Weakly-Supervised Lesion Detection in Video Capsule Endoscopy Based on a Bag-of-Colour Features Model

Michael Vasilakakis[1(✉)], Dimitrios K. Iakovidis[1], Evaggelos Spyrou[2], and Anastasios Koulaouzidis[3]

[1] Department of Computer Science and Biomedical Informatics, University of Thessaly, Lamia, Greece
vasilaka.inf@gmail.com, dimitris.iakovidis@ieee.org

[2] National Center for Scientific Research - Demokritos, Institute of Informatics and Telecommunications, Athens, Greece
espyrou@iit.demokritos.gr

[3] Endoscopy Unit, The Royal Infirmary of Edinburgh, Edinburgh, UK
akoulaouzidis@hotmail.com

Abstract. Robotic video capsule endoscopy (VCE) is a rapidly evolving medical imaging technology enabling more thorough examination and treatment of the gastrointestinal tract than conventional endoscopy technologies. Despite of the technological advances in this field, the reviewing of the large VCE image sequences remains manual and challenges experts' diagnostic capabilities. Video reviewing systems for automated lesion detection are still under investigation. Most of these systems are based on supervised machine learning algorithms, which require a training set of images, manually annotated by the experts to indicate which pixels correspond to lesions. In this paper, we investigate a weakly-supervised approach for lesion detection, which requires image-level instead of pixel-level annotations for training. Such an approach offers a considerable advantage with respect to the efficiency of the annotation process. It is based on state-of-the-art colour features, which, in this study, are extended according to the bag-of-visual-words model. The area under receiver operating characteristic achieved, reaches 81%.

Keywords: Video capsule endoscopy · Lesion detection · Colour features · Bag-of-Words · Weakly-supervised learning

1 Introduction

Video capsule endoscopy (VCE) enables the examination of the whole gastrointestinal (GI) tract in a non-invasive way. It is performed with a swallowable capsule endoscope (CE), which captures colour images during its approx. 12 h battery lifetime. Today's commercial CEs are passive, in the sense that they are moving by exploiting both the gravity and the peristaltic motion of the GI tract. However, several research prototypes have been proposed for active, robotic capsule endoscopy, which will enable thorougher examinations, easier lesion localization, and drug infusion [1].

T. Peters et al. (Eds.): CARE 2016, LNCS 10170, pp. 96–103, 2017.
DOI: 10.1007/978-3-319-54057-3_9

A major issue that is still unresolved, both in passive and active VCE is that it requires a lot of human effort for manually reviewing of the produced videos. Typically, each individual review lasts 45–90 min, during which, the reviewer's concentration should remain undivided for a careful inspection of the output video [2]. Such a tiring procedure is prone to human errors; a fact with serious consequences in the diagnostic yield, which is alarmingly low [3].

In order to cope with this problem, automated lesion detection methods based on computer vision algorithms have been proposed [4]. Most of these methods exploit supervised machine learning methodologies, capable of learning what is defined as normal and what is defined as an abnormal finding within the VCE video. The generation of datasets for training the learning machines requires that experts indicate which pixels correspond to normal or abnormal tissues within the VCE images. Considering that the videos produced by a VCE examination are composed of thousands of frames (usually of the order of 10^4), such a pixel-wise annotation task can prove very time-consuming and discouraging for annotation of large datasets by the experts.

A promising solution that could alleviate this problem is weakly-supervised learning, which involves training of a learning machine using weakly annotated data [5, 6]. In this paper weakly supervised learning is considered using images annotated at image-level instead of pixel-level. This way, a binary semantic label is assigned per video frame indicating whether its content is normal or abnormal. A drawback of such an approach is that the abnormal images can be tracked, but the localization of the lesion(s) within each abnormal frame remains a challenge. However, it is much more significant for the system to robustly detect which frames contain possible lesions than to localize the lesion within these frames, since this can be much easier done by the video reviewers.

The Bag-of-Words or Bag-of-Visual Words (BoW/BoVW) can be considered as a weakly supervised model built upon the notion of visual vocabularies. A visual vocabulary may be seen as a set of "exemplar" image patches (visual words), in terms of which any given image may be described. Typically, this vocabulary is built using a large corpus of representative images of the domain of interest and should be closely related to the problem at hand. The vocabulary may be seen as a means of quantization of the feature space i.e., the one of the local descriptors. Any unseen descriptor may then be easily quantized to its nearest visual word. The description of the whole image is formed by a histogram, counting the appearances of each visual word within it. Apart from the obvious advantage of BoW, i.e., that can be used as a weakly supervised approach as it has already been discussed, it also provides a fixed-size representation, a useful property for tasks such as classification using traditional classifiers e.g., feed-forward neural networks, support vector machines etc. Finally, the visual description provided by BoW may also be used on tasks such as inverted file indexing [7], visual retrieval etc.

An early application of BoW in capsule endoscopy has been investigated using speeded-up robust features (SURF) for polyp detection [8]. In [9] the performance of BoW was investigated using scale-invariant feature transform features (SIFT) and local binary patterns (LBP) for ulcer detection. A more complex feature extraction scheme for the construction required in BoW was proposed in [10]. This scheme was applied for polyp detection and includes extraction of SIFT, LBP, uniform LBP and histogram of oriented gradients (HoG) features from neighbourhoods of salient points detected using the SIFT key-point detector. In the context of bleeding detection, colour histograms extracted from various colour spaces were considered [11]. Colour along with textural information has also been exploited in [12] for detection of gastric and oesophageal cancer, gastritis, and oesophagitis. In that study superpixel segmentation was exploited for estimation of image descriptors from homogeneous regions. As in [12] the descriptors considered include colour histograms from various colour spaces as well as LBP-based textural signatures. Most of the aforementioned approaches are based on support vector machine (SVM) classifiers.

The BoW model was also exploited in the context of unsupervised segmentation of capsule endoscopy videos, based on probabilistic latent semantic analysis (pLSA) [13]. In the context of the analysis of higher resolution endoscopic images, BoW models have been proposed for browsing endoscopic imagery by semantic information [14], colonoscopy image classification [15], and classification of images obtained using chromoendoscopy and narrow-band imaging techniques.

Acknowledging the significance of incorporating an image-level instead of pixel-level annotation process in the development of training datasets for lesion detection systems in VCE, in this paper we investigate a novel BoW-based weakly-supervised learning approach using the state-of-the-art features that have been proposed in [15]. These features represent colour information both at pixel and region level in CIE-*Lab* colour space, and despite their simplicity they have been proved very effective in the detection of a diverse set of abnormalities [5, 17].

The rest of this paper is organized as follows: In Sect. 2 we describe the methodology we followed for the proposed weakly supervised classification scheme. We provide a brief description of both the generic BoW methodology and the approach we followed. Then, in Sect. 3 we demonstrate and discuss our experimental results. Finally, conclusions are drawn in Sect. 4, where we also discuss plans for future continuation of this work.

2 Methodology

BoW is a widely used method to model generic categories in detection, classification and recognition problems [18]. This method has been originally inspired by text document analysis techniques, and consists of calculating word frequencies. The first step of BoW is to describe an image as a set of “words”, which capture its visual content. To this goal, given an adequately large dataset, a set of features is extracted from every image and typically quantized using a clustering approach, e.g., the k-means algorithm [19]. Upon clustering, the centroids (or in some approaches the *medoids*, which opposed to centroids are actual members of the dataset) that have been determined, are used as a “visual vocabulary” and are often referred to as “visual words.”

Each feature is then translated (coded) into one of these visual words, i.e., to the nearest one in the feature space (typically based on the Euclidean distance). The next step involves a histogram construction, which describes the appearance frequency of every visual word within an image. Thus, this histogram is used to characterize the visual content of the image. Among the advantages of BoW, we should emphasize that it succeeds to reduce the problem of classifying a large number of high dimensional vectors from local point descriptors to a fixed-size, one dimensional vector without significant loss of visual information. Finally, any typical classification approach may be used for the classification of these histogram vectors. In this work we choose to use an SVM [20], trained with examples of histograms extracted from both normal and abnormal categories.

We use the well-known SURF (speeded up robust features) algorithm [21], in order to detect interest points and extract descriptions from patches around them. SURF is a powerful and fast descriptor scheme and has been successfully applied to a plethora of computer vision problems. It has been shown to achieve comparable repeatability and performance to other, more sophisticated schemes, at a lower computational cost. It combines a Hessian-Laplace region detector and a gradient orientation-based feature descriptor and is invariant to several image transformations and robust to illumination variations. For interest point selection, we also make use of a "naïve" approach known as "dense sampling". Following this approach, we select all pixels sampled using a regular grid (i.e., one with equal horizontal and vertical inter-pixel distances), which are then used as interest points. Although these points cannot be matched accurately, when compared e.g., to the SURF interest points, they carry valuable information regarding image content interpretation [22].

For the extraction of visual descriptions of patches around the interest points, we also evaluate the colour-based features of [16]. Images are first transformed to the CIE-*Lab* colour space and then, the following colour information is extracted from a square region centered at each point: (i) The *Lab* values of each interest point; (ii) The minimal and maximal values of each component. This results to a vector consisting of 9 values.

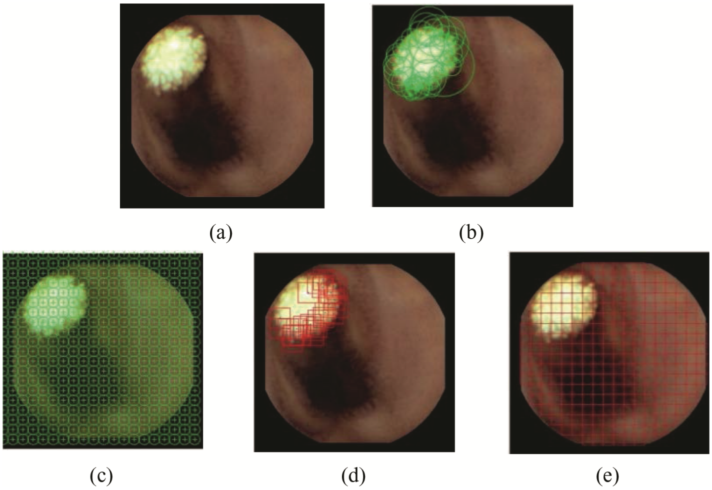

Fig. 1. Image examples of different uses of the algorithms: (a) A raw WCE image depicting lymphangiectasia; (b) SURF; (c) Dense SURF; (d) *Lab*; and (e) Dense *Lab*.

In Fig. 1(b) and (d) we illustrate the set of the SURF interest points extracted from a given VCE image, combined with SURF regions and fixed windows, respectively, whilst in Fig. 1(c) and (e) we illustrate the set of the dense interest points, also combined with SURF regions and fixed windows. One may easily observe that SURF points do not cover the visual properties of the whole image. Yet, the latter is achieved by the dense features.

3 Results

For the evaluation of the proposed weakly-supervised BoW approach, we performed experiments using a subset of dataset 2 from the publicly available KID database [23, 24]. This dataset displays a variety of different kinds of abnormalities. More precisely, the selected subset consists of 227 images of most common inflammatory lesions, e.g., as in Fig. 2(a) including ulcers, aphthae, mucosal breaks with surrounding erythema, cobblestone mucosa, stenoses and/or fibrotic strictures, and significant mucosal/villous oedema. It also includes a set of 1327 normal images derived from the small bowel (728 images), e.g., as in Fig. 2(b) (right), and the stomach (599 images), e.g., as in Fig. 2(b) (left).

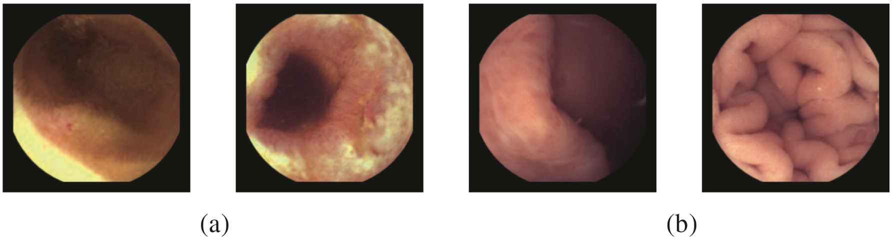

Fig. 2. Representative images from the dataset used in experiments: (a) Inflammatory lesion images, (b) Normal images from the stomach (left) and the small bowel (right)

In order to investigate whether BoW could be used as a reliable classification approach, we compare its performance in four different experiments. These differentiate on the method for the selection of interest points, the description of patches around the aforementioned points; and the colour space used. For the latter case we used greyscale images and also transformations of CIE-*Lab* (using standard illuminant D65), where *L and b* channels had been discarded, keeping only the colour information of *a*. We shall refer to the latter as the "*Lab* images". More specifically, the performed experiments are as follows: (i) SURF points and features on the greyscale image; (ii) dense points and SURF features on the *Lab* images; (iii) SURF points and colour features of [16]; (iv) dense points and colour features of [16]; and (v) the state-of–the-art method of [10], where image description is based on the combination of SIFT and compound local binary pattern features (CLBP). In each case, we extract interest points, then their descriptions, we create the visual vocabulary, which we use for image BoW description and finally train SVM classifiers. In every experiment we use 6-fold cross validation method and estimate the values of area under the receiver operating characteristic (AUC).

The visual vocabulary size ranged from 300 to 1200 words. For the experiments with dense SURF, we used multi-scale feature extraction with scale step 1.6, starting from scale 1.6, up to scale 6.4. We also experimented with various sizes of square regions, for the extraction of the colour features. We used 18 × 18 and 36 × 36 square areas. For dense feature extraction we used grid steps of 4, 10, 18 and 36 pixels, both horizontally and vertically. For the method of [10] we used CLBP of patch size 4 × 4 and 8 × 8. For the classification we used an SVM with RBF kernel.

Most notable results are summarized in Table 1. In this Table we may observe that best performance was achieved for the case of dense *Lab* features using a window size of 18 × 18 pixels and a visual vocabulary of 700 words. The best performance of standard SURF features (i.e., applied on grayscale images) was achieved using dense extraction and a vocabulary size of 800 words. However, this advantageous performance comes at cost of efficiency, since the number of samples obtained by dense SURF is higher (due to the regular sampling process). In addition, our approach had better results in comparison with of the state-of-the-art method of [10]. In any case the application of SURF on the *a* channel of CIE-*Lab* leads to an increase of AUC.

Table 1. Experimental Results; in dense (x), x denotes the step, in SURF (y), y denotes the colour space (g: greyscale, *a*: *a* channel of *Lab*). Note that in case of SURF feature description, image patches are selected by the algorithm, thus marked herein as "N/A"

Feature extraction	Feature description	Window size	Vocabulary size	AUC
dense (18)	*Lab* [15]	18 × 18	500	0.80
dense (4)	SURF (g)	N/A	800	0.70
dense (36)	*Lab* [15]	36 × 36	700	0.79
dense (18)	SURF (g)	18 × 18	800	0.69
dense (10)	*Lab* [15]	18 × 18	700	**0.81**
SURF (*a*)	*Lab* [15]	N/A	700	0.77
SURF (g)	SURF (g)	N/A	500	0.59
SIFT (g)	SIFT + CLBP [10]	4 × 4	500	0.73
SIFT (g)	SIFT + CLBP [10]	8 × 8	500	0.73
SIFT (g)	SIFT + CLBP [10]	8 × 8	700	0.74

4 Conclusions

In this paper we presented a weakly supervised classification scheme for automated lesion detection in VCE videos. We followed the BoW paradigm and created a visual dictionary encoding all extracted image features into visual words. A novel contribution of this paper is that we extended our state-of-the-art colour features [16, 17], according to the bag-of-visual-words model and created BoW image descriptions, which were used to train SVM classifiers. We evaluated four different feature extraction schemes, including a state-of-the-art approach, and investigated among others the use of colour and different sampling schemes. Our results indicate that standard SURF features are not capable of providing a reliable descriptor in the given problem. However, when applied to the *Lab* colour space, their performance is boosted. The latter are able to provide valuable results within the proposed weakly-supervised scheme, which could be used as an alternative to the demanding in terms of manual annotation effort, fully-supervised, schemes.

Open research topics in the area of BoW with application to weakly-supervised lesion detection include the construction of visual vocabularies (flat vs. hierarchical approaches, predefined vs. dynamically selected sizes), the selection of interest points (dense vs. salient vs. hybrid), the selection of patches surrounding interest points (shape, size, orientation) and of course their description (colour vs. greyscale vs. binary descriptors). We plan to perform a thorough systematic investigation to assess the effect of each part of BoW schemes to the overall results, within the context of lesion detection in VCE videos.

Acknowledgements. This research was supported by the special account of research grants of the Technological Educational Institute of Central Greece, Lamia, Greece.

References

1. Koulaouzidis, A., Iakovidis, D.K., Karargyris, A., Rondonotti, E.: Wireless endoscopy in 2020: Will it still be a capsule? World J. Gastroenterol. (WJG) **21**, 5119 (2015)
2. Koulaouzidis, A., Iakovidis, D.K., Karargyris, A., Plevris, J.N.: Optimizing lesion detection in small-bowel capsule endoscopy: from present problems to future solutions. Expert Rev. Gastroenterol. Hepatol. **9**, 217–235 (2015)
3. Zheng, Y., Hawkins, L., Wolff, J., Goloubeva, O., Goldberg, E.: Detection of lesions during capsule endoscopy: physician performance is disappointing. Am. J. Gastroenterol. **107**, 554–560 (2012)
4. Iakovidis, D.K., Koulaouzidis, A.: Software for enhanced video capsule endoscopy: challenges for essential progress. Nature Rev. Gastroenterol. Hepatol. **12**, 172–186 (2015)
5. Hoai, M., Torresani, L., la Torre, F.D., Rother, C.: Learning discriminative localization from weakly labeled data. Pattern Recogn. **47**, 1523–1534 (2014)
6. Blaschko, M., Vedaldi, A., Zisserman, A.: Simultaneous object detection and ranking with weak supervision. In: Advances in Neural Information Processing systems, pp. 235–243 (2010)
7. Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Object retrieval with large vocabularies and fast spatial matching. In: proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8. IEEE (2007)
8. Hwang, S.: Bag-of-visual-words approach based on SURF features to polyp detection in wireless capsule endoscopy videos. In: Proceedings of the 7th International Conference on Advances in Visual Computing (ISVC 2011), vol. 2, pp. 320–327 (2011)
9. Yu, L., Yuen, P.C., Lai, J.: Ulcer detection in wireless capsule endoscopy images. In: ICPR 2012, pp. 45–48. IEEE (2012)
10. Yuan, Y., Li, B., Meng, M.Q.-H.: Improved bag of feature for automatic polyp detection in wireless capsule endoscopy images. IEEE Trans. Autom. Sci. Eng. **13**, 529–535 (2016)
11. Yuan, Y., Li, B., Meng, M.Q.-H.: Bleeding frame and region detection in the wireless capsule endoscopy video. IEEE J. Biomed. Health Inf. **20**, 624–630 (2016)
12. Wang, S., et al.: Computer-aided endoscopic diagnosis without human specific labeling. IEEE Trans. Bio Med. Eng. **53**(11), 2347–2358 (2016)

13. Shen, Y., Guturu, P., Buckles, B.P.: Wireless capsule endoscopy video segmentation using an unsupervised learning approach based on probabilistic latent semantic analysis with scale invariant features. IEEE Trans. Inf Technol. Biomed. **16**, 98–105 (2012)
14. Kwitt, R., Vasconcelos, N., Rasiwasia, N., Uhl, A., Davis, B., Häfner, M., Wrba, F.: Endoscopic image analysis in semantic space. Med. Image Anal. **16**, 1415–1422 (2012)
15. Manivannan, S., Trucco, E.: Learning discriminative local features from image-level labelled data for colonoscopy image classification. In: 2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI), pp. 420–423. IEEE (2015)
16. Iakovidis, D.K., Koulaouzidis, A.: Automatic lesion detection in capsule endoscopy based on color saliency: closer to an essential adjunct for reviewing software. Gastrointest. Endosc. **80**, 877–883 (2014)
17. Iakovidis, D.K., Koulaouzidis, A.: Automatic lesion detection in wireless capsule endoscopy - a simple solution for a complex problem. In: 2014 IEEE International Conference on Image Processing (ICIP), pp. 2236–2240. IEEE (2014)
18. Sivic, J., Zisserman, A.: Efficient visual search of videos cast as text retrieval. IEEE Trans. Pattern Anal. Mach. Intell. **31**(4), 591–606 (2009)
19. Drake, J., Hamerly, G.: Accelerated k-means with adaptive distance bounds. In: 5th NIPS Workshop on Optimization for Machine Learning (2012)
20. Burges, C.J.: A tutorial on support vector machines for pattern recognition. Data Min. Knowl. Disc. **2**(2), 121–167 (1998)
21. Bay, H., Ess, A., Tuytelaars, T., Van Gool, L.: Speeded-up robust features (SURF). Comput. Vis. Image Underst. **110**(3), 346–359 (2008)
22. Tuytelaars, T.: Dense interest points. In: 2010 IEEE Conference on Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2281–2288. IEEE (2010)
23. Iakovidis, D.K., Koulaouzidis, A.: Software for enhanced video capsule endoscopy: challenges for essential progress. Nature Rev. Gastroenterol. Hepatol. **12**(3), 172–186 (2015)
24. Koulaouzidis, A., Iakovidis, D.K.: KID: Koulaouzidis-Iakovidis database for capsule endoscopy (2015). http://is-innovation.eu/kid