates a qualitative comparison between the first level SVM and our approach. Note in Fig. 3(a-c) that the single-frame approach of SVM classifies erroneously few ambiguous frames, instead our approach, correctly classifies these frames by propagating the classification of SVM from more certain frames towards ambiguous ones. The example in 3(d) shows a challenging case when our approach obtains an incorrect classification, however this is mainly due to the classification obtained by the first level SVM classifier, which in this example is erroneous for the whole subsequence.

In this experiment the window sizes was empirically set to $n = 10$ for the second layer classifier and $m = 5$ for the third layer classifier respectively.



GT:	A	A	A	A	A
SVM:	A	A	N	A	A
Our:	A	A	A	A	A

(a)

GT:	A	A	A	A	A
SVM:	A	A	N	A	A
Our:	A	A	A	A	A

(b)

GT:	N	N	N	N	N
SVM:	N	N	A	N	N
Our:	N	N	N	N	N

(c)

GT:	N	N	N	N	N
SVM:	A	A	A	A	A
Our:	A	A	A	A	A

(d)

N = Normal, A = Normal, SVM = SVM without temporal context, GT = Ground-Truth

Fig. 3: Qualitative example of the performance of the first level SVM and our proposed algorithm over four video subsequences (a-d) where our hierarchical classifier is able to correct the misclassified frames by enforcing temporal constraints.

Method	video 1	video 2	video 3	video 4	video 5
RF	58.6	62.2	90.3	63.9	**100**
RF-TC	**59.6**	**68.0**	**91.6**	**67.2**	**100**
improvement	**1.0**	**5.8**	**1.3**	**3.3**	0

Table 3: MCA per video with (2nd and 3rd rows) and without (1st row) the proposed hierarchical approach. RF was used as the first-level classifier.

3.3 Generalization to other classifiers

The goal of this section is to show the applicability of our approach with respect to other first-level classifiers, i.e., by replacing the SVM classifier (used in Sect. 3.2) with a Random Forest (RF) classifier.

Table 3 reports the MCA for RF with and without the temporal consistency. Adding the temporal consistency to the RF considerably improves the MCA for most of the videos. However, SVM without temporal information (Table 2) obtains better or very competitive results than RF without temporal information. When temporal consistency is added, SVM with temporal context performs better than RF with temporal context.

The number of trees in the RF classifier was set to 200 since we observed that increasing the number of trees leads to poor performance. This might happen because RF require very large training sets to perform optimally.

4 Conclusions

We presented here a novel three-layer classifier to detect normal-abnormal frames in a colonoscopy video. Differently from other methods, our approach hierarchically combines the accuracy and robustness of SVM with the temporal consistency of two temporal classifiers. Experimental evaluation over five challenging colonoscopic videos shown improved classification accuracy, with two cases with significant improvements of 8.5% and 14.9%, when comparing against a SVM approach without any temporal information. Future work will be directed towards investigating other classification approaches as well as quantifying the impact of uninformative frames in the classification process.

References

[1] : Cancer research uk. `info.cancerresearchuk.org/cancerstats`

[2] Bressler, B., Paszat, L.F., Chen, Z., Rothwell, D.M., Vinden, C., Rabeneck, L.: Rates of new or missed colorectal cancers after colonoscopy and their risk factors: A population-based analysis. Gastroenterology **132**(1) (2007)

[3] Lima, C., Barbosa, D., A.Ramos, A.Tavares, L.Montero, Carvalho, L.: Classification of endoscopic capsule images by using color wavelet features, higher order statistics and radial basis functions, IEEE EMBS (2008)

[4] Manivannan, S., Wang, R., Trucco, E.: Extended gaussian-filtered local binary patterns for colonoscopy image classification. In: IEEE ICCV Workshops. (2013)

[5] Manivannan, S., R.Wang, E.Trucco, A.Hood: Automatic normal-abnormal video frame classification for colonoscopy. In: IEEE ISBI. (2013)

[6] Engelhardt, S., Ameling, S., Paulus, D., Wirth, S.: Features for classification of polyps in colonoscopy, CEUR Workshop Proceedings (2010)

[7] Karkanis, S.A., Iakovvidis, D.K., Maroulis, D.E., Karras, D.A., Tzivras, M.: Computer aided tumor detection in endoscopic video using color wavelet features. IEEE transactions on IT in biomedicine **7** (2003)

[8] Maroulis, D.E., Iakovidis, D.K., Karkanis, S.A., Karras, D.A.: Cold: a versatile detection system for colorectal lesions in endoscopy video-frames. Computer Methods and Programs in Biomedicine **70** (2003)

[9] Cui, L., Hu, C., Zou, Y., Meng: Bleeding detection in wireless capsule endoscopy images by support vector classifier, IEEE Int. Conf. on Information and Automation (2010)

[10] Tjoa, M.P., Krishnan, S.: Feature extraction for the analysis of colon status from the endoscopic images. Biomedical engineering online (2003)

[11] Kumar, R., Zhao, Q., Seshamani, S., Mullin, G., Hanger, G., Dassopoulos, T.: Assessment of crohn's disease lesions in wireless capsule endoscopy images. Biomedical engineering online **59** (2012)

[12] Liedlgruber, M., Uhl, A.: Computer-aided decision support systems for endoscopy in the gastrointestinal tract: A review. Biomedical Engineering, IEEE Reviews in (2011)

[13] Ben-Hur, A., Weston, J.: A user's guide to support vector machines. In: Data Mining Techniques for the Life Sciences. Volume 609 of Methods in Molecular Biology. Humana Press (2010) 223–239

[14] Lin, H.T., Lin, C.J., Weng, R.: A note on platt's probabilistic outputs for support vector machines. Machine Learning **68**(3) (2007) 267–276

[15] Lowe, D.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision **60**(2) (2004) 91–110

[16] Wang, J., Yang, J., Yu, K., Lv, F., Huang, T., Gong, Y.: Locality-constrained linear coding for image classification. In: IEEE CVPR. (2010)

[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: I3A 1st workshop on Pattern Recognition Techniques for Indirect Immunoflurescence Images, ICPR. (2014)

[18] Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: LIBLINEAR: A library for large linear classification. Journal of Machine Learning Research **9** (2008) 1871–1874

[19] Vedaldi, A., Fulkerson, B.: VLFeat: An open and portable library of computer vision algorithms. `http://www.vlfeat.org/` (2008)