Evolving Systems (2020) 11:409–421
https://doi.org/10.1007/s12530-018-9236-x

ORIGINAL PAPER

Weakly supervised multilabel classification for semantic interpretation of endoscopy video frames

Michael D. Vasilakakis[1] · Dimitris Diamantis[1] · Evaggelos Spyrou[1] · Anastasios Koulaouzidis[2] · Dimtris K. Iakovidis[1]

Received: 5 January 2018 / Accepted: 17 May 2018 / Published online: 25 May 2018

Abstract

Several studies have addressed the problem of abnormality detection in medical images using computer-based systems. The impact of such systems in clinical practice and in the society can be high, considering that they can contribute to the reduction of medical errors and the associated adverse events. Today, most of these systems are based on binary classification algorithms that are "strongly" supervised, in the sense that the abnormal training images need to be annotated in detail, i.e., with pixel-level annotations indicating the location of the abnormalities. However, this approach usually does not take into account the diversity of the image content, which may include a variety of structures and artifacts. In the context of gastrointestinal video-endoscopy, addressed in this study, the semantics of the normal contents of the endoscopic video frames include normal mucosal tissues, bubbles, debris and the hole of the lumen, whereas the abnormal video frames may include additional semantics corresponding to lesions or blood. This observation motivated us to investigate various multi-label classification methods, aiming to a richer semantic interpretation of the endoscopic images. Among them, an image-saliency enabled bag-of-words approach and a convolutional neural network architecture enabling multi-scale feature extraction (MM-CNN) are presented. Weakly-supervised learning is implemented using only semantic-level annotations, i.e., meaningful keywords, thus, avoiding the need for the resource demanding pixelwise annotation of the training images. Experiments were performed on a diverse set of wireless capsule endoscopy images. The results of the experiments validate that the weakly-supervised multi-label classification can provide enhanced discrimination of the gastrointestinal abnormalities, with MM-CNN method to provide the best performance.

Keywords Endoscopy · Video analysis · Lesion detection · Weakly supervised learning · Multi-label classification · Bag-of-words · Convolutional neural networks

1 Introduction

Multi-label classification is a special case of data classification, where multiple labels may be assigned to a given instance. One may consider it as a generalization of multi-class classification, which enables the semantic characterization of data instances by labels that are not mutually exclusive.

One of the most exciting and continuously growing research areas in computer vision is the understanding of the visual image content. The majority of research efforts have turned to the automatic extraction of semantic annotations, aiming to imitate the way humans perceive and describe such content. This problem is often referred to as "bridging the semantic gap" (Smeulders et al. 2000) and consists of automatic extraction of high-level semantics from a given image, based on low-level features computed from raw data

✉ Dimtris K. Iakovidis
dimitris.iakovidis@ieee.org; diakovidis@dib.uth.gr; diakovidis@uth.gr

Michael D. Vasilakakis
vasilaka@uth.gr

Dimitris Diamantis
didiamantis@uth.gr

Evaggelos Spyrou
espyrou@uth.gr

Anastasios Koulaouzidis
akoulaouzidis@hotmail.com

[1] Department of Computer Science and Biomedical Informatics, University of Thessaly, Papasiopoulou 2-4, 35131 Lamia, Greece

[2] Endoscopy Unit, The Royal Infirmary of Edinburgh, Edinburgh, UK

(pixels). To this goal, semantic concepts are formalized, learned, and ultimately linked to their linguistic representation. The semantic content of a video frame may not be completely characterized by a single annotation, as it may contain more than one semantic concepts; therefore, the need for multi-label classification becomes evident.

Moreover, semantic interpretation of images can become even more challenging by extending its application to video scale. State-of-the-art works on semantic video interpretation are mainly based on supervised machine learning algorithms capable of classifying the contents of the video frames into semantically relevant categories (Li et al. 2016). In this paper we address the application domain of gastrointestinal (GI) video endoscopy, and we focus on wireless capsule endoscopy (WCE) as a more challenging application domain (Iakovidis and Koulaouzidis 2015). The video frames obtained from such an endoscopic examination are generally characterized as normal or abnormal depending on whether they contain abnormalities, such as lesions or blood. However, normal frames may include content belonging to a variety of semantic categories such as normal mucosa, bubbles, and debris. Also, the content of the abnormal video frames may include one or more kinds of abnormalities, as well as normal content.

Conventionally, the ground truth description of video frames in medical applications is performed by domain experts, using graphical annotations, i.e., the experts manually select and specify the boundaries of regions that correspond to the different semantic concepts of the video frames. However, such a process requires a large amount of effort, which may prove particularly costly, or even be inhibitory for large video frame collections. Therefore, there exists the need for more efficient approaches to video content annotation. Weakly-supervised learning (Hoai et al. 2014) comprises such an approach, as it requires only frame-level annotations for the training of the supervised classification system, i.e., the annotations indicate either the presence or the absence of the semantic concepts per training video frame.

A popular methodology to implement weakly-supervised learning is based on the Bag-of-visual-Words (BovW/BoW) feature extraction technique (Vasilakakis et al. 2016; Wang et al. 2015; Yu et al. 2012; Yuan et al. 2016a, b). BoW creates a histogram-like representation of a given video frame, based on a visual vocabulary that has been constructed using the available training set of video frames. More specifically, patches (i.e, regions) are extracted from video frames and low-level feature vectors are formed. Upon a clustering procedure, the aforementioned vocabulary is typically created by using cluster centroids. The latter are commonly referred to as "visual words." For a new video frame, each patch is assigned to a visual word based on the proximity of their corresponding feature vectors. Then, its visual content may be described by a histogram of these visual words, i.e., each word corresponds to a bin, and each bin counts the number of appearances of the corresponding word within the video frame. The classification of these fixed-size histograms can then be performed by supervised learning algorithms, such as support vector machines (SVMs) (Theodoridis and Koutroumbas 2008).

In this paper we investigate various multi-label classification methods, aiming to a richer semantic interpretation of WCE video frames. We present a saliency-enabled BoW-based methodology, as well as a convolutional neural network (CNN) architecture that provides multi-scale feature extraction for multi-label classification. The proposed multi-label classification approach considers that both normal and abnormal video frames, may include normal contents belonging to various semantic categories (as described earlier in this section). Considering that the video frame features extracted from these contents are usually different (e.g., bubbles include white reflections, debris has green/yellow hues) the proposed approach identifies them as members of separate classes, aiming to simplify the detection of abnormalities. Thus, for each video frame a more complete description is achieved using multiple semantic identifiers (labels). To the best of our knowledge, this is the first study evaluating semantic interpretation of endoscopy video frames using multi-label classification techniques. The only relevant previous work was performed with a significantly smaller dataset in a preliminary study (Vasilakakis et al. 2017).

The rest of this paper consists of 6 sections. In Sect. 2 we provide a brief medical background on GI video endoscopy, focusing on WCE. Then, in Sect. 3 we review relevant previous works, while Sect. 4 describes the dataset used in our experimental study. The proposed methodologies towards the multi-label semantic representation of video frames are presented in Sect. 5. The experimental results are presented and discussed in Sect. 6. Finally, the conclusions of our study are summarized in Sect. 7, where next steps for future research are also suggested.

2 Gastrointestinal video endoscopy

Gastrointestinal (GI) endoscopy is an ever-expanding group of well-established as well as developing techniques that allow physicians to inspect the GI tract. Traditional GI endoscopy is performed with flexible (conventional) endoscopes equipped with a charged-couple device (CCD) camera. However, there are segments of the GI tract such as the small-bowel that are not easily reached by conventional endoscopes due to length limitations. To overcome this, WCE has been proposed. It utilizes a swallowable device having the size of a large vitamin pill: the capsule endoscope (CE), which is equipped with a color video camera and a light source. The majority of commercially-available CEs are passive devices, in the sense that they move "naturally", i.e., by exploiting gravity as well as the peristaltic motion of the GI tract. During its 12-h long journey through the GI tract, a CE continuously captures thousands of color video frames, which are wirelessly transmitted to an external recorder (Koulaouzidis et al. 2015).

Although WCE provides wealth of information, since CEs capture approximately. 100K (on average) color frames, this also consists its major limitation (Iakovidis and Koulaouzidis et al. 2015). This occurs since the set of extracted frames needs to be manually reviewed by WCE readers, using specialized software. This process may range from 45 min up to a couple of hours, while their attention needs to be constantly undistracted. Often, non-medical clinical personnel, including nurses and clinical scientists, come to the rescue since clinicians' time is expensive and in high demand (Yung et al. 2017; Riphaus et al. 2009). Nonetheless and no matter who the reviewer is, WCE reading is a tiring, time-consuming procedure, hence prone to human errors. For example, it has been shown that the detection rate of clinically-significant findings by experts is limited to approximately 40% (Zheng et al. 2012). Thus, approaches offering fully-automated lesion detection are highly desirable (Iakovidis et al. 2014b). Such approaches have been previously applied in the context of known or suspected inflammatory bowel disease (IBD) (Koulaouzidis et al. 2013) with main aim to recognize typical abnormalities such as ulcers, aphthae etc.

On the matter of automatic lesion detection, several computer vision-based approaches have been proposed (Iakovidis and Koulaouzidis 2015) and are often based on supervised approaches. In these, experts annotate abnormal areas within the video frames while the rest of the frame is automatically classified as normal. Then, learning algorithms are applied on each category of abnormalities. Previously, we have presented a public, open-access database, namely KID which provides high-quality video frames annotated by medical experts (Koulaouzidis et al. 2017), and has been used for the evaluation of some previous works (Iakovidis and Koulaouzidis 2014).

3 Related work

The work related to the saliency-enabled BoW, and CNN-based methodologies presented in this paper includes: (a) salient point detection techniques; (b) weakly-supervised classification methods; and, (c) multi-label classification methods.

3.1 Salient point detection

One of the challenges within the problem of semantic description of WCE videos is the lack of standardized interpretation methods. Therefore, many research works begin by constructing a saliency map. Given such a map, they are then able to select a subset of a given image/video frame, i.e., regions that would be examined for potential existence of abnormalities. For example, Yuan et al. (2016b), proposed a saliency map extraction method for the detection of bleeding frames in WCE videos, by creating two saliency maps and by fusing color information of the *a* and the M channel of the CIE*Lab* and the CMYK color spaces, respectively, as well as heuristic properties of the "reddish" colors. Superpixel-based segmentation (Achanta et al. 2012) has been investigated by several research efforts among others,also for the detection of bleeding regions. For example, Fu et al. (2014) and Shi et al. (2015), extracted features and classified superpixels accordingly. On the other hand, Iakovidis et al. (2015) proposed an approach for selecting salient superpixels using hand-crafted features from the *a* channel of CIE*Lab*. Yuan et al. (2015) fused saliency maps that had been extracted based on multi-level superpixel representations, to detect ulcer. Several other approaches concentrated on the saliency detection of polyps. Yuan et al. (2017b) fused contrast-based and object center-based saliency maps and used strong classifiers. Also, Bernal et al. (2015) proposed the use of energy maps that indicated the likelihood of polyp presence. The problem of detection of multiple abnormalities within the same image/video frame has recently gained the attention of the research community. Yuan et al. (2017a) aimed to detect bleeding, polyp, ulcer and normal video frames. To this goal they calculated color SIFT features from each semantic category, separately extracted visual words from each and combined all words to obtain a visual dictionary. Video frames were also encoded by a novel adaptive saliency algorithm.

3.2 Weakly-supervised classification

Recently a preliminary study with a weakly-supervised CNN was performed (Georgakopoulos et al. 2016), to detect inflammatory lesions. Also, several weakly-supervised CNN-based approaches in the context of WCE have been proposed. Seguí et al. (2016) fed the CNN with RGB raw data along with their Hessian and Laplacian transformations. Chen et al. 2017 proposed a cascaded CNN scheme for recognizing the organs of the GI tract. Jia and Meng (2018) replaced the second fully connected layer of a CNN with an SVM to detect blood. Feature extraction was performed by a CNN to detect polyps in the study of Zhang et al. (2017). The network was pre-trained using non-medical video frames and an SVM was used for classification. Another approach towards weakly supervised learning is the BoW model. BoW has been shown to be an effective strategy to cope with the demand for annotated GIE video frames (Vasilakakis et al. 2016; Yuan et al. 2016a, b; Wang et al. 2015, 2016b). The vocabulary used for image/video frame representation is typically based on hand-crafted features. Color histograms have been used for bleeding detection (Yuan et al. 2016b), CIE-*Lab* features have been used in a previous work for inflammatory lesion detection (Vasilakakis et al. 2016), SIFT combined with complete LBP has been used for polyp detection (Yuan et al. 2016a), while several color and LBP histograms have been combined for gastric and oesophageal cancer, gastritis, and oesophagitis (Wang et al. 2016b).

3.3 Multi-label classification

Following the BoW feature extraction process, the content of the WCE video frames needs to be classified into semantic categories. Usually the classification of the endoscopic video frame content is performed into two categories, corresponding to normal and abnormal tissues, using binary classifiers, e.g., SVMs (Theodoridis and Koutroumbas 2008). However such approaches only provide an abstract categorization of the video frame content. This happens due to the initial assumption that every video frame belongs in exactly one of the aforementioned categories. This assumption does not consider the multiple semantics that may co-occur within a given video frame. For example, the semantics of a normal video frame besides mucosal tissues may include normal intestinal content such as bubbles and debris, and the lumen hole.

Let $Q = R^d$ be the feature space derived from a set of images used to train the supervised classification system. Also, let L denote the label space. In the binary case the classification system aims to learn a function $f : Q \rightarrow L$, by using the feature vectors $q_i \in Q, i = 1, 2, \dots, N$, that have been extracted from a set of N images and are labeled by l_j, $\in L$, $j = 1, 2$ from a training set $\{(q_i, l_i \vee 1 \leqslant i \leqslant N, \leqslant j \leqslant 2)\}$. In weakly supervised learning each label l_j refers to the semantic content of the whole image, since a given feature vector q_i describes the whole image rather than a specific region.

When tackling the problem of multiple-label classification, one may use a cascade of binary or multi-class classifiers on image regions (Georgakopoulos et al. 2016). In the latter case, an image is labeled with a single label $l_j \in L$, $j = 1, 2, \dots, m$; m denoting the number of the available classes to describe image content. However, such an approach does not take into account that the visual content of a single image may be described with more than one, different labels. Therefore, taking this observation into account, in this paper we propose a multi-label classification approach (Tsoumakas and Katakis 2007; Zhang and Zhou 2014).

More specifically, let v be a vector of m multiple labels for each $q_i \in Q, i = 1, 2, \dots, N$, where $v_j \in L, v_j = \left(l_1, l_2, \dots l_m\right)$, $j = 1, 2, \dots, z$. Each label is a binary flag denoting the presence of different kinds of image contents. In this paper a total of 5 labels are considered, indicating the presence of normal (l_1), abnormal (l_2), debris (l_3), bubbles (l_4), and lumen hole (l_5). The purpose of training a multi-label classifier is to learn a function $h : Q \rightarrow 2^L$.

There are two main learning strategies, namely the algorithm adaptation, and the problem transformation strategies (Tsoumakas and Katakis 2007). Algorithm adaptation tackles multiple labels by adapting existing learning algorithms from single- to multi-label. Examples of algorithms implementing this strategy include an adaptation of the k-nearest neighbor (k-NN) (Theodoridis and Koutroumbas 2008) a classifier for multi-label classification (MLkNN) (Zhang and Zhou 2007) and kernel methods, e.g., multi-label SVMs (Elisseeff and Weston 2001).

On the other hand, the problem transformation strategy deals with multi-label learning problem by reducing it into binary or multi-class categorization. This way, a traditional classifier, e.g., an SVM may be used. In this work we investigate the multi-label classification in the context of endoscopic video frame analysis using various problem transformation methods (Tsoumakas and Katakis 2007). More specifically, we use the binary relevance (Tsoumakas and Katakis 2007), the ranking and thresholding (Tsoumakas and Katakis 2007; Fürnkranz et al. 2008), the pairwise classification (Fürnkranz et al. 2008; Mencia and Furnkranz 2008) and the label combination (Read et al. 2008) methods.

The binary relevance method (Tsoumakas and Katakis 2007) trains different binary classifiers, each of which classifies the video frames according to a single label. In the context of the endoscopic video frame classification investigated in this paper, five binary classifiers are used to determine the existence of each of the five categories of content considered, e.g., the existence of abnormalities or not, the existence of debris or not, etc. However, this methodology does not consider possible dependencies between labels.

The label combination method (Read et al. 2008) transforms the task of multi-label learning into a standard, single-label, multi-class classification. It considers each different set of labels that exist in the multi-label data set as a single one. In this way, it treats every label combination in the training data as a unique class label in a binary label problem. Apart from the five classes that were referred earlier, there are also "classes" that derive from their combinations, e.g., a "new" class may be considered to be the set of video frames are labeled both as normal and debris. This artificial class is then denoted as the normal-debris class.

Ranking and thresholding methods (Tsoumakas and Katakis 2007; Fürnkranz et al. 2008) aim to transform the task of multi-label learning into a multi-class problem. In ranking the task is to order the set of labels. A threshold function is constructed from multi-label data, so that the topmost labels are more related with the new instance. A ranking of labels requires post-processing in order to give a set of labels, which is the proper output of a multi-label classifier.

The pairwise classification (Fürnkranz et al. 2008; Mencia and Furnkranz 2008) adopts the “one-*vs*-one” approach, where one classifier is associated with each pair of labels. This is contrary to the binary relevance approach of “one-*vs*-all” where one classifier is associated with the relevance of each label. Hence, instead of five binary problems, ten binary problems are formed, because there exist ten different pairs of labels. Typically, each pairwise problem is constructed from examples with which either labels (but not both) are associated, thus forming a decision boundary for these two labels.

Artificial neural networks (ANNs) have been traditionally used in single-label binary classification tasks. Zhang and Zhou (2006) proposed an adaptation of the error function of a back-propagation learning algorithm for multi-layer perceptron (MLP) architecture so as to account for multiple labels in the learning process. Within the last decade, deep learning and more specifically CNNs (LeCun et al. 1990) have shown high predictive capacity in the broader field of computer vision (Krizhevsky et al. 2012; Simonyan and Zisserman 2014; Szegedy et al. 2015) on single label datasets, such as ImageNet (Deng et al. 2009) and CIFAR (Krizhevsky 2009). CNNs have also been used in endoscopic single-label image classification. Georgakopoulos et al. (2016), developed a network capable of detecting inflammatory conditions in gastrointestinal images. Sekuboyina et al. (2017) also developed a CNN architecture for abnormality detection in WCE images, exploiting the importance of color information in the classification task. Another approach of using a CNN in a classification task is through a pre-trained network used for feature extraction exploiting the convolution layer feature maps. Zhang et al. (2017) utilized a pre-trained network, and more specifically CaffeNet (Jia et al. 2014), to transfer learning by extracting features from intermediate convolution layers of the network and then use them to train an SVM classifier.

Recently the usage of CNNs has been extended into multi-label classification problems. A common method to extend a CNN to multi-label classification is to transform it into multiple single-label classification problems by using one output neuron per label, as e.g., in the work of Gong et al. (2013), who explored various multi-label loss functions on a network similar to the one proposed by Krizhevsky et al. (2012). While in a typical multi-class classification problem a common practice is to use softmax activations on the output neurons, in multi-label classification problems this does not apply since the softmax function forces the output neurons to express the selected class as a probability, which depends on the rest of the classes. In multi-label classification the usage of sigmoid neurons is typically employed with a cross-entropy loss (Guillaumin et al. 2009), which expresses the probability of a given class as a Bernoulli distribution. Other approaches have also been proposed, such as the one of Wang et al. (2016a) which utilized a combination of a recurrent neural network (RNN) and a CNN to leverage the label dependencies that exist in natural images.

4 Weakly-labeled endoscopy dataset

For the evaluation of the methods investigated in this study, we use “Dataset 2” of the publicly available KID[1] database (Koulaouzidis et al. 2017). This dataset is composed of WCE video frames obtained from the whole GI tract using a MiroCam capsule endoscope with a resolution of 360 × 360 pixels. Abnormalities depicted within this dataset include 303 vascular (small bowel angiectasias and blood in the lumen), 44 polypoid (lymphoid nodular hyperplasia, lymphoma, Peutz–Jeghers polyps), 227 inflammatory (ulcers, aphthae, mucosal breaks with surrounding erythema, cobblestone mucosa, luminal stenoses and/or fibrotic strictures, and mucosal/villous oedema) lesion video frames, and 1778 normal video frames obtained from the esophagus, the stomach, the small bowel and the colon. In the context of this work, all frames were weakly annotated with the labels that indicate the presence of abnormal (l_1), debris (l_2), bubbles (l_3), and lumen hole (l_4). Figure 1 illustrates representative endoscopic frames for each one of the aforementioned labels.

The annotation in the context of this paper is weak; the content of a given image is globally annotated rather than in detailed pixel-level. The latter can be performed using existing state-of-the-art software tools, such as Ratsnake (Iakovidis et al. 2014a). However, video frame annotation using solely semantic labels, although it is significantly easier than the pixel-wise annotation process, can also become very time consuming.

In order to speed up such multi-label annotation tasks a novel software was developed, namely RATStream (Rapid Annotation Tool for video frame and video Streams), which enables time-efficient annotation of both video frames and video streams (Vasilakakis et al. 2017).

[1] KID dataset: https://is-innovation.eu/kid/.

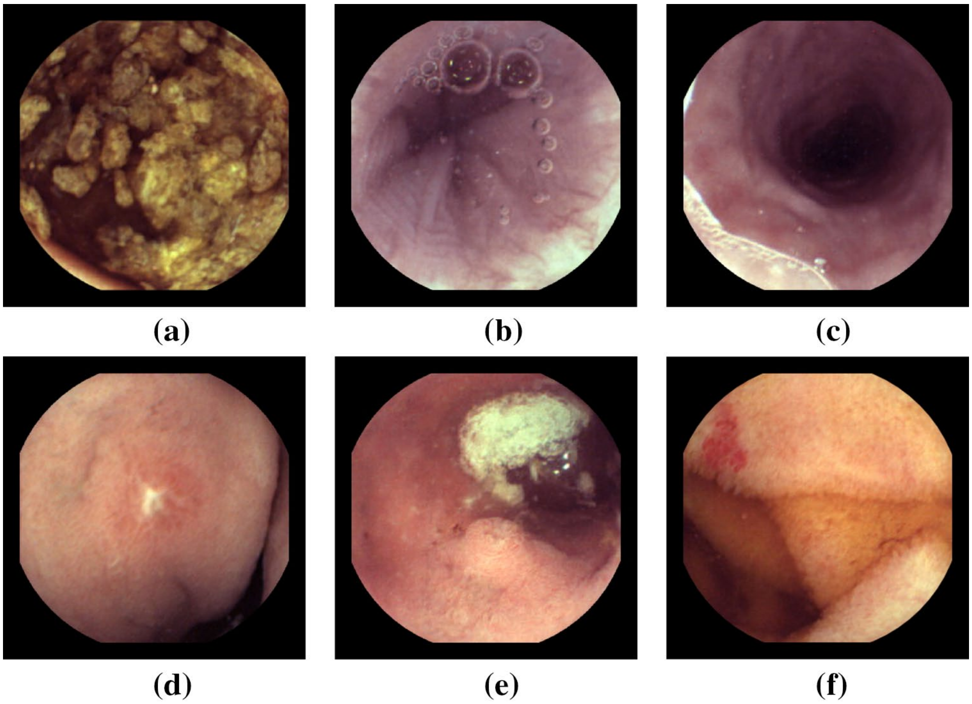


Fig. 1 Representative images from KID dataset. **a** Debris. **b** Bubbles. **c** Lumen hole. **d** Inflammation. **e** Polypoid. **f** Angiectasia

5 Weakly-supervised multi-label classification

Contrary to conventional supervised learning, weakly-supervised learning (Hoai et al. 2014) does not require explicit and detailed annotation. Instead, only video frame-level annotation of the semantics of the video frame is required. Thus, a given video frame may be annotated, e.g., with the semantic concept "abnormal", if it contains an abnormality, or with the semantic concept "normal", if it does not contain an abnormality. At a next level, further semantic concepts may be also added to the annotation process. In the context of lesion detection in WCE, different "normal" concepts can be associated with different normal intestinal content, e.g., "debris", "bubbles", whereas different "abnormal" concepts may include GI lesions, e.g., "inflammatory lesions", "vascular lesions", etc.

Weakly-supervised methods are often criticized because they do not always provide localization information regarding the detected classes. However, in the context of reviewing of large WCE videos, such an approach could significantly reduce the amount of effort required by the video reviewer, since it detects frames that possibly contain lesions. Since such frames are usually a rather small subset of the entire WCE video, the reviewers' task may be limited to localization of abnormalities within this subset, which is a less tiring task.

5.1 Salient point detection

The BoW approach can be based on a set of extracted patches surrounding dense points that result from a sampling process using a regular grid (i.e., one with equal horizontal and vertical inter-pixel distances). Such dense approaches are often criticized as "naïve" when compared to more sophisticated approaches, such as the SIFT (Lowe 2004) or SURF (Bay et al. 2008) interest points. However, it has been shown that they carry valuable information regarding semantic interpretation of visual content (Tuytelaars 2010). In the context of endoscopic video frame analysis the application of SURF on channel *a* of the CIE-*Lab* color space (*a*-SURF) resulted in salient points on all the abnormalities included in that study. Also, the results of a preliminary study (Vasilakakis et al. 2016) showed that dense sampling may be more time-consuming, but it can result in higher abnormality detection rates.

The dense sampling process using regular grid, extracts a large number of feature vectors. These feature vectors are not easily separable by a clustering algorithm. For this reason, there is a need to select some video frame points to extract fewer feature vectors without significant loss of information. A way to reduce these points in a video frame represented in CIE-*Lab* color space is to get points only from the video frame regions where a significant color change is observed. The purpose behind this idea is to discriminate and sample video frame regions, where a discontinuity in their color

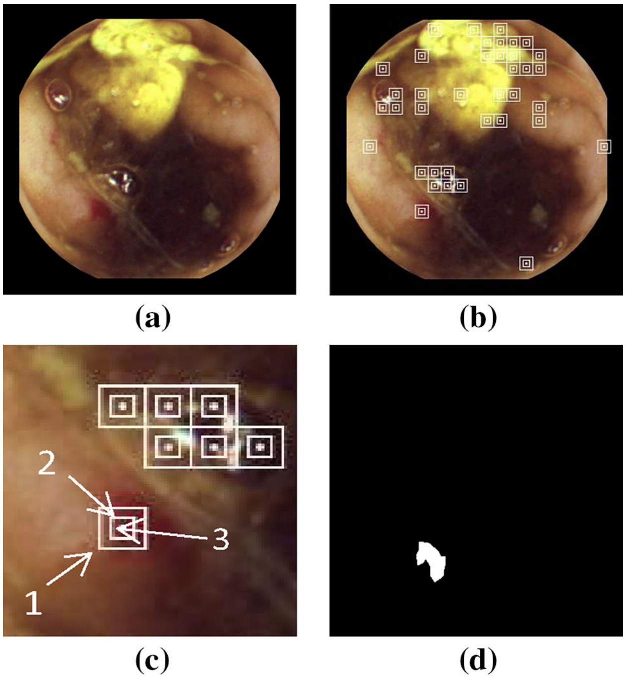


Fig. 2 DoM salient point detection. **a** Original video frame. **b** The remaining points after dense sampling, around which the Euclidean distances are estimated. **c** A magnified region of **b**, clearly indicating the outer window (1), the inner window (2) where the maxima are calculated, and the central point (3) of these windows. **d** Detailed graphical image annotation of **a**

description appears. The discontinuity in color of channel a, indicates the region as a region of interest. In that sense these regions can be considered as "salient" points. In order to detect such salient points, the difference between two maximum values in a-channel around the densely sampled points are considered.

The proposed difference of maxima (DoM) algorithm (Fig. 2) for salient point detection proceeds as follows: (a) Dense sampling of the video frame using window of size X. (b) For each point of the dense sampling grid calculate the maximum $a_{\max}$ and minimum $a_{\min}$ values of channel a from two windows of different sizes centered at each of these points; an outer window of size X, and an inner window of size $X/2$; (c) The central point of these windows (point 3 in Fig. 2c) is included into the set of salient points, if the Euclidean distance between the two vectors $(a_{\max}, a_{\min})_X$ and $(a_{\max}, a_{\min})_{X/2}$ is larger than the mean Euclidean distances estimated from all windows in the same video frame; otherwise, it is rejected.

5.2 BoW-SVM multi-label classification scheme

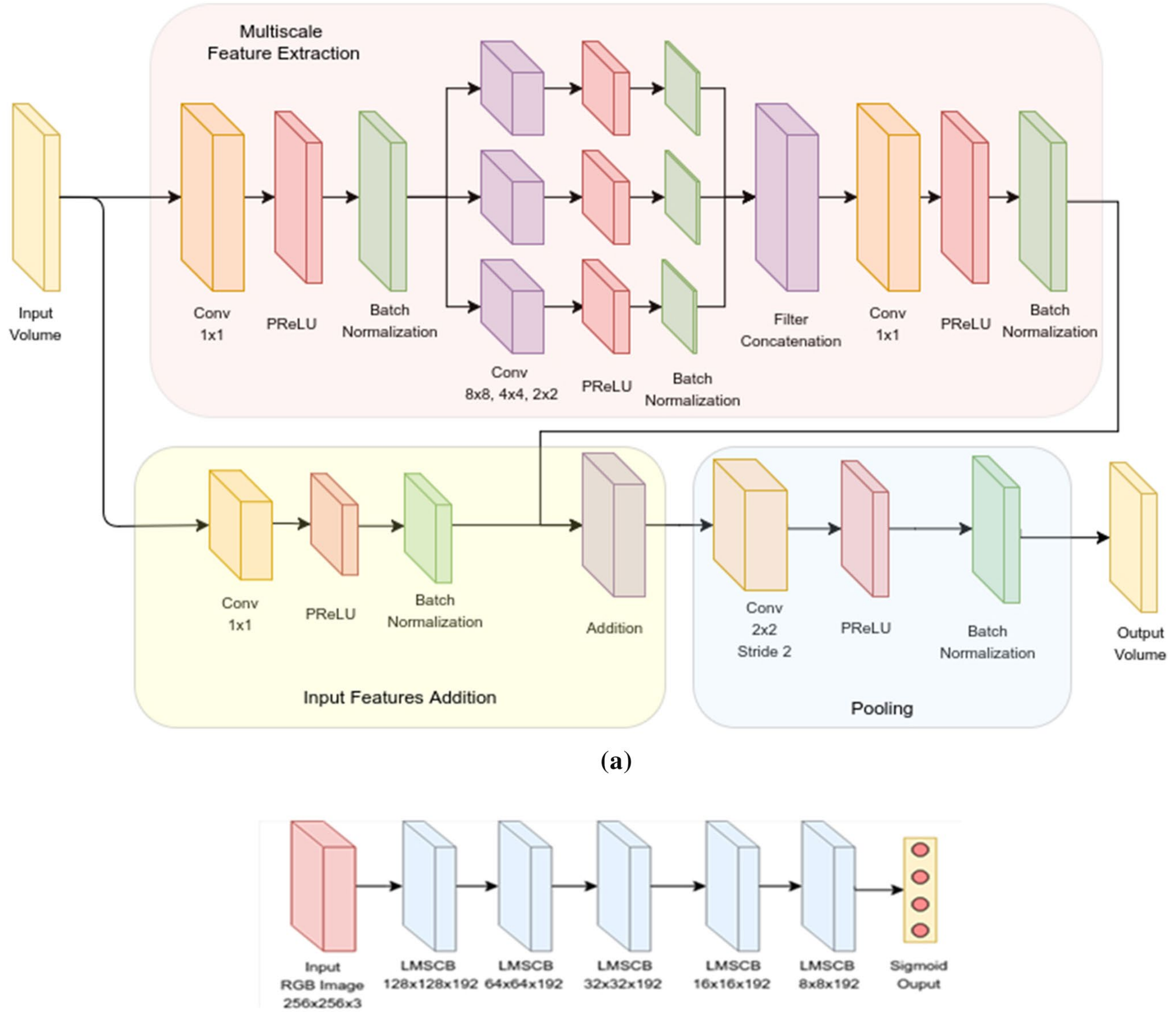
According to the BoW technique, a given image/video frame F_i is described as a set of visual "words" $\{w_i\}, i = 1, \ldots, N$ that originate from a visual "vocabulary." The latter comprises of a set of representative quantized low-level feature descriptions of a training video frame set. These descriptions derive from parts of video frames such as overlaid grids, regions (e.g., resulting upon a segmentation process) or patches (e.g., surrounding salient points). To this end, typical approaches adopt a clustering algorithm and the vocabulary consists of either the centroids or the medoids resulting from this process. These comprise the set of the visual words and accordingly a video frame is described (coded) using a histogram H_i calculated on these words. Put differently, each feature vector captures the frequencies of all words of the visual dictionary within the corresponding video frame. Another advantage of the BoW approach is that it results to a fixed-size video frame description (i.e., equal to N) which is appropriate to be used with typical machine learning approaches.

In this work, the adopted BoW approach is based on a set of extracted patches surrounding points that result from a salient point detection process. For the low-level description of the video frame patches, we choose to adopt a set of color-based features which has been previously applied to the problem of lesion detection and yielded superior results compared to the state-of-the-art approaches (Iakovidis and Koulaouzidis 2014). More specifically, a given video frame is first transformed to the CIE-*Lab* color space and then, from each patch we extract the *Lab* values of the salient point, and the max and min values of all three components within the entire patch. This way, a 9-dimensional color feature vector is extracted from each patch.

5.3 Multi-label multi-scale CNN classification scheme

CNNs have been widely used in computer vision, as their bio-inspired nature of their neuron arrangement enables them to capture both local and global features from the data. In this work we propose a novel fully convolutional neural network (FCNN), which is capable of extracting multi-scale features from an input image/video frame and combine them to perform multi-label classification in a weakly-supervised way.

The proposed multi-scale and multi-label CNN (MM-CNN) consists of five blocks (Fig. 3), which will be referred to as large-medium-small convolution blocks (LMSCBs). Each LMSCB performs multi-scale feature extraction and input volume enrichment. It consists of three components. The first one performs multi-scale feature extraction using three parallel convolution layers of 2 × 2, 4 × 4 and 8 × 8 filter sizes respectively. The convolution layers receive an input volume with 192 feature maps and outputs an output volume of 64 each feature maps. Each convolution operation is followed by parametric rectified linear units (PReLU) activations (He et al. 2015) along with batch normalization



(a)

(b)

Fig. 3 The proposed MM-CNN architecture, composed of LMSCBs. **a** The architecture of an LMSCB. The input volume is forwarded to the multi-scale feature extraction component and then to the addition operator. The final feature maps are then forwarded to the pooling component which results in a 50% dimensionality reduction. **b** The overall MM-CNN architecture composed of 5 LMSCB modules and 4 sigmoid output neurons, which are used for the multi-label classification

(Ioffe and Szegedy 2015) which prevents model overfitting and facilitates faster convergence. We used PReLU activations instead of traditional ReLU as the former may prevent the neuron saturation problem from which the latter suffer. The usage of batch normalization was employed to assist the training process of the network. The extracted features are then concatenated to form a single output vector which is used as the input of the next component. This approach has been inspired by the GoogLeNet (Szegedy et al. 2015), in which, the inception module, follows a similar multi-scale feature extraction approach. The second component is an addition operator, which adds the original input volume with the output volume of the first component. This forms an enhanced input volume for the following modules which allows the original input to be preserved across the LMSCBs. The last component performs down-sampling of the output volume using a convolution layer with a filter of size 2 × 2 and a stride equal to 2. We opted for this approach instead of the traditional max-pooling operation, in order to simplify the overall model architecture, since it has been shown that max-pooling can be replaced with a convolution layer of appropriate filter size and stride (Springenberg et al. 2014). To provide fine control over the input and output feature map volume of each component, we introduced a 1 × 1 convolution operation. This serves two purposes; it helps to keep the overall number of free-parameters manageable while it allows the addition operation of the second component to be valid.

The overall MM-CNN architecture consists of five layers of LMSCBs, as illustrated in Fig. 3b. In order to perform the task of multi-label classification, the network has 4 output sigmoid neurons; i.e., one for each label. A traditional approach in CNN architecture designs, is to append fully connected layers as the last layers of the network, which act as the classification layers of the model. While this approach has been used in several CNN architectures, such as (Krizhevsky et al. 2012; Simonyan and Zisserman 2014), it leads into a significant increase in the number of free parameters of the network and loses the spatial arrangement of the extracted features of the previous layers. For these reasons we chose to avoid using any fully connected layers in our network architecture. This helps to keep the number of free-parameters of the network lower; thus, reduces the overall complexity of the overall CNN model.

6 Experimental evaluation

Several experiments were performed to evaluate multi-label classification as a means for semantic interpretation of endoscopy video frames. Both the saliency-enabled BoW, and the MM-CNN methodologies were evaluated in comparison to state-of-the-art approaches.

In the case of the saliency-enabled BoW methodology, for each video frame, features have been extracted using the proposed DoM salient point detection method and the "naive" approach of dense feature extraction. For the proposed DoM salient point detection method we used image samples of 24 × 24 pixels. The BoW model was constructed with a visual vocabulary with sizes in the range from 500 to 2000 words using the *k-means* clustering algorithm (Drake and Hamerly 2012). The classification of the feature vectors obtained using the BoW method, was implemented by an SVM classifier. We have tested linear, polynomial and Radial Basis Function (RBF) kernels, and followed the grid-search approach (Chang and Lin 2011) to determine its optimal parameters. The RBF kernel provided the best results, for a minimum cost parameter $c = 10$ and $\gamma = 2^{-8}$.

In the case of the MM-CNN the training of the network was performed using the back-propagation algorithm with a batch size of 32 images and the root mean square propagation (RMSProp) (Hinton et al. 2012) optimizer with learning rate $l = 0.0001$ and fuzz factor $\varepsilon = 1e - 8$. Furthermore video frames from the KID dataset 2 have been cropped to 320 × 320 pixels by removing the excess surrounding black border. Then, they were downsized to a resolution of 256 × 256 pixels. The network has been implemented using the Keras (Chollet 2015) Python library backed by TensorFlow (Abadi et al. 2016) graph framework. It was trained using an NVIDIA GTX-960 enabled graphical processing unit (GPU) with 1024 CUDA (Nickolls et al. 2008) cores having 2 GB of RAM and clock frequency of 1127 MHz. It is worth mentioning that the entire training of the network for each fold took approximately 8 h. The early stopping technique was adopted to optimize the network's generalization performance, using 15% of the data as validation subset. The number of training epochs required per fold was approximately 2000. This could be considered as being relatively low when compared to other networks, e.g., the one of Simonyan and Zisserman (2014). Yet, it happens due to the low number of free-parameters of the overall architecture (Fig. 3).

To compare the classification performance of MM-CNN with the transfer learning approach in multi-label classification of WCE gastrointestinal tract images, we implemented the methodology followed by Zhang et al. (2017). More specifically, for the feature extraction we followed the same procedure as presented by the authors, while for the classification of the extracted features, we followed multi-label "one-*vs*-all" SVM with $c = 2^{-9}$and polynomial kernel. The parameters of the SVM where selected after a series of experiments in order to determine the optimal values for the domain.

The classification performance was thoroughly investigated using receiver operating characteristic (ROC) analysis (Fawcett 2006). These curves illustrate the tradeoff between sensitivity and specificity for various decision thresholds. In order to enable comparisons between the ROC curves, the area under the ROC (AUC) was used as a classification performance measure (Zhang and Zhou 2014), which unlike accuracy, comprises a relatively robust metric for datasets with imbalanced class distributions (Provost and Fawcett 1997). Experiments were performed using the 10-fold cross validation evaluation scheme, using SVMs as a binary classifier. Multi-label classification was implemented using a derivative of WEKA library (Witten et al. 2017) called MEKA (Read et al. 2017).

Initially, we examined the case of binary classification of the video frames into normal and abnormal classes. We investigate the performance of BoW method using the proposed DoM for salient point detection in comparison with the state of the art methodologies, of Yuan et al. (2016a), who used SIFT algorithm (Lowe 2004) for the detection of interest points and a concatenated feature vector of SIFT and LBP, or SIFT and CLBP (Yuan et al. 2016a), for the description of video frame regions. The comparison also includes the method of Vasilakakis et al. (2016), who used SURF algorithm for salient point detection in the a-channel (SURF(a)) of CIE-Lab and the dense BoW approach and the CNN (Zhang et al. 2017). The results with regards to lesion

Table 1 Binary classification results for lesion detection using various weakly supervised BoW methods and CNN with an SVM classifier

Methods	TP	FN	FP	TN	Sensitivity	Specificity	AUC
BoW + SURF(a)	232	342	214	1564	0.40	0.87	0.78
BoW + SIFT + LBP	181	393	210	1568	0.31	0.88	0.72
BoW + SIFT + CLBP	207	367	200	1578	0.36	0.88	0.78
BoW + dense	240	334	196	1582	0.42	0.88	0.8
CNN	299	275	259	1519	0.52	0.85	0.78
BoW + DoM	252	322	176	1602	0.44	0.90	0.81

The confusion matrix (true positives—TP, false negatives—FN, false positives—FP, and true negatives—TN), the sensitivity and the specificity are provided along with the AUC for each method

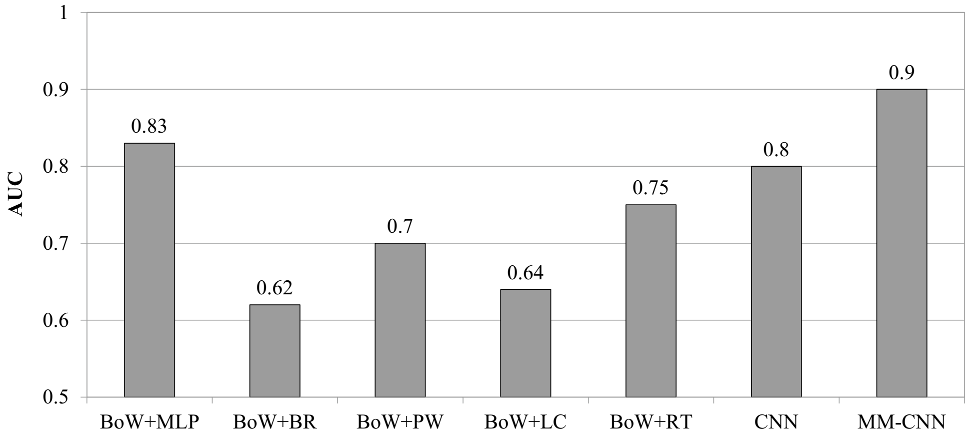


Fig. 4 Comparative multi-label lesion detection results for each multi-label method tested

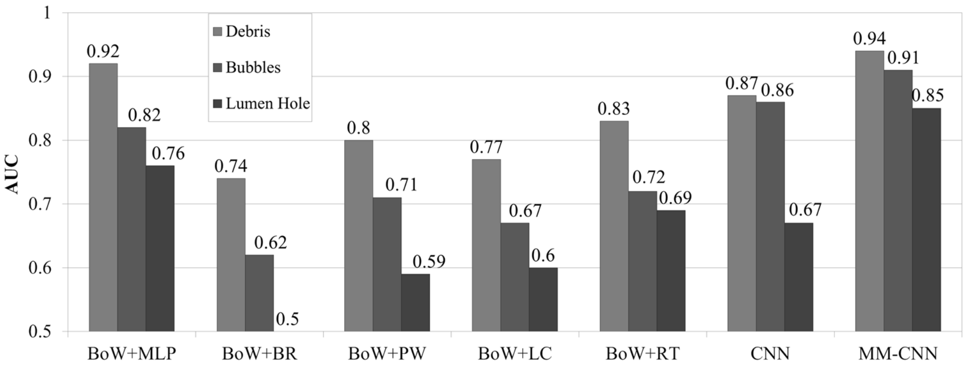


Fig. 5 Classification results for each semantic label in KID dataset 2 for each multi label method

detection are presented in Table 1. It can be noticed that the use of the proposed DoM algorithm increases the binary classification performance to an AUC of 0.81. All methods provide a low sensitivity. This indicates the difficulty of the lesion detection problem. The BoW method using DoM provided significantly higher specificity (less false positives) than all other methods. The higher sensitivity was obtained by CNN, at the cost of a higher false positive rate.

Multi-label classification was performed using the following labels: abnormal, debris, bubbles, and lumen hole. The use of DoM for multi-label classification, results in an even higher classification performance than the conventional binary classification scheme. Best results were obtained using the multi-layer perceptron (MLP) multi-label classification method with 100 hidden layer neurons, a learning rate of 0.1, trained with the features extracted from BoW model. The obtained AUC reached up to 0.83 using a vocabulary of 800 visual words. The results for all weakly methods using BoW features are presented in Fig. 4. The basic methods for multi-label classification, which used, were binary relevance (BR), label combination (LC), ranking and thresholding (RT) and pairwise classification (PC). For all multi label methods we used the same SVM with radial basis function kernel (RBF) with c = 10.

As in the binary classification experiments, these parameters were determined using the afore-mentioned kernels and grid-search approach. Also, Fig. 4 includes the results of CNN (Zhang et al. 2017) for multi-label classification in order to compare our proposed MM-CNN. It can be noticed that MM-CNN provided the highest performance compare to all the other approaches and achieved an AUC equal to 0.90.

The classification results per semantic category are presented in Fig. 5. It can be noticed that the result for debris are significantly higher than the results of bubbles and lumen hole. The reason is that the most video frames in KID dataset 2 had debris as content compare to the number of video frames that had bubbles and/or lumen hole.

It can also be noticed that the classification performance of the CNN is not always better than the BoW-based approaches, although it has been proved effective in the context of endoscopy (Zhang et al. 2017). This could be explained by the diversity of the KID dataset, which includes several different kinds of lesions, whereas the dataset used by Zhang et al. (2017) included only colorectal polyps.

7 Conclusions

We investigated multi-label classification methods for a richer semantic interpretation of endoscopy video frames. The rationale behind this was that the classification of the video frame contents into multiple semantic categories, could simplify the detection of contents corresponding to abnormalities, especially since the presence of intestinal content, such as debris and bubbles, is dominant in parts of the GI tract (Iakovidis and Koulaouzidis 2015). The results validate that the effect of using multiple labels can enhance abnormality detection.

Currently, several supervised methods exist for detection and removal of uninformative frames due to the presence of intestinal (Iakovidis and Koulaouzidis 2015). Drawbacks of such methods include the following: (a) they increase the computational complexity of the overall video analysis task, since the video needs to be processed in two steps (one for the removal of the uninformative frames and another one for the detection of the abnormalities); (b) by totally removing the video frames with the intestinal content there is a chance to also miss frames with abnormalities that are present with the intestinal content. Comparatively, the proposed multi-label classification approach provides information about both the presence of abnormalities and intestinal content in a single step.

Multi-label classification is a well-known classification approach; however, it has not been popular in image analysis mainly because, using conventional supervised methods, it requires a lot of effort in annotating in detail the training images with multiple labels. Motivated by the results of a previous work (Georgakopoulos et al. 2016), where the weakly-supervised approach to classification of GI lesions outperformed the conventional supervised approach, in this paper we introduced multi-label classification in the context of weakly-supervised learning. The results obtained show that by expressing the problem of abnormality detection as a multi-label classification problem can be beneficial.

We presented an extension of the BoW-based weakly supervised method (Vasilakakis et al. 2016) using DoM as an alternative to the conventional salient point detection algorithms. We also proposed MM-CNN, a novel CNN enabling multi-scale feature extraction for multi-label classification. The results obtained by both of these methods are better than those obtained by state-of-the-art methods. The best results were obtained by MM-CNN.

Future research directions include investigation of scalability of multi-label classification, when larger dataset is available for the learning task. Moreover, the effect in classification performance may have, when we choose different feature extraction methods during the stage of clustering in BoW procedure. Another interesting point for future research is the application of the proposed methods for distinguishing different kinds of abnormalities within the same images. However, to validate their effectiveness on this challenge, new datasets, with relevant images, need to be developed. Last but not least, additional experiments need to be done for the investigation of classification performance, when algorithm adaptation methods are used.

Acknowledgements The research presented in this paper was financially supported by the project "Klearchos Koulaouzidis" Grant no. 5151 and the Special Account of Research Grants of the University of Thessaly, Greece.

References

Abadi M, Agarwal et al (2016) Tensorflow: large-scale machine learning on heterogeneous distributed systems. arXiv preprint. arXiv:1603.04467

Achanta R, Shaji A, Smith K, Lucchi A, Fua P, Süsstrunk S (2012) SLIC superpixels compared to state-of-the-art superpixel methods. IEEE Trans Pattern Anal Mach Intell 34:2274–2282. https://doi.org/10.1109/tpami.2012.120

Bay H, Ess A, Tuytelaars T, Van Gool L (2008) Speeded-up robust features (SURF). Comput Vis Image Video Frame Underst 110:346–359. https://doi.org/10.1016/j.cviu.2007.09.014

Bernal J, Sánchez F, Fernández-Esparrach G, Gil D, Rodríguez C, Vilariño F (2015) WM-DOVA maps for accurate polyp highlighting in colonoscopy: validation vs. saliency maps from physicians. Comput Med Imaging Graph 43:99–111

Chang CC, Lin CJ (2011) LIBSVM: a library for support vector machines. ACM Trans Intell Syst Technol (TIST) 2(3):27

Chen H, Wu X, Tao G, Peng Q (2017) Automatic content understanding with cascaded spatial–temporal deep framework for capsule endoscopy videos. Neurocomputing 229:77–87

Chollet F (2015) Keras. GitHub. https://github.com/fchollet/keras

Deng J, Dong W, Socher R et al (2009) Imagenet: a large-scale hierarchical image database. In: Computer vision and pattern recognition. CVPR. IEEE Conference, pp 248–255

Drake J, Hamerly G (2012) Accelerated k-means with adaptive distance bounds. In: 5th NIPS workshop on optimization for machine learning

Elisseeff A, Weston J (2001) A kernel method for multi-labeled classification. NIPS 681–687

Fawcett T (2006) An introduction to ROC analysis. Pattern Recogn Lett 27:861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Fu Y, Zhang W, Mandal M, Meng M (2014) Computer-aided bleeding detection in WCE video. IEEE J Biomed Health Inform 18(2):636–642

Fürnkranz J, Hüllermeier E, Loza Mencía E, Brinker K (2008) Multilabel classification via calibrated label ranking. Mach Learn 73:133–153. https://doi.org/10.1007/s10994-008-5064-8

Georgakopoulos S, Iakovidis D, Vasilakakis M et al (2016) Weakly-supervised convolutional learning for detection of inflammatory gastrointestinal lesions. In: Imaging systems and techniques (IST), IEEE international conference. IEEE, pp 510–514

Gong Y, Jia Y, Leung T et al (2013) Deep convolutional ranking for multilabel image annotation. arXiv preprint. arXiv:1312.4894

Guillaumin M, Mensink T, Verbeek J, Schmid C (2009) Tagprop: discriminative metric learning in nearest neighbor models for image auto-annotation. In: Computer vision, 2009 IEEE 12th international conference, pp 309–316

He K, Zhang X, Ren S, Sun J (2015) Delving deep into rectifiers: surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE international conference on computer vision, pp 1026–1034

Hinton GE, Srivastava N, Swersky K (2012) Lecture 6a—overview of mini-batch gradient descent. In: Neural networks for machine learning, pp 31

Hoai M, Torresani L, De la Torre F, Rother C (2014) Learning discriminative localization from weakly labeled data. Pattern Recogn 47:1523–1534. https://doi.org/10.1016/j.patcog.2013.09.028

Iakovidis D, Koulaouzidis A (2014) Automatic lesion detection in capsule endoscopy based on color saliency: closer to an essential adjunct for reviewing software. Gastrointest Endosc 80:877–883. https://doi.org/10.1016/j.gie.2014.06.026

Iakovidis D, Koulaouzidis A (2015) Software for enhanced video capsule endoscopy: challenges for essential progress. Nat Rev Gastroenterol Hepatol 12:172–186. https://doi.org/10.1038/nrgastro.2015.13

Iakovidis D, Goudas T, Smailis C, Maglogiannis I (2014a) Ratsnake: a versatile image video frame annotation tool with application to computer-aided diagnosis. Sci World J 2014:1–12. https://doi.org/10.1155/2014/286856

Iakovidis D, Sarmiento R, Silva J, Histace A, Romain O, Koulaouzidis A, Dehollain C, Pinna A, Granado B, Dray X (2014b) Towards intelligent capsules for robust wireless endoscopic imaging of the gut. In: Imaging systems and techniques, IEEE international conference. IEEE, pp 95–100

Iakovidis D, Chatzis D, Chrysanthopoulos P, Koulaouzidis A (2015) Blood detection in wireless capsule endoscope images based on salient superpixels. In: Annual international conference of the IEEE Engineering in Medicine and Biology Society, EMBS, pp 731–734

Ioffe S, Szegedy C (2015) Batch normalization: accelerating deep network training by reducing internal covariate shift. In: International conference on machine learning, pp 448–456

Jia X, Meng M (2018) A deep convolutional neural network for bleeding detection in wireless capsule endoscopy images. In: 38th annual international conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp 639–642

Jia Y, Shelhamer E, Donahue J (2014) Caffe: convolutional architecture for fast feature embedding. In: Proceedings of the 22nd ACM international conference on multimedia, pp 675–678

Koulaouzidis A, Rondonotti E, Karargyris A (2013) Small-bowel capsule endoscopy: a ten-point contemporary review. World J Gastroenterol 19(24):3726–3746. 6

Koulaouzidis A, Iakovidis DK, Karargyris A, Rondonotti E (2015) Wireless endoscopy in 2020: will it still be a capsule? World J Gastroenterol 21(17):5119–5130

Koulaouzidis A, Iakovidis DK, Yung DE, Rondonotti E, Kopylov U, Plevris JN, Toth E, Eliakim A, Johansson GW, Marlicz W et al (2017) KID project: an internet-based digital video atlas of capsule endoscopy for research purposes. Endosc Int Open 5(06):E477–E483

Krizhevsky A (2009) Learning multiple layers of features from tiny images. Technical Report, Computer Science Department, University of Toronto. https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf

Krizhevsky A, Sutskever I, Hinton G (2012) Imagenet classification with deep convolutional neural networks. In: Proceedings of the 25th international conference on neural information processing systems (NIPS), Lake Tahoe, Nevada, vol 1, pp 1097–1105

Le Cun Y, Boser B, Denker J et al (1990) Handwritten digit recognition with a back-propagation network. In: Advances in neural information processing systems, pp 396–404

Li H, Liu L, Sun F et al (2016) Multi-level feature representations for video semantic concept detection. Neurocomputing 172:64–70. https://doi.org/10.1016/j.neucom.2014.09.096

Lowe D (2004) Distinctive image video frame features from scale-invariant keypoints. Int J Comput Vision 60:91–110. https://doi.org/10.1023/b:visi.0000029664.99615.94

Mencia E, Furnkranz J (2008) Pairwise learning of multilabel classifications with perceptrons. In: Neural networks, 2008. IJCNN 2008 (IEEE world congress on computational intelligence). IEEE international joint conference. IEEE, pp 2899–2906

Nickolls J, Buck I, Garland M, Skadron K (2008) Scalable parallel programming with CUDA. Queue 6:40. https://doi.org/10.1145/1365490.1365500

Provost F, Fawcett T (1997) Analysis and visualization of classifier performance: comparison under imprecise class and cost distributions. In: Proceedings of the third international conference on knowledge discovery and data mining (KDD'97), pp 43–48

Read J, Pfahringer B, Holmes G (2008) Multi-label classification using ensembles of pruned sets. Paper presented at the proceedings—IEEE international conference on data mining, ICDM, pp 995–1000

Read J, Reutemann P, Pfahringer B, Holmes G (2017) MEKA: a multi-label/multi-target extension to WEKA. J Mach Learn Res 17:1–5

Riphaus A, Richter S, Vonderach M, Wehrmann T (2009) Capsule endoscopy interpretation by an endoscopy nurse—a comparative trial. Zeitschrift für Gastroenterologie 47:273–276. https://doi.org/10.1055/s-2008-1027822

Seguí S, Drozdzal M, Pascual G, Radeva P, Malagelada C, Azpiroz F, Vitrià J (2016) Generic feature learning for wireless capsule endoscopy analysis. Comput Biol Med 79:163–172

Sekuboyina A, Devarakonda S, Seelamantula C (2017) A convolutional neural network approach for abnormality detection in wireless capsule endoscopy. In: Biomedical imaging (ISBI 2017). IEEE 14th international symposium, pp 1057–1060

Shi W, Chen J, Chen H, Peng Q, Gan T (2015) Bleeding fragment localization using time domain information for WCE videos. In: 2015 8th international conference on biomedical engineering and informatics, BMEI, pp 73–78

Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint. arXiv:1409.1556

Smeulders A, Worring M, Santini S et al (2000) Content-based imagevideo frame retrieval at the end of the early years. IEEE Trans Pattern Anal Mach Intell 22:1349–1380. https://doi.org/10.1109/34.895972

Springenberg J, Dosovitskiy A, Brox T, Riedmiller M (2014) Striving for simplicity: the all convolutional net. arXiv preprint. arXiv:1412.6806

Szegedy C, Liu W, Jia Y et al (2015) Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1–9

Theodoridis S, Koutroumbas K (2008) Pattern recognition. Elsevier/Academic Press, Amsterdam

Tsoumakas G, Katakis I (2007) Multi-label classification. Int J Data Wareh Min 3:1–13. https://doi.org/10.4018/jdwm.2007070101

Tuytelaars T (2010) Dense interest points. In: Computer vision and pattern recognition (CVPR). IEEE conference, pp 2281–2288

Vasilakakis M, Iakovidis DK, Spyrou E, Koulaouzidis A (2016) Weakly-supervised lesion detection in video capsule endoscopy based on a bag-of-colour features model. In: International workshop on computer-assisted and robotic endoscopy, pp 96–103

Vasilakakis M, Iakovidis D, Spyrou E et al (2017) Beyond lesion detection: towards semantic interpretation of endoscopy videos. In: International conference on engineering applications of neural networks. Springer, Cham, pp 379–390

Wang S, Cong Y, Fan H, Yang Y, Tang Y, Zhao H (2015) Computer aided endoscope diagnosis via weakly labeled data mining. In: Image processing (ICIP). IEEE international conference, pp 3072–3076

Wang J, Yang Y, Mao J et al (2016a) Cnn-rnn: a unified framework for multi-label image classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2285–2294

Wang S, Cong Y, Fan H, Liu L, Li X, Yang Y, Tang Y, Zhao H, Yu H (2016b) Computer-aided endoscopic diagnosis without human-specific labeling. IEEE Trans Biomed Eng 63(11):2347–2358

Witten I, Frank E, Hall M, Pal C (2017) Data mining, 1st edn. Morgan Kaufmann, Amsterdam

Yu L, Yuen P, Lai J (2012) Ulcer detection in wireless capsule endoscopy images. In: 21st international conference on pattern recognition (ICPR). IEEE, pp 45–48

Yuan Y, Wang J, Li B, Meng M (2015) Saliency based ulcer detection for wireless capsule endoscopy diagnosis. IEEE Trans Med Imaging 34(10):2046–2057

Yuan Y, Li B, Meng M (2016a) Improved bag of feature for automatic polyp detection in wireless capsule endoscopy images video frames. IEEE Trans Autom Sci Eng 13:529–535. https://doi.org/10.1109/tase.2015.2395429

Yuan Y, Li B, Meng M (2016b) Bleeding frame and region detection in the wireless capsule endoscopy video. IEEE J Biomed Health Inform 20(2):624–630

Yuan Y, Li B, Meng M (2017a) WCE abnormality detection based on saliency and adaptive locality-constrained linear coding. IEEE Trans Autom Sci Eng 14(1):149–159

Yuan Y, Li D, Meng MQH (2017b) Automatic polyp detection via a novel unified bottom-up and top-down saliency approach. IEEE J Biomed Health Inform. https://doi.org/10.1109/JBHI.2017.2734329

Yung D, Fernandez-Urien I, Douglas S, Plevris J, Sidhu R, McAlindon M, Panter S, Koulaouzidis A (2017) Systematic review and meta-analysis of the performance of nurses in small bowel capsule endoscopy reading. United Eur Gastroenterol J. https://doi.org/10.1177/2050640616687232

Zhang M, Zhou Z (2006) Multilabel neural networks with applications to functional genomics and text categorization. IEEE Trans Knowl Data Eng 18:1338–1351. https://doi.org/10.1109/tkde.2006.162

Zhang M, Zhou Z (2007) ML-KNN: a lazy learning approach to multi-label learning. Pattern Recogn 40:2038–2048. https://doi.org/10.1016/j.patcog.2006.12.019

Zhang M, Zhou Z (2014) A review on multi-label learning algorithms. IEEE Trans Knowl Data Eng 26:1819–1837. https://doi.org/10.1109/tkde.2013.39

Zhang R, Zheng Y, Mak T, Yu R, Wong S, Lau J, Poon C (2017) Automatic detection and classification of colorectal polyps by transferring low-level CNN features from nonmedical domain. IEEE J Biomed Health Inform 21(1):41–47

Zheng Y, Hawkins L, Wolff J, Goloubeva O, Goldberg E (2012) Detection of lesions during capsule endoscopy: physician performance is disappointing. Am J Gastroenterol 107:554–560. https://doi.org/10.1038/ajg.2011.46

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.