



Computer-aided colorectal tumor classification in NBI endoscopy using local features

Toru Tamaki^{a,*}, Junki Yoshimuta^a, Misato Kawakami^b, Bisser Raytchev^a, Kazufumi Kaneda^a, Shigeto Yoshida^c, Yoshito Takemura^d, Keiichi Onji^d, Rie Miyaki^d, Shinji Tanaka^c

^a Department of Information Engineering, Graduate School of Engineering, Hiroshima University, 1-4-1 Kagamiyama, Higashi-hiroshima, Hiroshima 739-8527, Japan

^b Faculty of Engineering, Hiroshima University, 1-4-1 Kagamiyama, Higashi-hiroshima, Hiroshima 739-8527, Japan

^c Department of Endoscopy, Hiroshima University Hospital, 1-2-3 Kasumi, Minami-ku, Hiroshima 734-8551, Japan

^d Department of Gastroenterology and Metabolism, Hiroshima University, 1-2-3 Kasumi, Minami-ku, Hiroshima 734-8551, Japan

ARTICLE INFO

Article history:

Received 1 October 2011

Received in revised form 26 July 2012

Accepted 20 August 2012

Available online 14 September 2012

Keywords:

Colorectal cancer

Colonoscopy

NBI

Pit-pattern

Bag-of-visual-words

ABSTRACT

An early detection of colorectal cancer through colorectal endoscopy is important and widely used in hospitals as a standard medical procedure. During colonoscopy, the lesions of colorectal tumors on the colon surface are visually inspected by a Narrow Band Imaging (NBI) zoom-videoendoscope. By using the visual appearance of colorectal tumors in endoscopic images, histological diagnosis is presumed based on classification schemes for NBI magnification findings. In this paper, we report on the performance of a recognition system for classifying NBI images of colorectal tumors into three types (A, B, and C3) based on the NBI magnification findings. To deal with the problem of computer-aided classification of NBI images, we explore a local feature-based recognition method, bag-of-visual-words (BoW), and provide extensive experiments on a variety of technical aspects. The proposed prototype system, used in the experiments, consists of a bag-of-visual-words representation of local features followed by Support Vector Machine (SVM) classifiers. A number of local features are extracted by using sampling schemes such as Difference-of-Gaussians and grid sampling. In addition, in this paper we propose a new combination of local features and sampling schemes. Extensive experiments with varying the parameters for each component are carried out, for the performance of the system is usually affected by those parameters, e.g. the sampling strategy for the local features, the representation of the local feature histograms, the kernel types of the SVM classifiers, the number of classes to be considered, etc. The recognition results are compared in terms of recognition rates, precision/recall, and *F*-measure for different numbers of visual words. The proposed system achieves a recognition rate of 96% for 10-fold cross validation on a real dataset of 908 NBI images collected during actual colonoscopy, and 93% for a separate test dataset.

© 2012 Elsevier B.V. All rights reserved.

1. Introduction

Colorectal cancer was the third-leading cause of cancer death in Japan in 2009, and the leading cause for death for Japanese women over the last 6 years (Ministry of Health et al., 2009). In the United States, a report estimates that 49,920 people have died from colorectal cancer in 2009 (National Cancer Institute, US National Institutes of Health, 2010) and in UK 26,423 people in 2008 (Cancer research UK, 2011). WHO has released projections (Health Statistics and Informatics Department, World Health Organization, 2008) in which the number of deaths in the world is estimated to be about 780,000 in 2015, and is expected to rise to 950,000 in 2030. Because of the increased

number of patients, skill/training required for medical examinations, and the need for objective evaluation in order to allow non-experts to achieve high diagnostic accuracy and reduce inter/intra-observer variability, it is very important to develop computerized systems able to provide supporting diagnosis for this type of cancer.

An early detection of colorectal cancer by *colonoscopy* or *colorectal endoscopy*, an endoscopic examination of the colon, is important and widely used in hospitals as a standard medical procedure. During colonoscopy, the lesions of colorectal tumors on the colon surface are often visually inspected by a Narrow Band Imaging (NBI) zoom-videoendoscope with a magnification factor of up to 100 (Tanaka et al., 2006). By using the visual appearance of colorectal tumors in endoscopic images, histological diagnosis is presumed based on classification schemes for *pit-patterns* (Kudo et al., 1994, 1996) and NBI magnification findings (Kanao et al., 2009; Oba et al., 2010).

* Corresponding author. Tel.: +81 82 424 7664.

E-mail address: tamaki@hiroshima-u.ac.jp (T. Tamaki).

URL: <http://home.hiroshima-u.ac.jp/tamaki/> (T. Tamaki).

A diagnosis by visual inspection, however, is affected by two main factors. One factor is the skill and familiarity of each inspector, i.e., performance differences between expert and non-expert endoscopists (Oba et al., 2010; Higashi et al., 2010; Chang et al., 2009). For example, Higashi et al. (2010) studied the effect of an intensive training program for non-experienced endoscopists (NEE), less-experienced (more than 5 years) endoscopists (LEE) who have never used NBI, and compared to high-experienced endoscopists (HEE) who had used NBI more than 5 years. After the training program, the LLE group improved their accuracy from 73% to 90%, which is close to the accuracy 93% of the HEE group (for NBI zoom-endoscopy). However, the NEE group improved from 63% only up to 74%, which is not a satisfactory result. The other factor concerns inter-observer variability and intra-observer variability (Meining et al., 2004; Mayinger et al., 2006; Oba et al., 2010), that is, whether different diagnoses would be made by different observers (inter-observer variability), or by the same observer at the different times (intra-observer variability). Table 2 shows an example of variability for NBI magnification findings (Oba et al., 2010), in which there was a high level of inter- and intra-observer variability.

Hence, a computer-aided system for supporting the visual inspection would be of great help for colonoscopy, due to the large number of images of colorectal tumors which should be classified in a regular inspection in an effort to detect cancer in its early stage. Fig. 1 shows a screen shot of an actual monitor at which endoscopists and doctors are looking during endoscopy. The monitor displays the live video from a videoendoscope (the largest image in Fig. 1). When an endoscopist presses a button on the endoscope, a frame of the video is captured as a still image, which is stored to a file and also displayed on the monitor below the live video (four still images captured are shown Fig. 1). Actually, endoscopists usually take a lot of pictures of a polyp for finding a good image of the polyp for a paper-based report with medical evidences (Aabakken, 2009). In the clinical workflow, it would be helpful if an *objective* diagnosis by a computer-aided system as a kind of second opinion could be provided onto the video monitor directly or for pictures taken during the examination (Takemura

et al., 2012). This can be used in two ways. First, it can assist in the endoscopist's decision-making which part of a tumor should be shot, or whether some more pictures need to be taken. For this purpose, processing a fixed region around the center of the live video would be enough because the region of interest should always be shot at the center. Second, even after the examination, an objective information is useful for supporting the doctors by allowing them to specify regions of interest in the captured still images to be processed. Many attempts have been done in this direction (i.e., classifying trimmed endoscopic images) (Gross et al., 2009a,b; Häfner et al., 2008, 2009a,b,c,d; Häfner et al., 2010b,a; Häfner et al., 2009e; Kwitt and Uhl, 2008; Kwitt et al., 2010; Stehle et al., 2009), however, most of them have not been verified for NBI images but only for pit-pattern images of a *chromoendoscopy*, which requires a dye-spraying process. Since those studies have developed different techniques of texture analysis specific to the visual appearance of pit-pattern images, it is not straightforward to extend them to NBI images.

In this paper, we focus on a recent recognition framework based on local features which has been used with great success for a range of challenging problems such as category-level recognition (Csurka et al., 2004; Nowak et al., 2006; Lazebnik et al., 2006) as well as instance recognition (Sivic and Zisserman, 2003; Chum et al., 2007) and also endoscopic image retrieval (André et al., 2012, 2011c,a,b, 2009). To deal with the problem of computer-aided classification of NBI images, we explore a local feature-based recognition method, *bag-of-visual-words* (BoVW or BoW), and provide extensive experiments on a variety of technical aspects. Our prototype system used in the experiments consists of a BoW representation of local features followed by Support Vector Machine (SVM) classifiers. BoW has been widely used for general object recognition and image retrieval as well as texture analysis. Local features such as Scale Invariant Feature Transform (SIFT) are extracted from an image and their distribution is modeled by a histogram of representative features (also known as visual words, VWs). A number of local features are extracted by using sampling schemes such as Difference-of-Gaussians (DoG-SIFT) and grid sampling (gridSIFT). In addition, in this paper we propose a new combination of local features and sampling schemes, *DiffSIFT* and *multi-scale gridSIFT*. Extensive experiments done by varying the parameters for each component are needed, for the performance of the system is usually affected by those parameters, e.g. the sampling strategy for the local features, the representation of the local feature histograms, the kernel types of the SVM classifiers, the number of classes to be considered, etc. The recognition results are compared in terms of recognition rates, precision/recall, and *F*-measure for different numbers of visual words. The proposed system achieves a recognition rate of 96% for 10-fold cross validation on a real dataset of 908 NBI images collected during actual colonoscopy, and 93% for a separated test dataset.

The rest of the paper is organized as follows: Section 2 reviews the medical aspects of colorectal cancers, two types of visual assessment strategies (pit-pattern classification and NBI magnification findings), and related work. Section 3 gives an outline of the bag-of-visual-words framework, and Section 4 describes the details of each component of the framework. Section 5 presents the experimental results obtained for 10-fold cross validation and a test dataset. Section 6 shows conclusions and discussions.

2. Colonoscopy and related work

2.1. Colorectal endoscopy and pathology

There are many ways of examination and screening of the colon. Those include colonoscopy, as well as Fecal Occult Blood Test

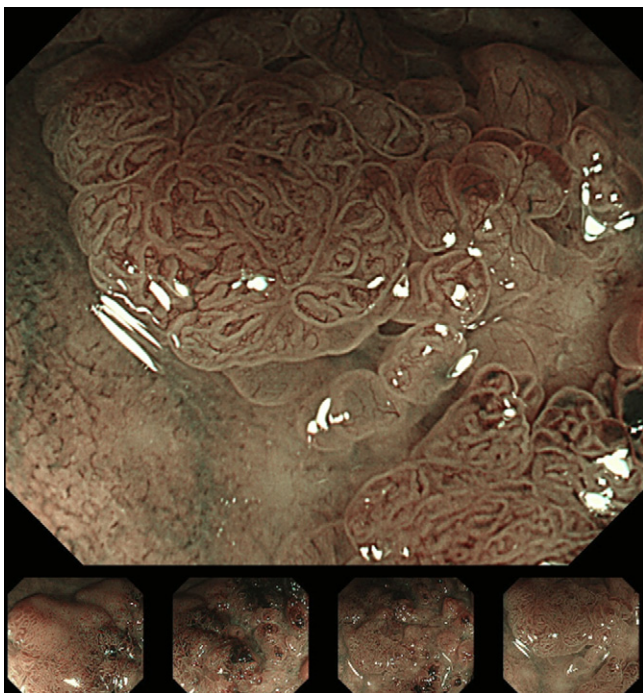


Fig. 1. A screen shot of an actual NBI zoom-videoendoscopy.

(FOBT) (Sanford and McPherson, 2009; Heitman et al., 2008), Digital Rectal Examination (DRE) (Gopalswamy et al., 2000), biomedical markers (Karl et al., 2008), CT colonography (Halligan and Taylor, 2007; Yoshida and Dachman, 2004), MR colonography (Shin et al., 2010; Beets and Beets-Tan, 2010), Marvin Positron Emission Tomography (PET) (Lin et al., 2011), Ultrasound (Padhani, 1999), Double Contrast Barium Enema (DCBE) (Johnson et al., 2004; Canon, 2008), Confocal Laser Endomicroscopy (CLE) (Kiesslich et al., 2007), and Virtual Endoscopy (Oto, 2002). Interested readers are referred to the following surveys on colorectal cancer screening and clinical staging of colorectal cancer (Jenkinson and Steele, 2010; Tweedle et al., 2007; Barabouli and Wong, 2005) and endoscopic imaging devices (Gaddam and Sharma, 2010).

The “gold standard” (Häfner et al., 2010a) for colon examination is *colonoscopy*, an endoscopy for the colon, which has been commonly used since the middle of the last century (Classen et al., 2010). Colonoscopy is an examination with a video-endoscope equipped with a CCD camera of size less than 1.4 cm in diameter at the top of the scope, as well as a light source and a tube for dye spraying and surgical devices for Endoscopic Mucosal Resection (EMR) (Ye et al., 2008).¹

Colorectal polyps detected during colonoscopy are histologically identified as cancers (*carcinomas*) or non-cancers through *biopsy*, by removing the lesions from the colon surface. The polyps are histologically classified into the following groups (Hirata et al., 2007b,a; Kanao et al., 2009; Raghavendra et al., 2010): *hyperplasias* (HP), *tubular adenomas* (TA), *carcinomas with intramucosal invasion* to *scanty submucosal invasion* (M/SM-s) and *carcinomas with massive submucosal invasion* (SM-m). TA, M/SM-s and SM-m are also referred to as *neoplastic*, and HP and normal tissue as *non-neoplastic*.

HP are non-neoplastic and hence not removed. TA are polyps and likely to develop into cancer (many cancers follow this *adenoma-carcinoma sequence* (Gloor, 1986)) and can be endoscopically resected (i.e., removed during colonoscopy). M/SM-s cancers can also be endoscopically resected, while SM-m cancers should be surgically resected (Watanabe et al., 2012). Since SM-m cancers deeply invade the colonic surface, it is difficult to remove them completely by colonoscopy due to the higher risk of *lymph node metastasis*.

SM-m cancers need to be discriminated from M/SM-s cancers and TA polyps *without* biopsy because the risk of *complications* during colonoscopy should be minimized by avoiding unnecessary biopsy. Biopsy also tends to be avoided because the biopsied tumor develops to a *fibrosis* (Matsumoto et al., 2010) which causes a *perforation* at the time of surgery. While there is controversy about the benefit of avoiding biopsy (Gershman and Ament, 2012), recent advances in the field of medical devices (Gaddam and Sharma, 2010) would allow the endoscopic assessment of histology to be established and documented without biopsy in the near future (Rex et al., 2011).

2.2. Pit-pattern classification

In order to determine a histological diagnosis by using the visual appearance of colorectal tumors in endoscopic images, the *pit-pattern classification* scheme was proposed by Kudo et al. (1994, 1996) and later modified by Kudo and Tsuruta as described in Imai et al. (2001). A *pit pattern* refers to the shape of a *pit* (Bank et al., 1970), the opening of a colorectal crypt, and can be used for the visual inspection of mucosal surface. During a *chromoendoscopy*, indigo carmine dye spraying or crystal violet staining are

used to enhance the microscopic appearances of the pit patterns illuminated by a white light source. Fig. 2a shows an image of a colon taken by an endoscope without staining, while Fig. 2b and c show images stained by two different dyes. In (b) and (c), the structure of the mucosal surface on the polyp is well enhanced and the visibility is much better than in white light colonoscopy (a). Pit-pattern analysis started in the 1970s (Bank et al., 1970; Kosaka, 1975), and developed over the next 20 years (Kudo et al., 1994, 1996; Imai et al., 2001). Currently, the most widely used classification categorizes pit-patterns into types I to V. Types III and V are further divided (Fig. 3) into III_S (S: Smaller) and III_L (L: Larger), V_I (I: Irregular) and V_N (N: Non-structure).

The pit-pattern classification has been used to differentiate non-neoplastic colorectal lesions from neoplastic ones (for example, Fu et al., 2004), and to guide therapeutic decisions. Indicated diagnosis (Tanaka et al., 2006; Kanao, 2008) roughly corresponds to: follow up (no resection) (type I and II), endoscopic resection (type III_S, III_L, and IV), surgery (type V_N), and further examinations (type V_I).

2.3. NBI magnification findings

Narrow Band Imaging (NBI) (Gono et al., 2003, 2004; Machida et al., 2004; Sano et al., 2006) is a recently developed videoendoscopic system that uses RGB rotary filters placed in front of a white light source to narrow the bandwidth of the spectral transmittance. The central wavelengths of the RGB filters are set to 540 and 415 nm with a bandwidth of 30 nm, since the hemoglobin in the blood absorbs lights of these wavelengths. NBI provides a limited penetration of light to the mucosal surface, and enhances the microvessels and their fine structure on the colorectal surface (see Fig. 2d). NBI enables an endoscopist to quickly switch a white light colonoscopy image to an NBI colonoscopy image when examining tumors, while a chromoendoscopy requires a cost for spraying, washing and vacuuming dye and water, and prolongs the examination procedure.

NBI was introduced to gastro or esophageal examinations in the last decade (Panossian et al., 2011) as dye-spraying needs to be avoided due to its irritancy. NBI has been also used for colonoscopy since around 2004 (Gono et al., 2003, 2004; Machida et al., 2004; Sano et al., 2006). By using NBI, the pits are also indirectly observable, for the microvessels between the pits are enhanced in black, while the pits are left in white. The high visibility of the microvessels (Ignjatovic et al., 2011; Higashi et al., 2010; Chang et al., 2009) has led to the wide use of NBI both in pit-pattern analysis (Hirata et al., 2007b) and microvessel analysis (Hirata et al., 2007a).

Several categorizations of *NBI magnification findings*, diagnosis based on magnifying NBI endoscopic images, have been recently developed by different medical research groups (Oba et al., 2011):

- *Hiroshima University Hospital* (Kanao et al., 2009; Oba et al., 2010): three main types (A, B, and C) and subtypes (C1, C2, and C3).
- *Sano Hospital* (Sano et al., 2009; Ikematsu et al., 2010): three main types (I, II, and III) and subtypes (IIIA and IIIB) based on the density of capillary vessels, lack of uniformity, ending and branching of the vessels.
- *Showa University Northern Yokohama Hospital* (Wada et al., 2009): six main types (A to F) associated with two subtypes (1 and 2) based on the thickness, network structure, density and sparsity of the vessels.
- *The Jikei University School of Medicine* (Saito et al., 2011; Tamai et al., 2011): four main types (1, 2, 3 and 4) and subtypes (3 V and 3 I) based on detail and regularity of the vessels.

¹ Also such devices are used for Endoscopic Piecemeal Mucosal Resection (EPMR) (Tamegai, 2007), or Endoscopic Submucosal Dissection (ESD) (Saito et al., 2007).

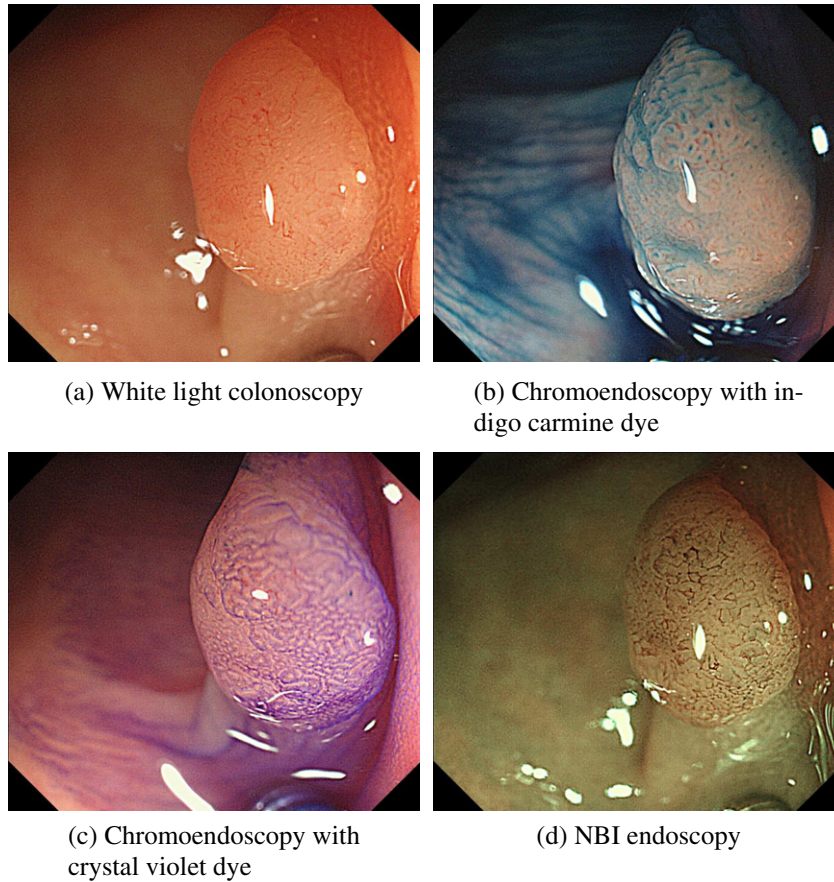


Fig. 2. Images showing different colonoscopy processes.

I		Round pit (normal pit)	
II		Asteroid pit	
III _s		Tubular or round pit that is smaller than the normal pit (Type I)	
III _L		Tubular or round pit that is larger than the normal pit (Type I)	
IV		Dendritic or gyrus-like pit	
V _I		Irregular arrangement and sizes of III _s , III _L , IV type pit pattern	
V _N		Loss or decrease of pits with an amorphous structure	

Fig. 3. Pit-pattern classification of colorectal lesions (Takemura et al., 2010).

Among those, in this paper we use the classification proposed by Tanaka's group (Kanao et al., 2009; Oba et al., 2010) at Hiroshima University Hospital. It divides the microvessel structures in an

NBI image into types A, B, and C (see Fig. 4). In type A, microvessels are not observed, or slightly observed but opaque with very low contrast (typical images are shown in the top row of Fig. 5). In type

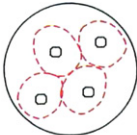

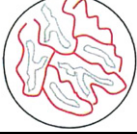


Type A		Microvessels are not observed or extremely opaque.
Type B		Fine microvessels are observed around pits, and clear pits can be observed via the nest of microvessels.
Type C	1 	Microvessels comprise an irregular network, pits observed via the microvessels are slightly non-distinct, and vessel diameter or distribution is homogeneous.
	2 	Microvessels comprise an irregular network, pits observed via the microvessels are irregular, and vessel diameter or distribution is heterogeneous.
	3 	Pits via the microvessels are invisible, irregular vessel diameter is thick, or the vessel distribution is heterogeneous, and avascular areas are observed.

Fig. 4. NBI magnification findings (Kanao et al., 2009).

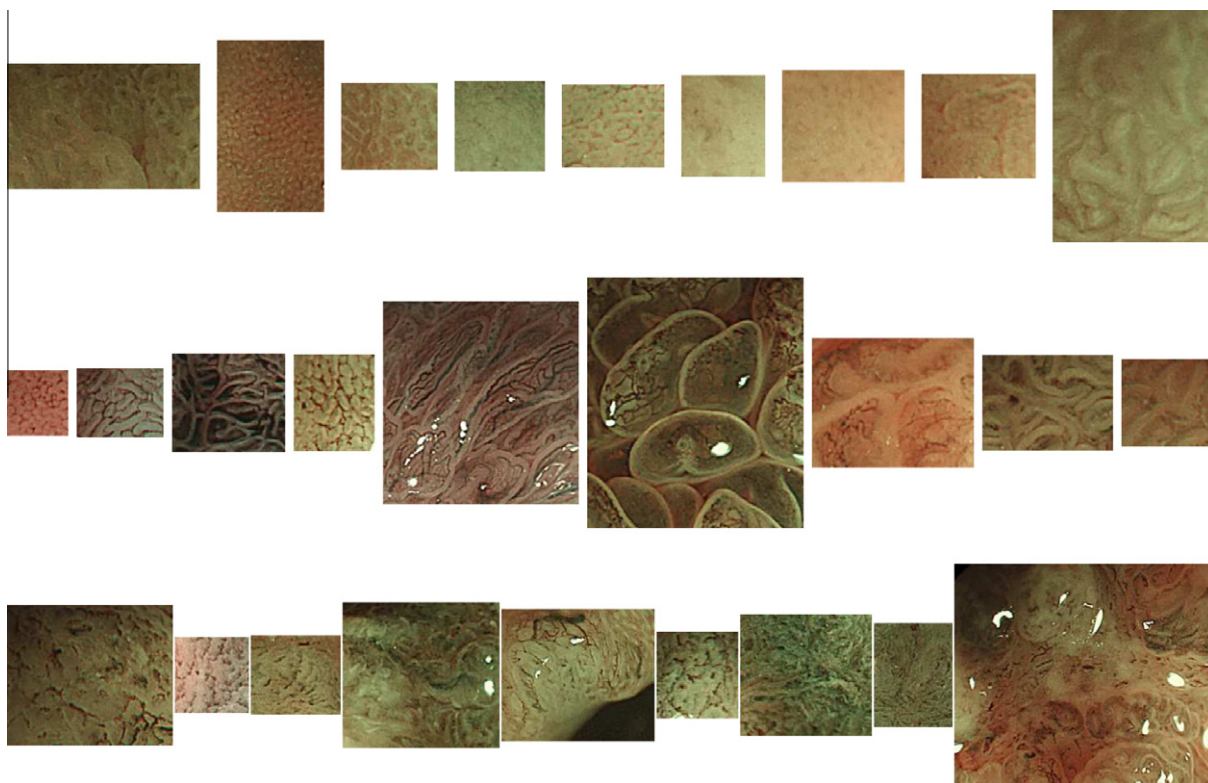


Fig. 5. Examples of NBI images of types A (top row), B (middle row), and C3 (bottom row).

B, fine microvessels are visible around clearly observed pits (the middle row of Fig. 5). Type C is divided into three subtypes C1, C2, and C3 according to detailed texture. In type C3, which exhibits the most irregular texture, pits are almost invisible because of the

irregularity of tumors, and microvessels are irregular and thick, or heterogeneously distorted (the bottom row of Fig. 5).

This classification has been shown to have a strong correlation with histological diagnosis (Kanao et al., 2009), as shown in Table 1.

Table 1

Relationship between NBI magnification findings and histologic findings (from Kanao et al. (2009)).

Type	HP	TA	M/SM-s	SM-m
A	80%	20%		
B		79.7%	20.3%	
C1		46.7%	42.2%	11.1%
C2			45.5%	54.5%
C3				100%

Table 2

(Top) Inter-observer variability and (bottom) intra-observer variability in assessment of NBI magnification findings (from Oba et al. (2010)).

Group A	Group B		
	C1	C2	C3
C1	85.4%	14.6%	
C2	13.2%	76.3%	10.5%
C3		12.7%	87.3%

Group B (1st)	Group B (2nd)		
	C1	C2	C3
C1	94.0%	6.0%	
C2	20.4%	71.4%	8.2%
C3		21.4%	78.6%

80% of type A corresponded to HP, and 20% to TA. 79.7% of type B corresponded to TA, and 20.3% to M/SM-s. For type C3, on the other hand, 100% were of SM-m. Therefore, it is very important to detect type C3 among other types, not just as a two-class problem of neoplastic (type A) and non-neoplastic (other types) lesions (Stehle et al., 2009; Tischendorf et al., 2010).

Note that we exclude types C1 and C2 (in a similar manner to Kanao et al., 2009) from most of the experiments in this paper. One reason is that the inter- and intra-observer variability is not small and adjacent levels are interchangeably recognized as shown in Table 2. In particular, type C2 has a large variability, which leads to poor performance of classification systems. Actually, our method does not perform well for this five-class classification problem (as detailed in Section 5.7.1): the recognition rate is at most 75%, and recall rates for types C1, C2, and C3 are about 50%, 10%, and 50%, respectively. This means that most images of type C2 are misclassified to different classes.

Another reason to exclude types C1 and C2 is that the subtypes C1, C2, and C3 are defined as the degree of irregularity of the microvessel network, rather than by discrete labels. This means that categorizing samples to intermediate labels (C1 and C2) makes little sense. Instead, two ends of irregularity (B and C3) are better to be used to indicate numerical values between them as the development of a polyp. The prediction of the numerical value between B and C2 is one of our future tasks.

However, still it is not easy to discriminate between the remaining three classes in a recognition task. Several examples of images in the dataset used in our experiments are shown in Fig. 5 (see also Section 5.1). In particular, images of type B have a wide variety of textures, for which a simple texture analysis might not work well. Moreover, some images of types B and C3 have similar appearances to each other as shown in Fig. 6, which makes even the three-class classification quite a challenging task.

2.4. Related work and contributions

To support colorectal cancer diagnosis with computer-aided systems, many approaches have been discussed and proposed from a variety of aspects.

Endoscopy frame selection: Oh et al. (2007) proposed a method for selecting visually preferable in-focus frames and excluding out-of-focus and blurred frames.

3D reconstruction from videoendoscopy: Hirai et al. (2011) used the folds of the colonic surfaces for matching in order to reconstruct the 3D geometry of the colon.

Polyp detection: Karkanis et al. (2003) used color wavelets followed by LDA, and Maroulis et al. (2003) used neural networks for detection of polyps in colorectal videoendoscopy. Sundaram et al. (2008) detect colon polyps by using 3D shape information reconstructed by CT images.

Biopsy image recognition: Tosun et al. (2009) used unsupervised texture segmentation for detecting cancer regions. Al-Kadi (2010) developed a method using Gaussian Markov random fields, and Gunduz-Demir et al. (2010) used a local object-graph to segment colon glands.

Endoscopic image retrieval: André et al. (2012), André et al. (2011c), André et al. (2011a), André et al. (2011b), and André et al. (2009) used a content-based video retrieval method for *in vivo* endomicroscopy.

While a huge amount of research on medical image analysis has been done including the work mentioned above, only few groups have worked on automatic visual inspection of colonoscopy by using the pit-pattern classification. We should note the work by the Multimedia Signal Processing and Security Lab, Universität Salzburg which is one of the most active groups in this field (Häfner et al., 2006, 2008, 2009a,b,c,d, 2010b,a, 2009f; Häfner et al., 2009e, 2008; Kwitt et al., 2010; Kwitt and Uhl, 2007a). In Häfner et al. (2010a), the edges of Delaunay Triangles produced by Local Binary Patterns (LBP) of RGB channels were used as features, achieving a recognition rate of 93.3%. In Häfner et al. (2010b), morphology and edge detection were applied to images for extracting 18 features followed by feature selection, and recognition rates of 93.3% for a two-class problem and 88% for six classes were achieved. Kwitt et al. (2010) introduced a generative model that involves prior distributions as well as posteriors, and employed a two-layered cascade-type classifier that achieved 96.65% for two classes and 93.46% for three classes. Other work from this group includes texture analysis with wavelet transforms (Häfner et al., 2009f; Kwitt and Uhl, 2007a), Gabor wavelets (Kwitt and Uhl, 2007b), histograms (Häfner et al., 2006), and others. In our previous work (Takemura et al., 2010), we have used shape analysis of extracted pits, such as area, perimeter, major and minor axes of a fit ellipse, diameter, and circularity.

There is much less research on automatic classification of NBI endoscopy images, compared to research based on pit-patterns, because NBI systems became popular only after 2005. To the best of our knowledge, only the group at the Institute of Imaging and Computer Vision at RWTH Aachen University has reported some studies including colorectal polyp segmentation (Gross et al., 2009a) and localization (Breier et al., 2011). They used Local Binary Patterns (LBP) of NBI images, as well as vessel features extracted by edge detection (Gross et al., 2009b), and vessel geometry features extracted by segmentation (Stehle et al., 2009; Tischendorf et al., 2010). They classified NBI images of colorectal tumors into non-neoplastic and neoplastic: images were directly related with histological diagnosis, and no visual classification scheme was introduced.

In contrast, the present paper makes the following two contributions. First, we investigate if the visual inspection schemes are valid for NBI magnification findings, like has been shown for pit-pattern classification. Previous works on NBI image recognition (Gross et al., 2009b; Stehle et al., 2009; Tischendorf et al., 2010) are based on histopathological results for classifying tumors into non-neoplastic and neoplastic. However, it is very important to discriminate M/SM-s and SM-m cancers (both are neoplastic)

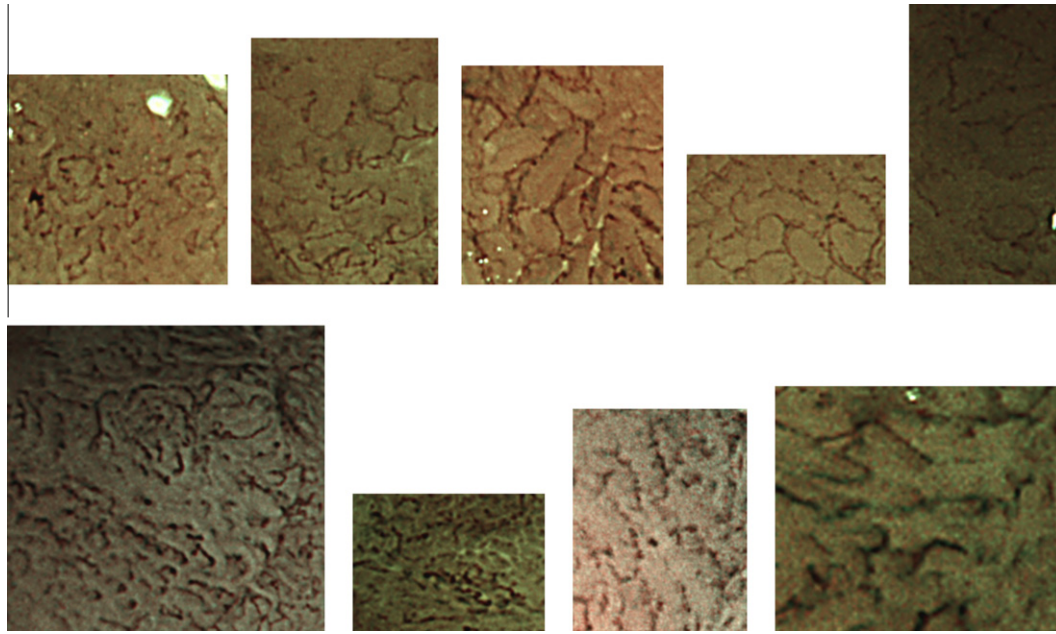


Fig. 6. Some similar images of types B and C3.

without biopsy or EMR, as described in Section 2.1. Second, due to the large intra-class variation of the visual appearance of NBI images, we propose to use for representation the bag-of-visual-words (BoW) framework. Since BoW has been successfully applied to a variety of recognition problems, including generic object recognition and texture analysis, it is natural to expect that BoW would perform well for NBI images that have a wide range of variation in texture of colorectal tumors. In the following sections, we outline the process of recognition with BoW and address some technical aspects to be considered in the extensive experiments which is the significantly extended version of our prior work (Tamaki et al., 2011).

It should be noted that BoW has been also applied to image retrieval of probe-based Confocal Laser Endomicroscopy (pCLE) by André et al. (2012, 2011c,a,b, 2009) for real-time diagnosis from *in vivo* endomicroscopy. For video and image retrieval, they proposed “Bag of Overlap-Weighted Visual Words” constructed from multi-scale SIFT descriptors combined with the k -nearest neighbors retrieval. Their emphasis is on a “retrieval” approach (Müller et al., 2009, 2004) since it enables endoscopists to directly compare the current image and retrieved images for diagnosis.

In contrast, our approach targets classification for providing an objective diagnosis and a kind of “second opinion” to avoid oversights during colonoscopy and assist an endoscopist’s decision-making (Takemura et al., 2012), as well as for real-time *in vivo* endoscopic diagnosis (Rex et al., 2011). To demonstrate our concept, a prototype recognition system for NBI video sequences has been constructed, which is shown in the last section of this paper.

3. Outline of bag-of-visual-words

A bag-of-visual-words (BoW) is a representation of images mainly used for generic object recognition or category recognition. The use of bag-of-visual-words for generic visual categorization (Csurka et al., 2004; Nowak et al., 2006; Lazebnik et al., 2006) and instance recognition (Sivic and Zisserman, 2003; Chum et al., 2007) has been motivated by the success of the bag-of-words method for text classification (Joachims, 1998; Tong and Koller, 2002; Lodhi et al., 2002; Cristianini et al., 2002) in which a document or text is represented by a histogram of words appearing

regardless of word order. Similarly, BoW represents an image as a histogram of representative local features extracted from the image regardless of their location. Each representative local feature, called a *visual word* (or codeword, visterm, visual texton), is the center a cluster of local features. A set of visual words (codewords) is often called a visual vocabulary (or a codebook).

Fig. 7 shows an overview of the recognition process with BoW. Local features are extracted from images, and then clustered with vector quantization to represent an image with a histogram of visual words. This approach includes the following steps:

1. Training phase
 - (a) Extracting feature points from the training images.
 - (b) Computing feature vectors (descriptors) for each feature point.
 - (c) Clustering feature vectors to generate visual words.
 - (d) Representing each training image as a histogram of visual words.
 - (e) Training classifiers with the histograms of the training images.
2. Test phase
 - (a) Extracting feature points from a test image.
 - (b) Computing feature vectors (descriptors) for each feature point.
 - (c) Representing the test image as a histogram of visual words.
 - (d) Classifying the test image based on its histogram.

BoW can be divided into three main components: detection and description of local features, visual word representation, and classification.

Detection and description of local features is the first step where information is extracted from the images in the form of feature vectors. This step can be further divided into two steps: *detection*, or *sampling* (detecting the location where the features are to be extracted), and *description* (how the features are to be represented). In the next section, we focus on the Scale Invariant Feature Transform (SIFT) (Lowe, 1999, 2004), which is a standard feature descriptor and known to perform better (Mikolajczyk and Schmid, 2003, 2005) than other features such as PCA-SIFT (Ke and Sukthankar, 2004), for example. For detection, both the Difference-of-Gaussians

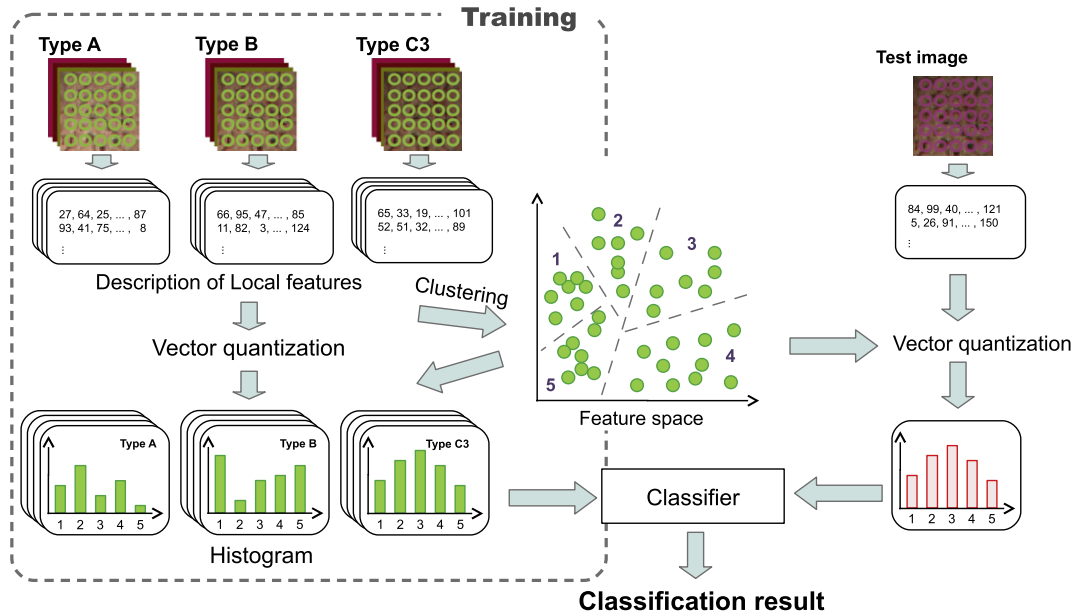


Fig. 7. Overview of bag-of-visual-words.

(DoG) detector (Lowe, 1999, 2004) and grid sampling (Nowak et al., 2006; Fei-Fei and Perona, 2005; Bosch et al., 2007) are investigated in this paper. Other detectors (Tuyltelaars and Mikolajczyk, 2007) such as Harris (Harris and Stephens, 1988), Harris–Laplace (Mikolajczyk and Schmid, 2004) and dense interest points (Tuyltelaars, 2010) are also good alternatives to DoG, and investigating them is left for future work. SURF (Bay et al., 2006, 2008) is known to be faster than SIFT in detection speed, however, a GPU implementation of SIFT is available (Wu, 2010), and the speed of detection depends on the sampling strategy. We have seen in preliminary experiments that grid sampling of SIFT is reasonably fast.

Visual word representation involves clustering of the extracted local features for finding the visual words (the cluster centers), and histogram representation of the images. For clustering a large number of local features, hierarchical *k*-means (Nister and Szwed, 2006) has been widely used for representing a vocabulary tree. From the viewpoint that the histogram of visual words can be considered as a representation of the local feature distributions, many vocabulary-based approaches (Amores, 2010) and variants

have been recently proposed, such as Gaussian Mixture Model (GMM) (Campbell et al., 2006; Farquhar et al., 2005; Perronnin et al., 2006; Perronnin, 2008; Zhou et al., 2008), kernel density estimation (van Gemert et al., 2008), soft assigning (Philbin et al., 2008) and global Gaussian approach (Nakayama et al., 2010). In this paper, we employ a simple vocabulary-based approach as it has a lower computational cost than the other methods, and investigate two types of histogram representations.

Classification is used to classify the test image based on its histogram representation. In this paper, Support Vector Machine (SVM) (Vapnik, 1998; John Shawe-Taylor, 2000; Schölkopf and Smola, 2002; Steinwart and Christmann, 2008) is used with different kernel types. Other classifiers, such as Naive Bayes (Csurka et al., 2004; Bishop, 2006), Probabilistic Latent Semantic Analysis (pLSA) (Hofmann, 1999; Qiu and Yanai, 2008; Yanai and Qiu, 2009), or Multiple Kernel Learning (Sonnenburg et al., 2006; Joutou and Yanai, 2009) have also been used for category recognition.

4. Technical details

In this section, we describe some technical aspects explored in the experiments for classifying NBI images based on the NBI magnification finding by using the BoW representation.

4.1. Local features

4.1.1. DoG-SIFT

SIFT (Lowe, 1999, 2004) is a local feature descriptor invariant to shift, rotation, scale and intensity change. Difference-of-Gaussians (DoG) is used to detect keypoints (Fig. 8a) where the DoG response is maximum in space (location) and scale (different width of Gaussian) (Schmid and Mohr, 1997; Lindeberg, 1994; Koenderink, 1984). The maximum direction of the intensity gradients around each keypoint is computed as its orientation. Then, a histogram of gradients within the patch, the local region centered at the keypoint, is computed. Usually 8 directions of gradients in 4 × 4 blocks in a patch are used to make a 128-dimensional vector as a SIFT descriptor.

To remove unreliable keypoints, the eigenvalues of the patch are used for eliminating points on edges (Lowe, 2004). Keypoints

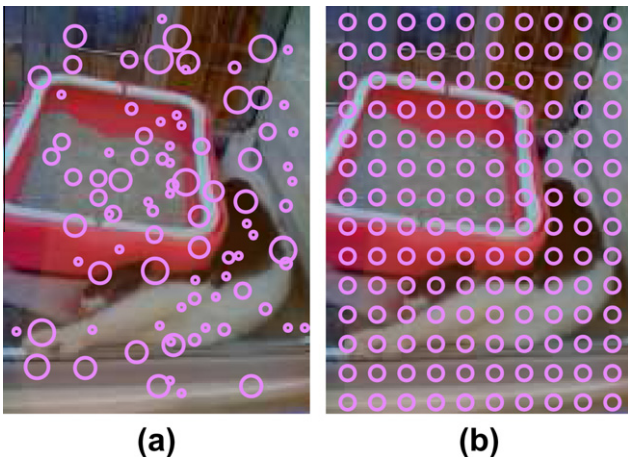


Fig. 8. DoG and grid sampling of features. (a) Keypoints detected by DoG. (b) Grid sampling.

are discarded if the Hessian H of the associated patch does not satisfy

$$\frac{\text{Tr}(H)^2}{\text{Det}(H)} < \frac{(r+1)^2}{r}, \quad (1)$$

where $r = 10$ (Lowe, 2004).

DoG responses are also used for eliminating points with low contrast. Keypoints at location \hat{x} are discarded if the value of the DoG response

$$D(\hat{x}) = D + \frac{1}{2} \frac{\partial D^T}{\partial x} \hat{x}, \quad (2)$$

is less than a threshold $D_{th} = 0.03$ (Lowe, 2004), which means low contrast and hence unstable extrema.

4.1.2. gridSIFT

SIFT descriptors at sparsely detected keypoints by DoG can be used for recognition, however, densely sampled SIFT descriptors are known to perform better (Fei-Fei and Perona, 2005; Jurie and Triggs, 2005; Nowak et al., 2006; Herve et al., 2009). The SIFT features obtained by grid sampling are referred to as *gridSIFT* (Nowak et al., 2006; Fei-Fei and Perona, 2005; Bosch et al., 2007): features are sampled at points on a regular grid over the image (Fig. 8b). For gridSIFT, the following parameters need to be specified:

- *Grid spacing* (Fig. 10) affects how densely the features are extracted. Smaller spacing generates more features while at the same time increasing computation and storage cost, but generally performs better.
- The *scale* of a patch around each grid point affects the extent of the spatial information involved in the feature at the point. Although DoG finds the optimal scale for each keypoint, gridSIFT requires a range of scales to be specified at each grid point. In this paper, we refer the scale of a SIFT descriptor to the width of a square block (see Fig. 9).

We use each descriptor of different scales at the sample point as a single feature vector (Fig. 10a): for example, if there are 100 points on the grid and 4 scales are being used, then we have 400 gridSIFT descriptors, where each descriptor is a 128 dimensional vector. An alternative way to define the descriptor is to combine the descriptors of different scales at the sample point into a single feature vector (Fig. 10b): e.g., in this case we have 100 descriptors, each being a 512 dimensional vector (we call this variant *multi-scale gridSIFT*).

Note that we do not use spatial information as in Spatial Pyramid Matching (Lazebnik et al., 2006) and Pyramid Histogram Of visual Words (PHOW) (Bosch et al., 2007). And also we do not use orientation information as in SIFT descriptors; i.e., orientations of SIFT descriptors of gridSIFT are fixed. Since the images in the NBI

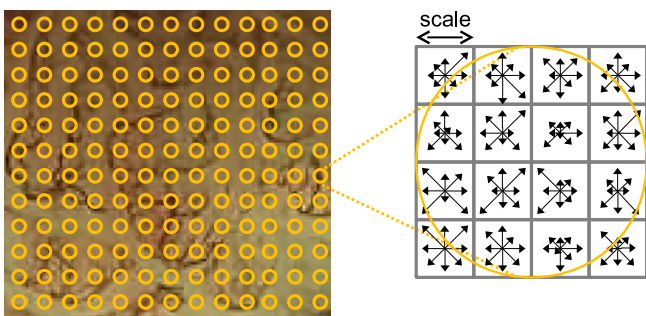
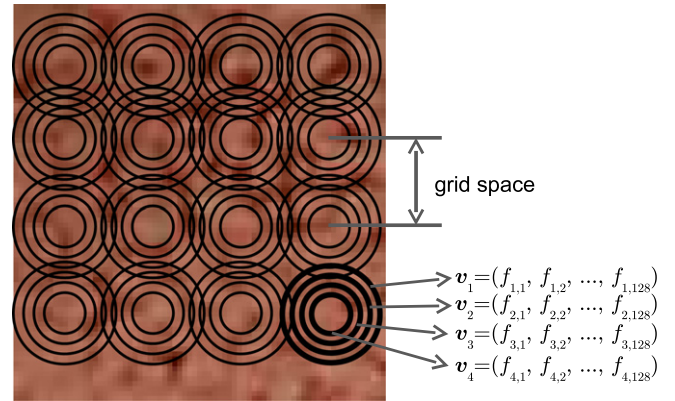
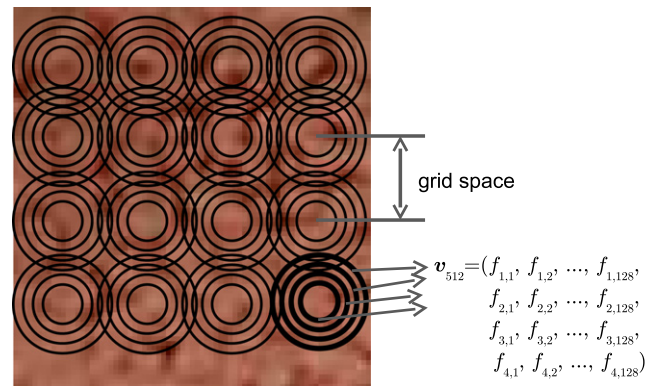


Fig. 9. Scale in gridSIFT.



(a)



(b)

Fig. 10. gridSIFT descriptor. (a) gridSIFT. (b) Multi-scale gridSIFT.

dataset used in the experiments have been taken by a randomly rotated endoscope and trimmed to different sizes, spatial information in NBI images are less informative than those in images used for category recognition.

Orientation and scale invariances of local descriptors are useful for matching images of the same scene which have been rotated or scaled. Those invariances are generally useful for recognition tasks and therefore merit further investigation in a future work, but have not been used in this paper. Random sampling (Nowak et al., 2006) is not used, since it also seems to be less effective for randomly rotated and shifted texture images, like the NBI images.

4.1.3. DiffSIFT

By a simple extension of gridSIFT, we propose a new combination of SIFT descriptors and grid sampling, which we call *DiffSIFT*. DiffSIFT is inspired by Relational Binarized HOG (Matsushima et al., 2010) in which Histograms of Oriented Gradients (HOG) (Dalal and Triggs, 2005) of two different regions are subtracted and binarized. Similarly, DiffSIFT is computed by the difference between two gridSIFT descriptors at adjacent grid points (Fig. 11). If multiple scales are used, subtraction is done for descriptors with the same scale. Note that the orientations of the SIFT descriptors of DiffSIFT are also fixed.

We refer as horizontal DiffSIFT (hDiffSIFT) the method which uses subtraction of horizontally adjacent descriptors, and vertical DiffSIFT (vDiffSIFT) when subtracting vertically adjacent descriptors. When both descriptors are used at the same time, we call them hvDiffSIFT descriptors.

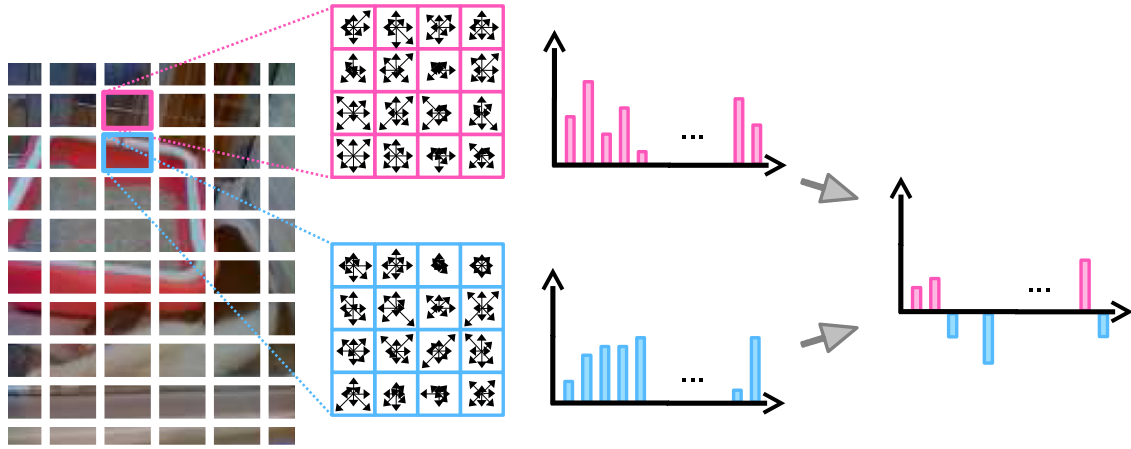


Fig. 11. DiffSIFT descriptor computed as the difference between two SIFT descriptors at adjacent grid points.

4.2. Visual word representation

4.2.1. Number of visual words

The number of visual words (or vocabulary size, codebook size) is a crucial parameter for better performance and needs to be tuned for each application: for example, 200 to 400 for natural scene categorization (Lazebnik et al., 2006), 6000 to 10,000 for object retrieval (Sivic and Zisserman, 2003), 10,000 to 1,200,000 for landmark object retrieval (Philbin et al., 2007). It has been reported (Philbin et al., 2007; Nowak et al., 2006; Zhang et al., 2007) that increasing the size improves performance, however, the performance has a peak at a certain size and appears flat or even degrades for larger sizes.

The best size of the vocabulary obviously depends on the size of the dataset, while a relatively large number of visual words is necessary in general. Since NBI images have never been explored in terms of visual words, it is necessary to study the effect which vocabulary size has on recognition performance. When the vocabulary size is huge, the computational cost for clustering is very high. To deal with a large vocabulary size, here we use two methods: hierarchical k -means clustering and class-wise concatenation of visual words.

4.2.2. Hierarchical k -means clustering

Hierarchical k -means can be used for clustering of large-size data and in (Nister and Stewenius, 2006) has been used for vocabulary tree construction. In our experiments, we need to cluster several millions of features densely sampled from almost one thousand training images, and explore vocabularies of size up to several thousands visual words. Hence, the use of hierarchical k -means is necessary for reducing the computational cost for clustering.

4.2.3. Class-wise concatenation of visual words

Zhang et al. (2007) create *global visual words* by concatenation of the visual words of each class (Fig. 12). Instead of clustering together all features from all classes, they perform k -means clustering separately for each class to form a class-wise vocabulary. Then a visual word representation of an image is made by concatenating the representations for each class-wise vocabulary. Similarly, it is possible to combine multi-level vocabularies by using different number of clusters (Quelhas and Odobez, 2007) to form a global visual word representation. In this case, results show that a performance peak is reached for a certain vocabulary size, and using multi-level vocabularies leads to improved performance.

We use a concatenation of vocabularies for three classes (corresponding to types A, B, and C3 NBI images) to form a global visual

word representation. Consequently, the dimensionality of the histograms representing the images is between $3 \times 2^2 = 12$ and $3 \times 2^{13} = 24,576$. In our case, however, the computational cost for vocabulary construction for each class is reduced to $O(\frac{n}{3}kd)$ on average, resulting in a significant reduction of the computational cost compared with the global vocabulary case. In the cross validation dataset used in the experiments, the total number of gridSIFT features is 4,123,706 (with grid spacing of 10 pixels), which requires about 2 GB memory just for storing all features, and even more for clustering when all features are used for constructing a global vocabulary. The class-wise concatenation of visual words requires much smaller amount of memory for storing features: in the dataset used, type A had 1,423,841, type B 2,148,225, and type C3 551,640 features.

4.3. Classifiers

4.3.1. SVM

The Support Vector Machine (SVM) classifier is used for classifying a two-class problem by using slack variables (soft margin) and nonlinear kernels (the kernel trick). Interested readers can find more details in textbooks like (Steinwart and Christmann, 2008; Schölkopf and Smola, 2002; John Shawe-Taylor, 2000; Vapnik, 1998).

4.3.2. Kernel types

We use the following five kernels for SVM: Radial Basis Function (RBF), linear, χ^2 , and histogram intersection (HI):

$$k_{RBF}(\mathbf{x}_1, \mathbf{x}_2) = \exp(-\gamma \|\mathbf{x}_1 - \mathbf{x}_2\|^2), \quad (3)$$

$$k_{linear}(\mathbf{x}_1, \mathbf{x}_2) = \mathbf{x}_1^T \mathbf{x}_2, \quad (4)$$

$$k_{\chi^2}(\mathbf{x}_1, \mathbf{x}_2) = \exp\left(-\frac{\gamma}{2} \sum_i \frac{(x_{1i} - x_{2i})^2}{x_{1i} + x_{2i}}\right), \quad (5)$$

$$k_{\chi^2_{sig}}(\mathbf{x}_1, \mathbf{x}_2) = -\sum_i \frac{(x_{1i} - x_{2i})^2}{x_{1i} + x_{2i}}, \quad (6)$$

$$k_{HI}(\mathbf{x}_1, \mathbf{x}_2) = \sum_i \min(x_{1i}, x_{2i}), \quad (7)$$

where γ is a scaling parameter which should be tuned for each problem.

The RBF kernel (or Gaussian kernel) is the most commonly used kernel, while Zhang et al. (2007) reported that the χ^2 kernel performs best. Note that there are some variations of the χ^2 kernel. Here we use a RBF type kernel χ^2 (Chapelle et al., 1999; Fowlkes et al., 2004) and a negative type kernel χ^2_{sig} (Schölkopf and Smola, 2002; Haasdonk and Bahlmann, 2004), while still there are some

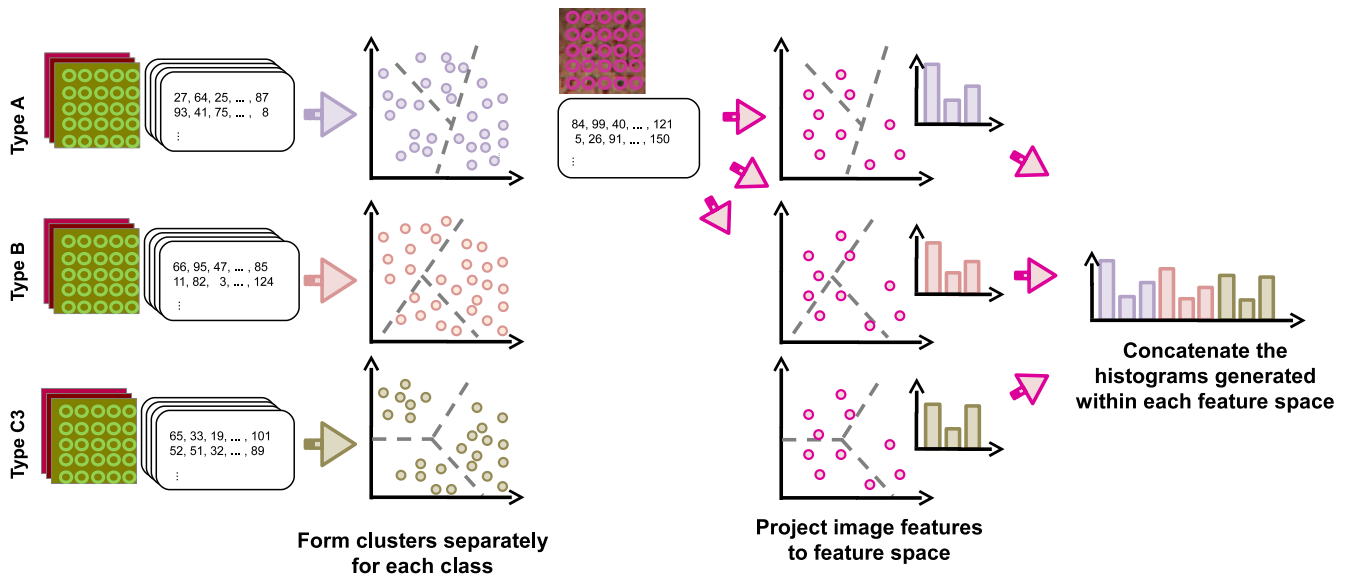


Fig. 12. Concatenating the visual words for each class in the training set.

other variants (Vedaldi and Zisserman, 2012). The histogram intersection kernel is also used for histogram classification (Swain and Ballard, 1991; Grauman and Darrell, 2005; Lazebnik et al., 2006) as well as the χ^2 kernel. Although the linear kernel does not provide any nonlinearity for SVMs, it has the advantage that it can be computed very fast and no kernel parameters are involved. Note that a class of kernels including the histogram intersection can be also computed as fast as a linear kernel by explicitly embedding vectors in a feature space, as shown by Maji et al. (2008).

Scaling (or standardization, normalization) of the feature vectors is also one of the important factors affecting performance (Graf and Borer, 2001; Hsu et al., 2003; Hastie, 2009). We normalize training and test data; i.e., each element in feature vectors is linearly scaled into $[-1, +1]$ for linear and RBF kernels, but not for histogram-based (χ^2 , χ_{ng}^2 and histogram intersection) kernels because this scaling affects the counts in the bins of a histogram.

In our experiments, an SVM classifier is trained for different values of the regularization parameter² C and the scaling parameter γ , by using 5-fold cross validation on the training set. The ranges of the parameter values are set based on preliminary experiments to $C = 2^{-3}, 2^{-1}, \dots, 2^{19}$ and $\gamma = 2^{-19}, 2^{-17}, \dots, 2^0, 2^1$, then the values that give the best recognition rate are chosen.

4.3.3. Multi-class classification

To handle the three-class problem, the one-against-one strategy is employed. There are two main strategies for extending a two-class classifier such as SVM to a multi-class classifier: one-against-all (or one-versus-all, one-versus-the-rest) and one-against-one (one-versus-one).

One-against-all (Vapnik, 1998; Lee et al., 2001; Weston and Watkins, 1999) is to train k SVM classifiers with one class as a positive training set, and all other classes as a negative training set. This strategy has the problem that the dataset sizes for each SVM training are unbalanced. *One-against-one* (Platt et al., 2000; Allwein et al., 2000) is to train two-class SVM classifiers for each pair of classes and then combine the results from all classifiers. The disadvantage of this strategy is that the number of classifiers is proportional to the square of the number of classes c : i.e., $\frac{c(c-1)}{2}$.

² This parameter controls the trade-off between the margin and the slack variables' penalties. See Steinwart and Christmann (2008), for example.

We employ the one-against-one strategy because both strategies seem to achieve similar performance (Hsu and Lin, 2002; Milgram et al., 2006), and for the three-class problem both strategies result in training the same number of classifiers. Also, we would like to avoid the unbalancedness of the one-against-all strategy because the dataset used in our experiments is already unbalanced (explained below). Currently we do not use any weights for the SVM classifier to deal with the unbalancedness, leaving this problem for future research.

5. Experimental Results

In this section we report the experimental results obtained by the proposed system. In the following several subsections, we explain the dataset, the evaluation measures, experimental settings and classification results for each experiment.

5.1. Dataset

We have collected 908 NBI images (as shown in Table 3) of colorectal tumors as a dataset for n -fold cross validation (we use $n = 10$ which is a typical choice Hastie, 2009). Examples of images in the dataset are shown in Fig. 5. Note that each image corresponds to a colon polyp obtained from a different subject: no images share the same tumor. In addition, we have collected another 504 NBI images as a separated test dataset. The cross validation dataset was collected before April 2010, while the test dataset after May 2010. This is similar to a practical situation in which samples collected at a certain period are used for training, and samples taken after that are then classified.

Every image was collected during an NBI colonoscopy examination. The instruments used were a magnifying videoendoscope system CF-H260AZ/I (Olympus Optical Co, Ltd, Tokyo, Japan), which provides up to 75x optical magnification. Then the images were digitized into 1440×1080 pixels and stored on an Olympus

Table 3
NBI image dataset.

Dataset	Type A	Type B	Type C3	Total
Cross validation	359	462	87	908
Test	156	294	54	504
Total	515	756	141	1412

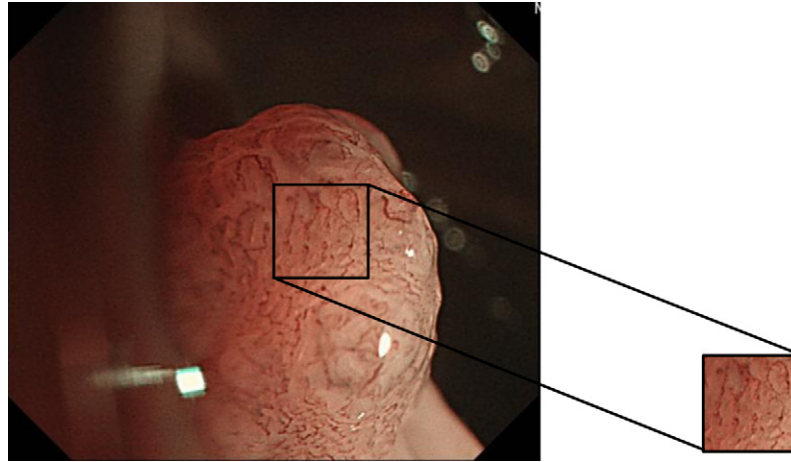


Fig. 13. Dataset construction by trimming a rectangle (right) from an NBI videoendoscope image (left).

EICP-D HDTV recorder. Care was taken to ensure that the lighting conditions, zooming and optical magnification were kept as similar as possible across different images. Therefore, the microvessel structures on the surface of the colorectal tumors can be regarded as being approximately of the same size in all images. The captured images were then trimmed by medical doctors and endoscopists to a rectangle so that the rectangle contains an area in which typical microvessel structures appear (see Fig. 13). For this reason, the size of the images is not fixed and varies from about 100×300 to 900×800 pixels. Image labels were provided by at least two medical doctors and endoscopists who are experienced in colorectal cancer diagnosis and familiar with pit-pattern analysis (Kudo et al., 1994, 1996; Imai et al., 2001) and NBI classifications (Kanao et al., 2009). Note that these images were collected not for these experiments, but for actual medical reporting (Aabakken, 2009) in case of surgery or comparison during a follow-up period, discussion with endoscopists, and informed consent with patients.

The study was conducted with an approval from the Hiroshima University Hospital ethics committee, and an informed consent was obtained from the patients and/or family members for the endoscopic examination.

5.2. Evaluation

In our experiments, 10-fold cross validation is used for evaluating the performance. For each experiment, recognition rate, recall, precision, and F -measure are calculated from a confusion matrix (Table 4) as follows:

$$\text{Recognition rate} = \frac{\sum_i m_{ii}}{\sum_{i,j} m_{ij}}, \quad (8)$$

$$\text{Recall}_j = \frac{m_{ji}}{\sum_i m_{ji}}, \quad (9)$$

$$\text{Precision}_j = \frac{m_{ji}}{\sum_j m_{ji}}, \quad (10)$$

$$F\text{-measure}_j = \frac{2\text{Recall}_j \times \text{Precision}_j}{\text{Recall}_j + \text{Precision}_j}, \quad (11)$$

where $i, j \in \{A, B, C\}$.

Table 4
Confusion matrix.

		Estimated category		
		Type A	Type B	Type C3
True category	Type A	m_{AA}	m_{AB}	m_{AC}
	Type B	m_{BA}	m_{BB}	m_{BC}
	Type C3	m_{CA}	m_{CB}	m_{CC}

Usually precision and recall rates are defined for two-class problems. We use those rates specific to each type. As we stated in Section 2, images of type C3 should be correctly identified because of the high correlation to SM-m cancers (Kanao et al., 2009). Therefore, the class-specific recall rate, in particular for type C3, is important to measure the performance of the system. In the following subsections, mainly results for recognition rates are shown in figures, with Recall_{C3} in tables. Recall, precision, and F -measure are available in the Supplemental material.

5.3. Implementation

For generating the gridSIFT, DiffSIFT, and DoG-SIFT descriptors, we use VLFeat (Vedaldi and Fulkerson, 2008), which is widely used for large image databases such as (Deng et al., 2009). For the hierarchical k -means clustering we use also the VLFeat implementation.

For the SVM classifiers, we use libSVM (Chang and Lin, 2011), a publicly available implementation of SVM, by adding our implementation of χ^2 and intersection kernels.

5.4. Experimental results

In the following three subsections we describe the results of 10-fold cross validation, the results on the test dataset, and some other additional experiments:

- 10-fold cross validation (Section 5.5)
 - class-wise concatenation of visual words
 - DoG-SIFT
 - different grid spacings
 - SVM kernels
 - different scales of gridSIFT descriptors
 - multi-scale gridSIFT
 - DiffSIFT
- Results on the test dataset (Section 5.5)
 - SVM kernels
- Additional experiments (Section 5.7)
 - 5-class problem.

5.5. Results for the 10-fold cross validation

First we evaluate the performance of the system using a 10-fold cross validation (Hastie, 2009) on 908 NBI images, as shown in Table 3. For each experiment, the dataset is randomly divided into 10 folds of 90 images each (i.e., 8 images are randomly excluded).

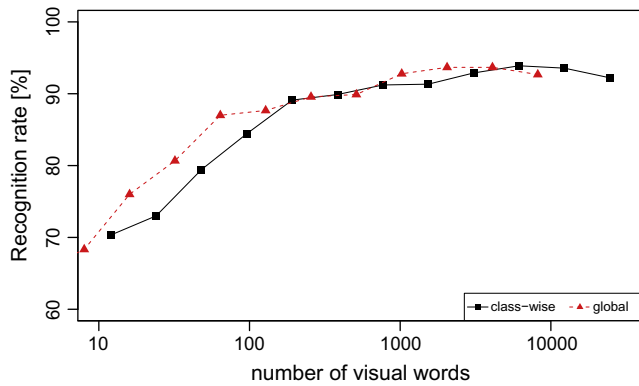


Fig. 14. Comparison between global and class-wise concatenated visual words. gridSIFT (grid spacing of 10 pixels, scale 5, 7, 9, 12 pixels), global or class-wise visual word concatenation, linear kernel SVM.

Table 5

Performance of gridSIFT with different visual word types (grid spacing of 10 pixels, scale 5, 7, 9, 12 pixels, linear kernel SVM).

Type	Recognition rate	VWs	Recall _{C3}
Global	93.67	2048	67.60
Class-wise	93.89	6144	64.71

At each fold, 810 images are used for training and 90 images for validation.

5.5.1. Visual words with and without class-wise concatenation

First we checked whether class-wise concatenation performs as well as the global vocabulary. Fig. 14 shows the performance curves for the two methods when increasing the number of visual words (VWs). Table 5 shows the maximum recognition rates, the number of visual words, and recall rate for type C3. We can see in Fig. 14 that both curves show similar performances, and the difference of maximum recognition rates is small. Therefore, we chose the use of class-wise concatenation of visual words for successive experiments for its smaller computational cost.

5.5.2. DoG-SIFT

Here we compare the performance of the gridSIFT descriptors with SIFT descriptors detected by DoG (DoG-SIFT). Fig. 15 and Table 6 show the performance of DoG-SIFT.

There are two parameters in DoG, for eliminating edges ($r = 10$ as a default value) and low contrast ($D_{th} = 0.03$), as described in Section 4.1.1. These parameters control the number of detected keypoints and might affect the performance, because more keypoints usually lead to better performance. When the low contrast parameter D_{th} is set to a very small value (0.003 or 0.006) compared to the default value (0.03), there is no significant difference between the results with $r = 5, 10$ ($p = 0.097$), $r = 5, 15$ ($p = 0.734$) and $r = 10, 15$ ($p = 0.289$) for the results in Fig. 15 (top),³ hence the recognition rate is insensitive to the edge parameter r and the low contrast parameter D_{th} in this range. However, when D_{th} is larger than this range, no keypoints are detected in some training and test images and the performance is drastically degraded, as shown in Fig. 15 (bottom) for $D_{th} = 0.01$. Therefore, D_{th} must be set to a smaller value rather than the default value (0.03).

In the results obtained for DoG-SIFT, the following two observations can be made in comparison with gridSIFT. First, DoG-SIFT

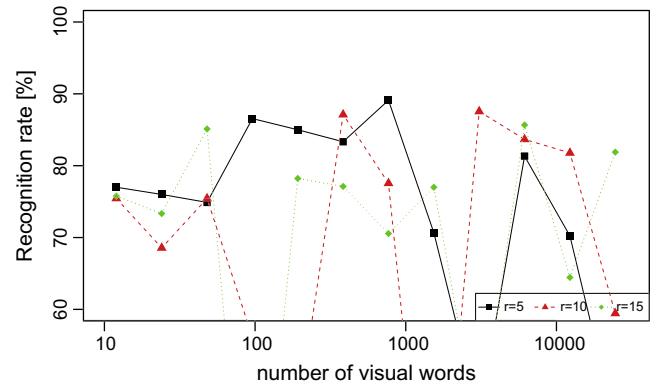
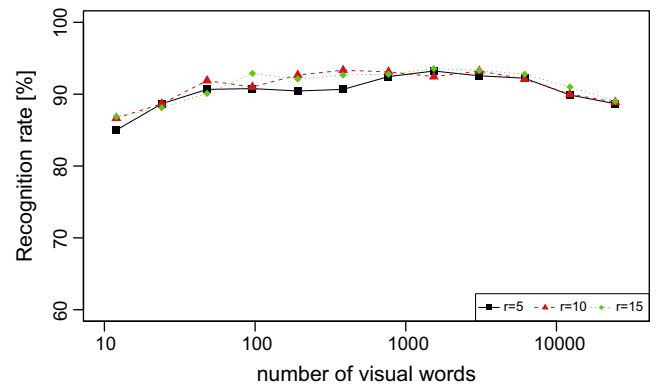
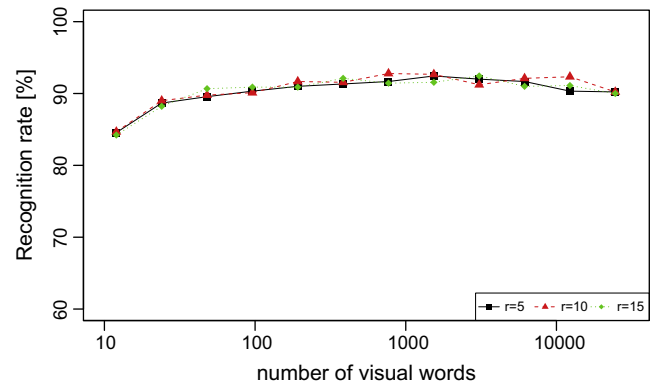


Fig. 15. Performance of DoG-SIFT with values of $r = 5, 10, 15$ for each contrast threshold D_{th} of (top) 0.003, (middle) 0.006, and (bottom) 0.01 (linear kernel SVM).

Table 6

Performance of DoG-SIFT with values of $r = 5, 10, 15$ (linear kernel SVM).

D_{th}	r	Recognition rate	VWs	Recall _{C3}
0.003	5	92.44	1536	58.14
	10	92.78	768	65.12
	15	92.44	3072	55.81
0.006	5	93.22	1536	68.60
	10	93.33	384	73.56
	15	93.55	1536	72.09
0.01	5	89.11	768	67.81
	10	87.56	3072	69.77
	15	85.67	6144	63.22

shows better performance when a small number of visual words is used: the recognition rate is 86.69% ($r = 15, D_{th} = 0.006$, VWs = 12), which is much better than the results for gridSIFT for the same number of visual words. Second, while the peak performance for DoG-SIFT is 93.55% ($r = 15, D_{th} = 0.006$), gridSIFT

³ The two-tailed paired t -test of $df = 11$ over 12 different sizes of vocabularies with 5% significance level ($\alpha = 0.05$) is used in the successive experiments. The null hypothesis (H_0) is that there is no statistically significant difference.

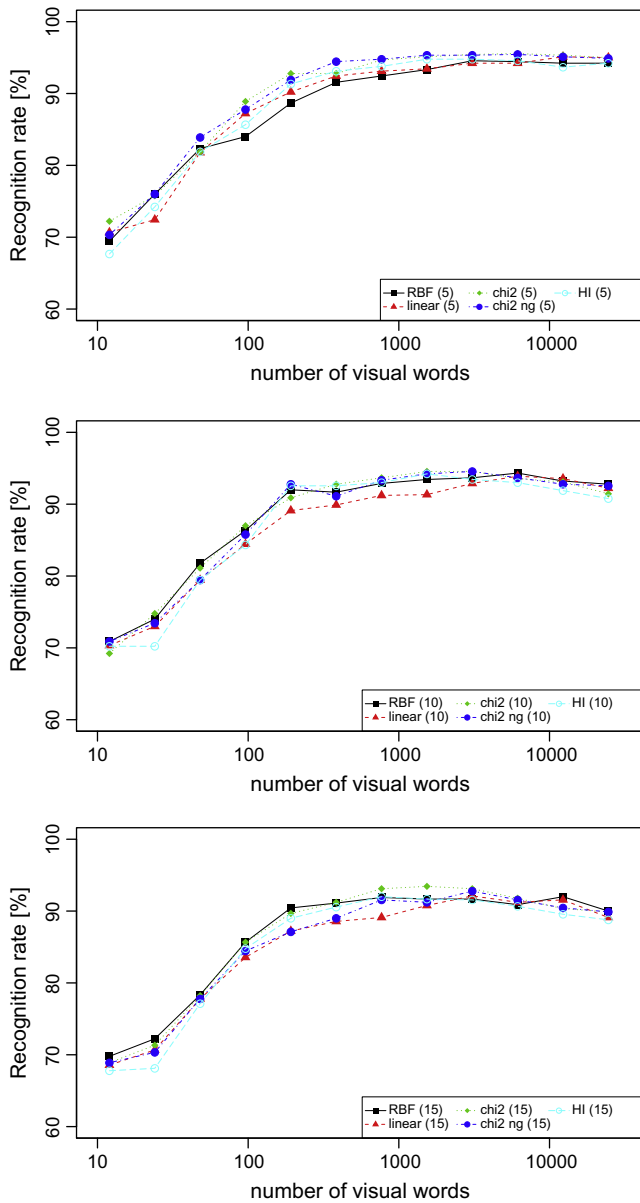


Fig. 16. Performance with different grid spacing and different kernel types. gridSIFT (grid spacing of (5), (10), (15) pixels, scale 5, 7, 9, 12 pixels), RBF, linear, χ^2 (chi2), χ^2_{ng} (chi2 ng), and histogram intersection (HI) kernel SVM.

performs better than that, as shown in the following experiments. Therefore, gridSIFT is preferable when a relatively large number of visual words are available. However, a smaller size may be obtained when orientation invariance is used (see Section 4.1.2).

5.5.3. gridSIFT with different grid spacing

Here we explore the effect of grid spacing on gridSIFT: with smaller spacing a lot of descriptors are extracted while with larger (hence sparse) spacing descriptors are fewer. We use a regular grid with spacing of 5, 10, and 15 pixels. For 5 pixels, the total number of descriptors is about 9 times larger than for spacing of 15 pixels.

Fig. 16 shows the performances of the SVM classifiers with RBF, linear, χ^2 , χ^2_{ng} , and HI kernels. Table 7 shows the maximum recognition rates, the number of visual words, and recall rate for type C3. The performance improves roughly 2% to 3% by using smaller spacing for each kernel type. Table 8 shows the results of statistical significance tests for each pair of different grid spacings. These results

Table 7

Performance of gridSIFT with different grid spacing and SVM kernel types (scale 5,7,9,12 pixels).

Kernel	Spacing	Recog. rate	VWs	Recall _{C3}
RBF	5	94.56	3072	75.58
	10	94.33	6144	64.37
	15	92.00	12,288	46.51
Linear	5	95.11	12,288	73.56
	10	93.89	6144	64.71
	15	92.11	3072	52.94
χ^2	5	95.56	6144	71.76
	10	94.56	1536	69.77
	15	93.44	1536	61.63
χ^2_{ng}	5	95.44	6144	74.71
	10	94.56	3072	68.60
	15	92.78	3072	53.49
HI	5	94.78	1536/3072	66.67/63.10
	10	94.22	1536	58.14
	15	92.00	768	50.00

support the observation of performance improvement for smaller spacing (except for the RBF kernel with grid spacing of 5 and 10 pixels). For types A and B, recall and precision rates do not appear to change significantly and good performance is achieved even for spacing of 15 pixels. In contrast, for type C3 the recall rate improves by 10% to 20% as spacing becomes smaller. The performance might be further improved by using spacing smaller than 5 pixels, however, we did not conduct such experiments because of the high computational cost due to the huge number of descriptors: already 16,268,314 features were used for spacing of 5 pixels, while there are 44,915,615 features for spacing of 3 pixels.

5.5.4. SVM kernels

As shown above, the performance difference for different SVM kernels is apparently small. Table 7 summarizes the performance difference for five SVM kernel types, also shown in Fig. 16 for grid spacing of 10 pixels. The χ^2 kernel performs best, but the difference is small. Recognition rate difference is at most 4.89% (for 96 visual words), and the maximum performances differ only about 1%. Table 8 shows the results of statistical significance tests for each pair of different kernels. For grid spacing of 5 pixels, where all kernels perform best as described above, RBF, linear and HI kernels are not significantly different, and χ^2 and χ^2_{ng} kernels are better than those.

We use the linear kernel for the successive experiments hereafter for reducing the computation time of the experiments. SVM with a linear kernel has much smaller computational cost while the nonlinear RBF and χ^2 kernels also need the scaling parameter γ to be selected by cross validation. χ^2_{ng} and HI kernels are also attractive alternatives because no such parameter is involved, however those are nonlinear.

5.5.5. Different combinations of multiple scales in gridSIFT

The scale, or size, of a patch for computing the gridSIFT descriptors affects the range of the local region involved. All experiments above used four different scales: 5, 7, 9, and 12 pixels. Here we explore different combinations of scales for examining which scales are most effective for the performance of the system.

Fig. 17 shows the performance for combinations of 1, 2, and 3 scales separately. The results seem to indicate that:

- using only a single scale results in a worse performance than when multiple scales are used. In particular, the smallest and the largest scales 3 and 12 perform worst.
- combinations of multiple scales lead to decreased performance when scale 12 is included.

Table 8
p-Values of significance tests for gridSIFT for each pair of different grid spacing and SVM kernel types (scale 5,7,9,12 pixels). Gray boxes show $p > 0.05$, which means no significant differences at the 5% significance level. "0.000" means $p < 0.5 \times 10^{-3}$.

	RBF			linear			χ^2		χ^2_{ng}			HI			
	5	10	15	5	10	15	5	10	5	10	15	5	10	15	
RBF	5	0.748	0.024	0.418	0.010	0.000	0.004	0.808	0.039	0.000	0.925	0.001	0.378	0.306	0.004
	10		0.000	0.493	0.001	0.000	0.000	0.928	0.001	0.000	0.446	0.000	0.570	0.044	0.000
	15			0.001	0.278	0.003	0.000	0.000	0.791	0.000	0.000	0.029	0.003	0.063	0.007
Linear	5				0.000	0.000	0.001	0.561	0.002	0.002	0.347	0.000	0.964	0.050	0.000
	10					0.000	0.000	0.015	0.475	0.000	0.014	0.002	0.008	0.538	0.016
	15						0.000	0.000	0.007	0.000	0.000	0.132	0.000	0.001	0.820
χ^2	5							0.000	0.000	0.953	0.000	0.007	0.000	0.000	0.000
	10								0.000	0.000	0.627	0.000	0.456	0.074	0.000
	15									0.000	0.000	0.005	0.002	0.049	0.000
χ^2_{ng}	5										0.000	0.000	0.001	0.000	
	10											0.000	0.356	0.045	0.000
	15												0.000	0.003	0.433
HI	5													0.047	0.000
	10														0.000
	15														0.000

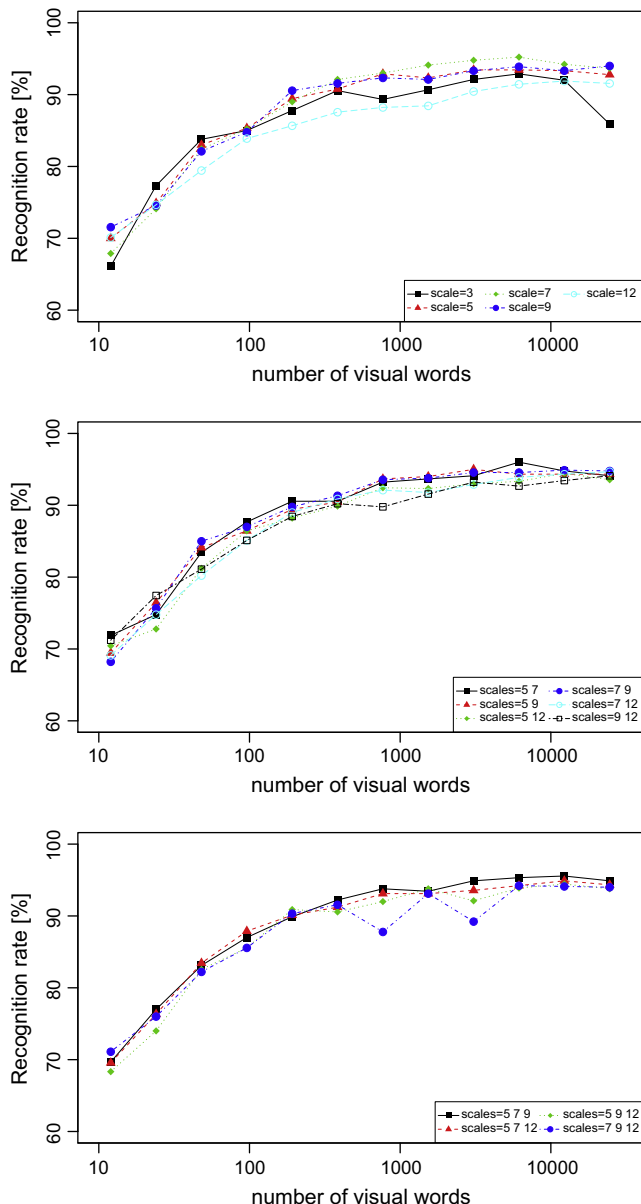


Fig. 17. Performance when single scales are used. gridSIFT (grid spacing of 5 pixels, [single, two, three] combinations of scales 3, 5, 7, 9 or 12 pixels, linear kernel SVM).

Table 9
Performance of gridSIFT with different combinations of scales (grid spacing of 5 pixels, linear kernel SVM).

Scale (s)	Recognition rate	VWs	Recall _{C3}
3	92.89	6144	60.71
5	93.44	3072	68.24
7	95.22	6144	69.41
9	94.00	24,576	56.98
12	91.89	12,288	61.90
(5,7)	96.00	6144	77.01
(5,9)	95.00	3072	76.47
(5,12)	94.44	12,288	63.53
(7,9)	94.89	12,288	72.41
(7,12)	94.67	24,576	65.12
(9,12)	94.11	24,576	68.24
(5,7,9)	95.33	6144	75.58
(5,7,12)	93.56	3072	75.86
(5,9,12)	94.44	12,288	71.26
(7,9,12)	94.11	12,288	70.59
(5,7,9,12)	95.11	12,288	73.56

- in contrast, performance is improved when the combination includes scale 7.

Table 9 shows the maximum recognition rates, the number of visual words, and type C3's recall rate. Using only scale 7 achieves visual recognition rate of 95%, and scales 5 and 9 also achieve more than 93%, hence each of those scales is expected to contribute to a better performance. However, type C3 recall rates for single scales are not better than those for multiple scales. When multiple scale combinations include scale 12, type C3 recall rates are degraded to less than 70% and recognition rates are also slightly degraded. The best performance was obtained by the combination of scales 5 and 7, with recognition rate of 96% and type C3 recall rate of 77.01%. Table 10 shows the results of statistical significance tests for each pair of different combination of scales, and these results support the observations listed above. While the best performance was obtained by the combination (5,7), this is not significantly different with a single scale of 7 or combinations (5,9), (7,9), (5,7,9), (5,7,12), and (7,9,12). In terms of computation time, a single scale is preferable because fewer features are used, while combinations of two or more scales can still be used.

5.5.6. Multi-scale gridSIFT

When multiple scales are used for computing the gridSIFT descriptors, multiple descriptors for a single point are obtained. For gridSIFT, those descriptors are then individually clustered. Here

Table 10

p-Values of significance tests for gridSIFT for each pair of different scale combinations (grid spacing of 5 pixels, linear kernel SVM). Gray boxes show $p > 0.05$, which means no significant differences at the 5% significance level. "0.000" means $p < 0.5 \times 10^{-3}$.

	5	7	9	12	(5,7)	(5,9)	(5,12)	(7,9)	(7,12)	(9,12)	(5,7,9)	(5,7,12)	(5,9,12)	(7,9,12)	(5,7,9,12)
3	0.051	0.037	0.067	0.333	0.008	0.006	0.206	0.005	0.176	0.155	0.002	0.006	0.082	0.169	0.239
5		0.341	0.396	0.000	0.001	0.002	0.357	0.008	0.476	0.534	0.000	0.004	0.954	0.750	0.074
7			0.759	0.002	0.102	0.135	0.148	0.063	0.131	0.335	0.007	0.217	0.359	0.478	0.002
9				0.000	0.014	0.132	0.167	0.136	0.169	0.257	0.014	0.114	0.625	0.477	0.054
12					0.000	0.000	0.001	0.000	0.001	0.000	0.000	0.000	0.002	0.002	0.003
(5,7)								0.734	0.003	0.017	0.641	0.418	0.009	0.051	0.000
(5,9)								0.010	0.012	0.032	0.095	0.938	0.033	0.157	0.000
(5,12)									0.012	0.858	0.971	0.001	0.004	0.416	0.891
(7,9)										0.013	0.053	0.252	0.647	0.011	0.130
(7,12)											0.928	0.000	0.004	0.460	0.951
(9,12)												0.003	0.024	0.607	0.887
(5,7,9)													0.048	0.003	0.042
(5,7,12)														0.014	0.099
(5,9,12)															0.692
(7,9,12)															0.098
															0.597

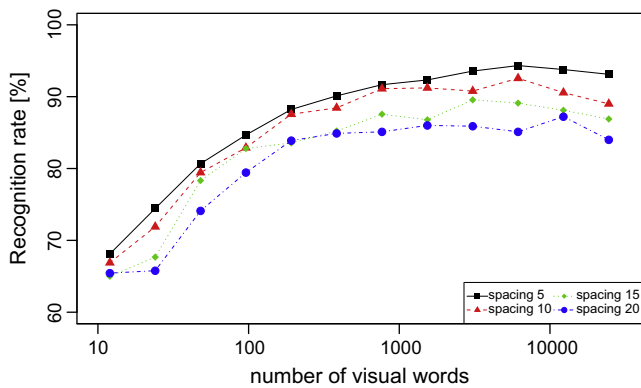


Fig. 18. Performance of multi-scale gridsift with different grid spacing. gridsift (grid spacing of 5, 10, 15 pixels, scale 5, 7, 9, 12 pixels, linear kernel SVM).

Table 11

Performance of gridsift and multi-scale gridsift (grid spacing of 5 pixel, linear kernel SVM, scale 5, 7, 9, 12 pixels, linear kernel SVM).

	Recognition rate	VWs	Recall _{C3}
gridsift	95.11	12,288	73.56
Multi-scale gridsift	94.33	6144	64.71

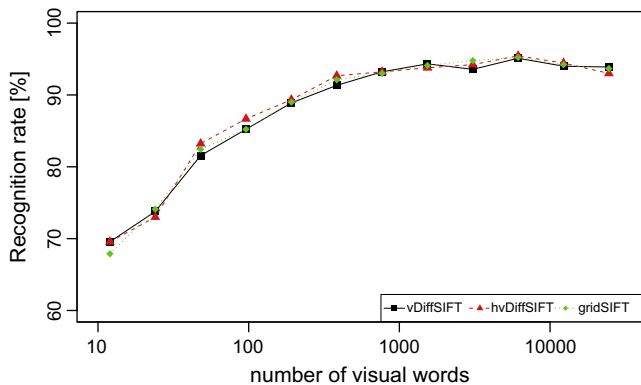


Fig. 19. Performance comparison between DiffSIFT and gridsift on the cross-validation dataset. DiffSIFT and gridsift (grid spacing of 5 pixels, scale 7 pixels, linear kernel SVM).

Table 12

Performance comparison between DiffSIFT and gridsift. (grid spacing of 5 pixels, scale 7 pixels, linear kernel SVM).

Feature	Recognition rate	VWs	Recall _{C3}
gridsift	95.22	6144	69.41
vDiffSIFT	95.11	6144	74.42
hvDiffSIFT	95.44	6144	71.26

we investigate an alternative way for constructing a descriptor of larger size by combining the features obtained at different scales into a single feature vector (multi-scale gridsift, see Section 4.1.2). We use four scales (5, 7, 9 and 12 pixels): four 128 dimensional gridsift feature vectors are concatenated to make a single 512 dimensional vector.

Fig. 18 shows the resulting performance. In Table 11 are given the maximum recognition rates, the number of visual words, and the recall rates for type C3. Overall, the performance of multi-scale gridsift follows similar trends to that of gridsift: performance improves as smaller grid spacing is used, no significant difference between different kernel types. A notable exception, however, can be seen in the decrease of the recall rate and *F*-measure for type C3. Therefore, multi-scale gridsift seems to be inferior to gridsift. This is also supported by the result of the significance test which is $p = 0.007 < 0.05$.

5.5.7. DiffSIFT

Here we show the results obtained with the DiffSIFT descriptors proposed in this paper. To evaluate the effectiveness of DiffSIFT as a descriptor, we use a single scale 7 and compare it with gridsift.

Fig. 19 compares the performance of DiffSIFT and gridsift using the same parameters, and Table 12 shows the maximum recognition rates, the number of visual words, and recall rate for type C3. The difference is slight and hvDiffSIFT improves the recognition rate only by 0.22% in comparison to gridsift. Table 13 shows the results of statistical significance tests and indicates no significant

Table 13

P-values of significance tests between gridsift, vDiffSIFT, and hvDiffSIFT (grid spacing of 5 pixels, scale 7 pixels, linear kernel SVM).

	vDiffSIFT	hvDiffSIFT
gridsift	0.608	0.352
vDiffSIFT		0.189

difference between DiffSIFT and gridSIFT. But vDiffSIFT improves type C3 recall rate by about 5%, as shown in Table 12.

5.6. Results on the test dataset

Next, we evaluate the performance on the test dataset as shown in Table 3. Now all of the 908 NBI images from the dataset used for the cross validation were used to construct the visual words and to train the classifiers. The additional dataset of 504 NBI images was used for the evaluation. Recall that the training dataset (used for the cross validation) and the test dataset were collected during different periods, as described in Section 5.1.

There are two reasons why we perform experiments with a separated test dataset in addition to the experiments with the cross validation dataset. First, separating a training dataset from a test dataset is similar to the practical situation in which samples collected at a certain period are used for training, and samples taken after that are then classified. Second, using all samples in a training dataset would lead to parameter tuning instead of validation. Therefore, many recent papers on machine learning use three kinds of datasets: training, validation, and test sets. A training set is used for parameter tuning, and those are evaluated by using a validation dataset. After the parameters are fixed, then a (novel) test set is used for evaluation of generalization. If we evaluate a method using only training and test sets, the resulting parameters would overfit both the training and test sets. Therefore, we used cross-validation for most experiments, and then we used fixed parameters (such as kernels, grid spacing, and scale combination) for the test set.

5.6.1. Performance of gridSIFT with different kernels

To see how gridSIFT works for the test (i.e., a novel) dataset, we use the parameter settings that give the best performance found in the cross validation experiments: spacing of 5 pixels and combination of two scales, 5 and 7.

Fig. 20 and Table 14 show the performance. A recognition rate of above 90% is achieved when the number of visual words is larger than 192 for most kernels and 1536 for the linear kernel, although this is lower than the rates obtained by cross validation.

Overall, RBF and HI kernels appear to perform better than the linear kernel as type C3 recall rate is improved. Table 15 shows the results of statistical significance tests. RBF, χ^2 and χ_{ng}^2 kernels perform better than the linear kernel while the performance of the HI kernel is not significantly different from the linear kernel.

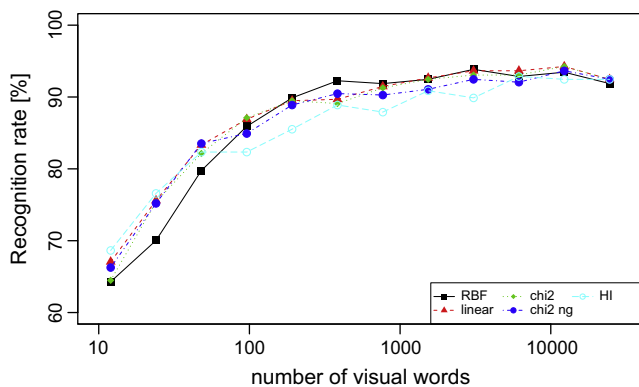


Fig. 20. Performance of gridSIFT with different kernels on the test dataset. gridSIFT (grid spacing of 5 pixels, two scales 5 and 7 pixels), RBF, linear, χ^2 (chi2), χ_{ng}^2 (chi2 ng), and histogram intersection (HI) kernel SVM.

Table 14

Performance of gridSIFT with different kernels on the test dataset (grid spacing of 5 pixels, scale 5, 7 pixels).

Kernel	Recognition rate	VWs	Recall _{C3}
RBF	93.65	12,288	66.67
Linear	93.45	12,288	62.96
χ^2	94.25	12,288	61.11
χ_{ng}^2	94.25	12,288	61.11
HI	93.85	3072	68.52

Table 15

P-values of significance tests for gridSIFT for different kernels on the test dataset (grid spacing of 5 pixels, scales 5, 7 pixels).

	Linear	χ^2	χ_{ng}^2	HI
RBF	0.034	0.009	0.492	0.731
Linear		0.001	0.027	0.340
χ^2			0.045	0.140
χ_{ng}^2				0.430

5.7. Other experiments

5.7.1. Preliminary results for the 5-class problem

As mentioned before, the NBI magnification findings classify NBI images into five types: A, B, C, C1, and C3 (Fig. 4 Kanao et al., 2009). We have used only three types A, B and C3 due to the on-going arguments on the classification and the ambiguity between B, C1, C2 and C3 (Table 1). However, we think that it is also worthwhile to tackle the five-class problem because this would demonstrate the ability of the BoW classification scheme, and also might contribute to a further development of classification schemes in the medical research field from the side of the engineering field.

To the cross validation dataset (Table 5), we add 215 images of type C1 and 71 images of type C2. Totally 1194 images were used and evaluated by 10-fold cross validation: each of the 10 folds had 119 images, and four images were randomly excluded. We used gridSIFT with parameters that performed well for the three-class problem.

Fig. 21 shows the performance for the five-class problem. recognition rate is at most 75%, which is lower by 20% than the three-class problem. Type A retains good recall rates, while type B and C decrease recall rates by 15% and 10%, respectively. Recall rates for types C1 and C2 are 45% and 10%, which means that most images of those types are classified to different classes.

The NBI magnification findings discriminate between types C1 and C2 by considering vessel irregularity, diameter, and

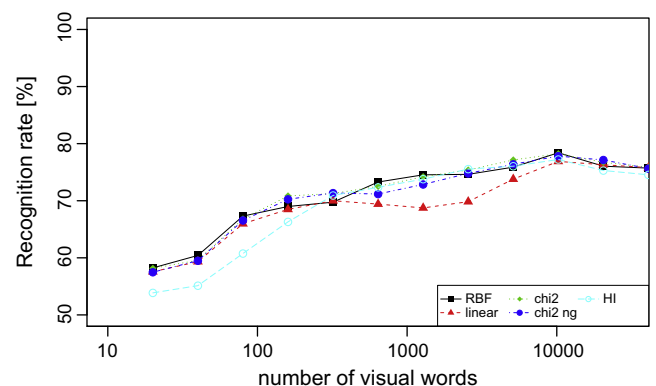
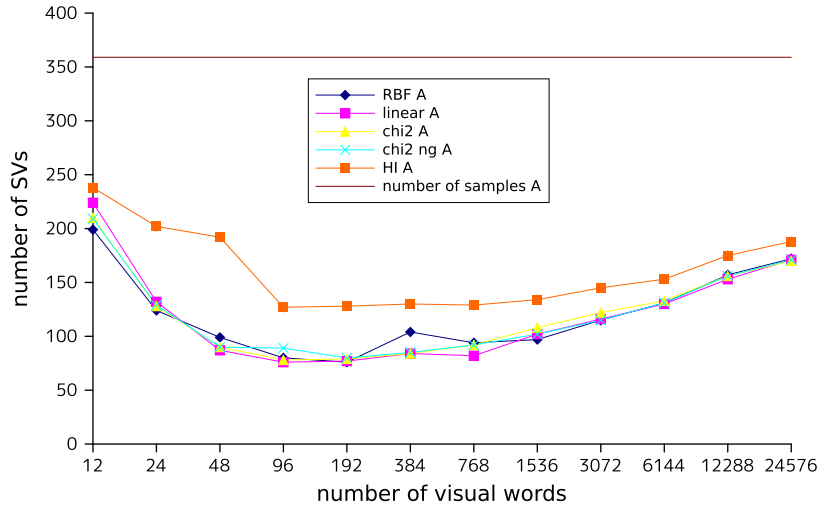
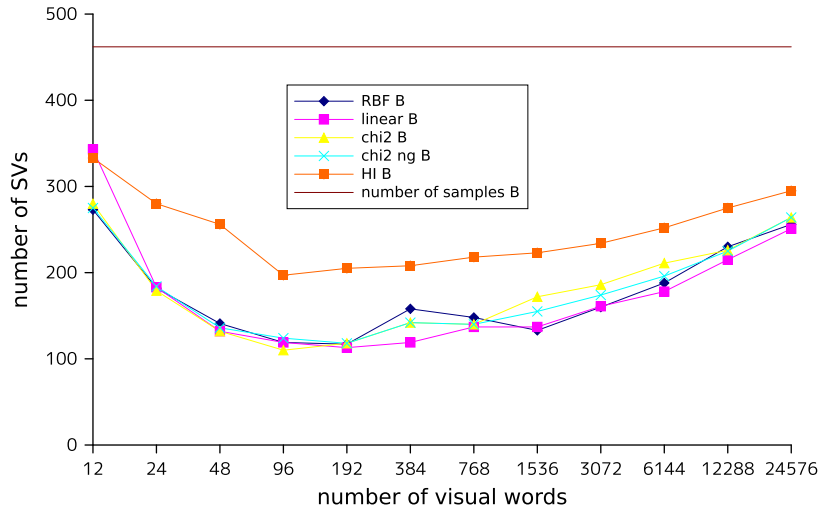


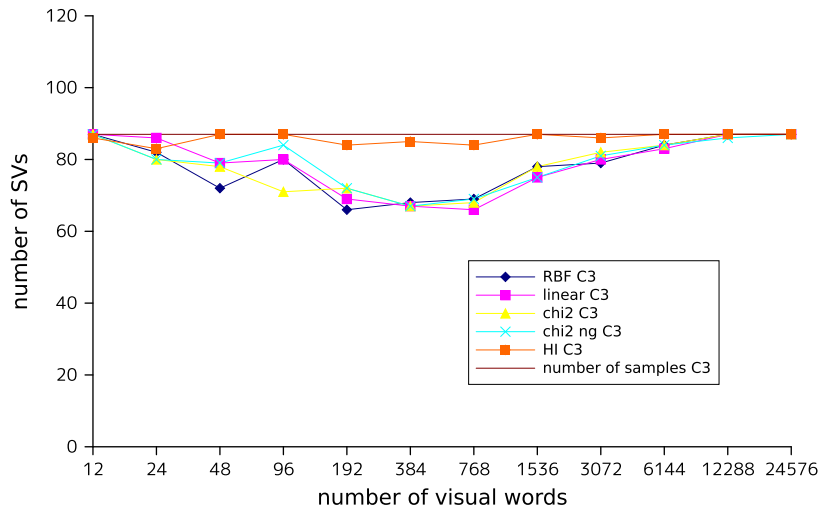
Fig. 21. Performance for the 5-class problem. gridSIFT (grid spacing of 5 pixels, scale 5, 7 pixels, different kernels SVMs).



(a)



(b)



(c)

Fig. 22. Number of support vectors for each class. (a) Type A (359 samples). (b) Type B (462 samples). (c) Type C3 (87 samples). gridSIFT (grid spacing of 5 pixels, scale 5, 7 pixels) on the whole CV dataset.

homogeneity (Fig. 4, Kanao et al., 2009). Therefore, for describing such texture properties, texon-based features (Winn et al., 2005; Shotton et al., 2008) might be better than an intensity gradient-based descriptor such as SIFT, because those subtypes C1, C2, and C3 are defined using the degree of irregularity of the observed microvessel network (see Fig. 4). Obviously the SIFT descriptor does not capture such information.

6. Summary and discussions

In this paper, we focused on a recent recognition framework based on local features, bag-of-visual-words (BoW), and provided extensive experiments on a variety of technical aspects, related to its application in the context of colorectal tumor classification in NBI endoscopic images. The prototype system used in the experiments consisted of a bag-of-visual-words representation of local features followed by Support Vector Machine (SVM) classifiers. Extensive experiments with varying the parameters for each component were needed, for the performance of the system is usually affected by those parameters, as summarized below.

Visual words representation with class-wise concatenation (section 5.5.1) has been shown to achieve similar performance with the global vocabulary (Table 5), with much smaller computational cost (Section 4.2.3).

DoG-SIFT (Section 5.5.2) has shown better performance when a relatively small number of visual words were used (Fig. 15). Although the performance was insensitive to the number of visual words, as well as to SIFT parameters such as the edge threshold, the peak performance for DoG-SIFT was not better than that of gridSIFT, as shown in successive experiments.

The performance of gridSIFT has been explored with different grid spacing (Section 5.5.3) and different SVM kernel types (Section 5.5.4). The χ^2 kernel performed best, but the difference was small (Table 7) and the maximum performances differ only by about 1%. On the other hand, the performance improved roughly 2–3% by using smaller spacing for each kernel type (also Table 7).

Different combinations of four different scales (5, 7, 9, and 12 pixels) of gridSIFT also have been explored (Section 5.5.5). The best performance was obtained by the combination of scales 5 and 7 (Table 9). Type C3 recall rates for single scales were not better than those for multiple scales, which was degraded however when multiple scale combinations included too large a scale (i.e., scale 12). The use of multi-scale gridSIFT resulted in the decrease of the recall rate and *F*-measure for type C3 (Section 5.5.6), while the performance of multi-scale gridSIFT followed similar trends to that of gridSIFT (Fig. 18).

The DiffSIFT descriptors, proposed in this paper, improved type C3 recall rate by about 5% (Table 12), while the difference between the maximum recognition rates was small.

For the test dataset (Section 5.6), recognition rates of above 90% were achieved by using gridSIFT (Fig. 20). For a five-class dataset (Fig. 21), the recall rates for types C1 and C2 were quite poor, indicating that texon-based features might be better than an intensity gradient-based descriptor, such as SIFT, for describing the texture properties of C1 and C2 (Fig. 4, Kanao et al., 2009).

It can be concluded that gridSIFT and linear kernel SVM seem to be sufficient for a practical system when small grid spacings (i.e., 5 and 7) and relatively large number of visual words (roughly 6000 to 10,000) are used, as performance peaks (above 95%) were reached for vocabularies of that size. These sizes are somewhat moderate compared with sizes reported for other systems (see Section 4.2.1). However, a question arises whether vocabularies of such a large size (e.g., 10 times larger than the number of training samples) may tend to over-train. This means that almost all of the training samples would be used as support vectors (SVs). To figure out the percentage of SVs compared to training samples, we counted the SVs when the entire CV dataset (908 samples in total) was used for training. Fig. 22 shows the number of SVs for each class. For Type A and B, the percentages of SVs are about 60% to 70% for vocabularies of larger size, while it increases to almost 100% for Type C3, obviously due to the unbalanced number of training samples. Interestingly, the HI kernel requires more SVs than other kernels, which may indicate over-training.

Also, the computation time for recognition is reasonably fast: about 60 ms for a test image (about 15 fps). This includes all steps, i.e., reading an image, extracting gridSIFT features, computing a visual word representation, and classification with linear kernel SVM. This enables us to develop a real time recognition/assessment system, which is expected to be of a great help for colonoscopy (Rex et al., 2011). Currently, we are developing a prototype system of NBI videoendoscopy by feeding a video sequence frame by frame to our prototype recognition system to classify images in real time during endoscopic examination of the colon. A video output of the prototype system is available as part of the supplemental material. Snapshots and outputs over 200 frames are shown in Fig. 23. A rectangular window of 120×120 size at the center of each frame is classified by the system proposed in this paper. The output of each frame is not a label but posterior probabilities of three classes (Huang et al., 2006). In Fig. 23, three colored temporal curves represent the posterior probabilities. The polyp in the video is diagnosed as Type B, and the posterior probability of Type B is larger

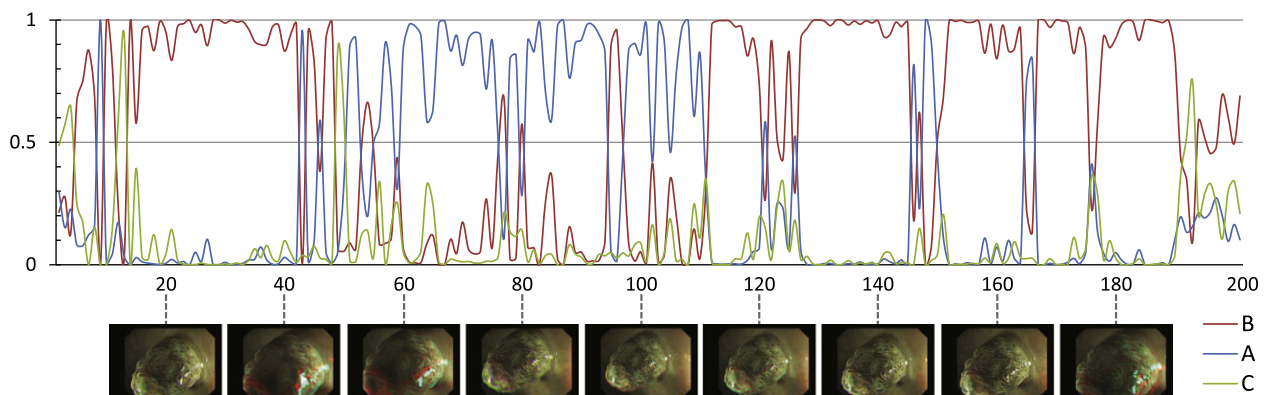


Fig. 23. A prototype system recognizing video sequences. Posterior probabilities for 200 frames are shown in different colored curves, with associated snapshots from the video. Because the polyp in the video is diagnosed as Type B, the posterior probability of Type B (red curve) should be close to 1 over the sequence (but currently not). Each frame is classified by the system described in this paper. Frame rate is about 17 fps (768 visual words, grid spacing of 5 pixels, scale 5, 7 pixels, and linear kernel SVM). The full sequence is available online as supplemental material.

than the others at about 60% of the 200 frames. This result is a very promising first step toward the final goal of achieving endoscopic diagnosis (Rex et al., 2011). At the same time, a huge number of factors need to be investigated further, such as the stability of the output, motion blur, focus, size of the window, color bleeding between frames due to the nature of NBI, highlight areas, and more. We are presently working on the evaluation of the prototype system.

Last but not least, our future work includes the prediction of histology by using the proposed method. While our emphasis in this paper is on the classification based on the NBI magnification findings for supporting a diagnosis by visual inspection, we already have preliminary but promising results for predicting histology as a two-class problem classifying NBI images into Type A and B-C3 (Takemura et al., 2012), and a three-class problem classifying chromoendoscopic pit-pattern images into TA, M/SM-s, and SM-m (Onji et al., 2011). We are currently planning more detailed experiments for validating the prediction ability of the system.

Acknowledgement

This work was supported in part by JSPS KAKENHI Grant Number 24591026.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at <http://dx.doi.org/10.1016/j.media.2012.08.003>.

References

- Aabakken, L., 2009. Reporting and Image Management. Wiley-Blackwell. chapter 20. Colonoscopy: Principles and Practice, 2 edition.
- Al-Kadi, O.S., 2010. Texture measures combination for improved meningioma classification of histopathological images. *Pattern Recognition* 43, 2043–2053.
- Allwein, E.L., Schapire, R.E., Singer, Y., 2000. Reducing multiclass to binary: a unifying approach for margin classifiers. *Journal of Machine Learning Research*, 113–141.
- Amores, J., 2010. Vocabulary-based approaches for multiple-instance data: a comparative study. *International Conference on Pattern Recognition*, 4246–4250.
- André, B., Vercauteren, T., Ayache, N., 2011a. Content-based retrieval in endomicroscopy: toward an efficient smart atlas for clinical diagnosis. In: *Proceedings of the MICCAI Workshop – Medical Content-based Retrieval for Clinical Decision (MCBR-CDS'11)*. Springer.
- André, B., Vercauteren, T., Buchner, A.M., Wallace, M.B., Ayache, N., 2011b. Retrieval evaluation and distance learning from perceived similarity between endomicroscopy videos. In: *Proceedings of Medical Image Computing and Computer Assisted Intervention (MICCAI)*. Springer, Heidelberg, pp. 297–304.
- André, B., Vercauteren, T., Buchner, A.M., Wallace, M.B., Ayache, N., 2011c. A smart atlas for endomicroscopy using automated video retrieval. *Medical Image Analysis* 15, 460–476.
- André, B., Vercauteren, T., Buchner, A.M., Wallace, M.B., Ayache, N., 2012. Learning semantic and visual similarity for endomicroscopy video retrieval. *IEEE Transactions on Medical Imaging* 31, 1276–1288.
- André, B., Vercauteren, T., Perchant, A., Wallace, M.B., Buchner, A.M., Ayache, N., 2009. Introducing space and time in local feature-based endomicroscopic image retrieval. In: *Proceedings of the MICCAI Workshop – Medical Content-based Retrieval for Clinical Decision (MCBR-CDS'09)*. Springer, pp. 18–30.
- Bank, S., Cobb, J., Burns, D., Marks, I., 1970. Dissecting microscopy of rectal mucosa. *The Lancet* 295, 64–65.
- Barabouthi, D.G., Wong, W.D., 2005. Clinical staging of rectal cancer. *Seminars in Colon and Rectal Surgery* 16, 104–116.
- Bay, H., Ess, A., Tuytelaars, T., Gool, L.V., 2008. Speeded-up robust features (SURF). *Computer Vision and Image Understanding* 110, 346–359.
- Bay, H., Tuytelaars, T., Van Gool, L., 2006. SURF: speeded up robust features. In: *Leonardis, A., Bischof, H., Pinz, A. (Eds.), Computer Vision – ECCV 2006, Lecture Notes in Computer Science*, vol. 3951. Springer, Berlin/Heidelberg, pp. 404–417.
- Beets, G.L., Beets-Tan, R.G., 2010. Pretherapy imaging of rectal cancers: ERUS or MRI? *Surgical Oncology Clinics of North America* 19, 733–741.
- Bishop, C., 2006. *Pattern Recognition and Machine Learning*. Springer, New York.
- Bosch, A., Zisserman, A., Muoz, X., 2007. Image classification using random forests and ferns. In: *IEEE 11th International Conference on Computer Vision*, 2007. ICCV 2007, pp. 1–8.
- Breier, M., Gross, S., Behrens, A., Stehle, T., Aach, T., 2011. Active contours for localizing polyps in colonoscopic NBI image data. In: *Proc. of Medical Imaging 2011: Computer-Aided Diagnosis*, pp. 79632M–79632M-10.
- Campbell, W., Sturim, D., Reynolds, D., Solomonoff, A., 2006. SVM based speaker verification using a GMM supervector kernel and NAP variability compensation. In: *2006 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2006. ICASSP 2006 Proceedings, p. 1.
- Cancer research UK, 2011. CancerStats, Incidence 2008 – UK. <<http://info.cancerresearchuk.org/cancerstats/incidence>>.
- Canon, C.L., 2008. Is there still a role for double-contrast barium enema examination? *Clinical Gastroenterology and Hepatology* 6, 389–392.
- Chang, C.C., Hsieh, C.R., Lou, H.Y., Fang, C.L., Tiong, C., Wang, J.J., Wei, I.V., Wu, S.C., Chen, J.N., Wang, Y.H., 2009. Comparative study of conventional colonoscopy, magnifying chromoendoscopy, and magnifying narrow-band imaging systems in the differential diagnosis of small colonic polyps between trainee and experienced endoscopist. *International Journal of Colorectal Disease* 24, 1413–1419.
- Chang, C.C., Lin, C.J., 2011. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* 2, 27:1–27:27.
- Chapelle, O., Haffner, P., Vapnik, V.N., 1999. Support vector machines for histogram-based image classification. *IEEE Transactions on Neural Networks* 10, 1055–1064.
- Chum, O., Philbin, J., Isard, M., Zisserman, A., 2007. Scalable near identical image and shot detection. In: *Proceedings of the 6th ACM international conference on Image and Video Retrieval*. ACM, New York, NY, USA, pp. 549–556.
- Classen, M., Tytgat, G.N.J., Lightdale, C.J., 2010. *Gastroenterological Endoscopy*. Thieme Medical Publisher, second ed.
- Cristianini, N., Shawe-Taylor, J., Lodhi, H., 2002. Latent semantic kernels. *Journal of Intelligent Information Systems* 18, 127–152.
- Csurka, G., Dance, C.R., Fan, L., Willamowski, J., Bray, C., 2004. Visual categorization with bags of keypoints, in: *European Conference on Computer Vision (ECCV2004) Workshop on Statistical Learning in Computer Vision*, pp. 59–74.
- Dalal, N., Triggs, B., 2005. Histograms of oriented gradients for human detection, in: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2005. CVPR 2005, pp. 886–893.
- Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L., 2009. ImageNet: a large-scale hierarchical image database. In: *IEEE Conference on Computer Vision and Pattern Recognition*, 2009. CVPR 2009, pp. 248–255.
- Farquhar, J., Szedmak, S., Meng, H., Shawe-Taylor, J., 2005. Improving bag-of-keypoints image categorisation: generative models and pdf-kernels. Technical report, Department of Electronics and Computer Science, University of Southampton.
- Fei-Fei, L., Perona, P., 2005. A bayesian hierarchical model for learning natural scene categories. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2005. CVPR 2005, pp. 524–531.
- Fowlkes, C., Belongie, S., Chung, F., Malik, J., 2004. Spectral grouping using the Nyström method. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 26, 214–225.
- Fu, K., Sano, Y., Kato, S., Fujii, T., Nagashima, F., Yoshino, T., Okuno, T., Yoshida, S., Fujimori, T., 2004. Chromoendoscopy using indigo carmine dye spraying with magnifying observation is the most reliable method for differential diagnosis between non-neoplastic and neoplastic colorectal lesions: a prospective study. *Endoscopy* 36, 1089–1093.
- Gaddam, S., Sharma, P., 2010. New trends in endoscopic imaging. *Gastroenterology & Endoscopy News* 8.
- van Gemert, J., Geusebroek, J.M., Veenman, C., Smeulders, A., 2008. Kernel codebooks for scene categorization. In: *Forsyth, D., Torr, P., Zisserman, A. (Eds.), Computer Vision – ECCV 2008, Lecture Notes in Computer Science*, vol. 5304. Springer, Berlin/Heidelberg, pp. 696–709.
- Gershman, G., Ament, M., 2012. *Practical Pediatric Gastrointestinal Endoscopy*. Wiley-Blackwell (second ed.).
- Gloor, F.J., 1986. The adenoma–carcinoma sequence of the colon and rectum. *Sozial- und Preventivmedizin/Social and Preventive Medicine* 31, 74–75.
- Gono, K., Obi, T., Yamaguchi, M., Ohshima, N., Machida, H., Sano, Y., Yoshida, S., Hamamoto, Y., Endo, T., 2004. Appearance of enhanced tissue features in narrow-band endoscopic imaging. *Journal of Biomedical Optics* 9, 568–577.
- Gono, K., Yamazaki, K., Doguchi, N., Nonami, T., Obi, T., Yamaguchi, M., Ohshima, N., Machida, H., Sano, Y., Yoshida, S., Hamamoto, Y., Endo, T., 2003. Endoscopic observation of tissue by narrowband illumination. *Optical Review* 10, 211–215.
- Gopalswamy, N., Newaz, S., Gianti, S., Bhutani, A., Markert, R., Swamy, L., 2000. Digital rectal examination as a part of colorectal cancer screening in hospitalized veterans. *The American Journal of Gastroenterology* 95, 2534–2535.
- Graf, A., Borer, S., 2001. Normalization in support vector machines. In: *Radig, B., Florczyk, S. (Eds.), Pattern Recognition, Lecture Notes in Computer Science*, vol. 2191. Springer, Berlin/Heidelberg, pp. 277–282.
- Grauman, K., Darrell, T., 2005. The pyramid match kernel: discriminative classification with sets of image features, in: *Tenth IEEE International Conference on Computer Vision*, 2005. ICCV 2005, vol. 2, pp. 1458–1465.
- Gross, S., Kennel, M., Stehle, T., Wulff, J., Tischendorf, J., Trautwein, C., Aach, T., 2009a. Polyp segmentation in NBI colonoscopy. In: *Meinzer, H.P., Deserno, T.M., Handels, H., Tolxdorff, T., Brauer, W. (Eds.), Bildverarbeitung für die Medizin 2009*. Springer, Berlin/Heidelberg, pp. 252–256 (Informatik aktuell).
- Gross, S., Stehle, T., Behrens, A., Auer, R., Aach, T., Winograd, R., Trautwein, C., Tischendorf, J., 2009b. A comparison of blood vessel features and local binary patterns for colorectal polyp classification. In: *Proc. of Medical Imaging 2008: Computer-Aided Diagnosis*, SPIE. pp. 72602Q–72602Q-8.

- Gunduz-Demir, C., Kandemir, M., Tosun, A.B., Sokmensuer, C., 2010. Automatic segmentation of colon glands using object-graphs. *Medical Image Analysis* 14, 1–12.
- Haasdonk, B., Bahlmann, C., 2004. Learning with distance substitution kernels. In: Rasmussen, C., Bulthoff, H., Scholkopf, B., Giese, M. (Eds.), *Pattern Recognition – Proc. of the 26th DAGM Symposium, Lecture Notes in Computer Science*, vol. 3175. Springer, Berlin/Heidelberg, pp. 220–227.
- Häfner, M., Gangl, A., Kwitt, R., Uhl, A., Vecsei, A., Wrba, F., 2009a. Improving pit-pattern classification of endoscopy images by a combination of experts. In: Yang, G.Z., Hawkes, D., Rueckert, D., Noble, A., Taylor, C. (Eds.), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2009, Lecture Notes in Computer Science*, vol. 5761. Springer, Berlin/Heidelberg, pp. 247–254.
- Häfner, M., Gangl, A., Liedlgruber, M., Uhl, A., Vecsei, A., Wrba, F., 2009b. Combining gaussian markov random fields with the discrete wavelet transform for endoscopic image classification. In: *Proceedings of the 16th International Conference on Digital Signal Processing*. IEEE Press, Piscataway, NJ, USA, pp. 177–182.
- Häfner, M., Gangl, A., Liedlgruber, M., Uhl, A., Vecsei, A., Wrba, F., 2009c. Pit pattern classification using extended local binary patterns. In: 9th International Conference on Information Technology and Applications in Biomedicine, 2009. ITAB 2009, pp. 1–4.
- Häfner, M., Gangl, A., Liedlgruber, M., Uhl, A., Vecsei, A., Wrba, F., 2009d. Pit pattern classification using multichannel features and multiclassification. In: Exarchos, T., Papadopoulos, A., Fotiadis, D. (Eds.), *Handbook of Research on Advanced Techniques in Diagnostic Imaging and Biomedical Applications*. IGI Global, Hershey, PA, USA, pp. 335–350.
- Häfner, M., Gangl, A., Liedlgruber, M., Uhl, A., Vecsei, A., Wrba, F., 2010a. Classification of endoscopic images using Delaunay triangulation-based edge features. In: Campilho, A., Kamel, M. (Eds.), *Image Analysis and Recognition. Lecture Notes in Computer Science*, vol. 6112. Springer, Berlin/Heidelberg, pp. 131–140.
- Häfner, M., Gangl, A., Liedlgruber, M., Uhl, A., Vecsei, A., Wrba, F., 2010b. Endoscopic image classification using edge-based features. In: *Proc. of 20th International Conference on Pattern Recognition (ICPR2010)*, IEEE, pp. 2724–2727.
- Häfner, M., Kendlbacher, C., Mann, W., Taferl, W., Wrba, F., Gangl, A., Vecsei, A., Uhl, A., 2006. Pit pattern classification of zoom-endoscopic colon images using histogram techniques. In: *Signal Processing Symposium, 2006. NORSIG 2006. Proceedings of the 7th Nordic*, pp. 58–61.
- Häfner, M., Kwitt, R., Uhl, A., Gangl, A., Wrba, F., Vecsei, A., 2009e. Feature extraction from multi-directional multi-resolution image transformations for the classification of zoom-endoscopy images. *Pattern Analysis & Applications* 12, 407–413.
- Häfner, M., Kwitt, R., Uhl, A., Wrba, F., Gangl, A., Vecsei, A., 2009f. Computer-assisted pit-pattern classification in different wavelet domains for supporting dignity assessment of colonic polyps. *Pattern Recognition* 42, 1180–1191.
- Häfner, M., Kwitt, R., Wrba, F., Gangl, A., Vecsei, A., Uhl, A., 2008. One-against-one classification for zoom-endoscopy images. In: 4th IET International Conference on Advances in Medical, Signal and Information Processing, 2008. MEDSIP 2008, pp. 1–4.
- Halligan, S., Taylor, S.A., 2007. CT colonography: results and limitations. *European Journal of Radiology* 61, 400–408.
- Harris, C., Stephens, M., 1988. A combined corner and edge detector. In: *Proc. of 4th Alvey Vision Conference (AVC1988)*, pp. 147–151.
- Hastie, T., 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, second ed. Springer, New York.
- Health Statistics and Informatics Department, World Health Organization, 2008. *Global Burden of Disease: 2004 update*. <<http://www.who.int/evidence/bod>>.
- Heitman, S.J., Au, F., Manns, B.J., McGregor, S.E., Hilsden, R.J., 2008. Nonmedical costs of colorectal cancer screening with the fecal occult blood test and colonoscopy. *Clinical Gastroenterology and Hepatology* 6, 912–917.
- Herve, N., Boujemaa, N., Houle, M.E., 2009. Document description: what works for images should also work for text? In: *Multimedia Content Access: Algorithms and Systems III*, SPIE, pp. 72550B–72550B-12.
- Higashi, R., Uraoka, T., Kato, J., Kuwaki, K., Ishikawa, S., Saito, Y., Matsuda, T., Ikematsu, H., Sano, Y., Suzuki, S., Murakami, Y., Yamamoto, K., 2010. Diagnostic accuracy of narrow-band imaging and pit pattern analysis significantly improved for less-experienced endoscopists after an expanded training program. *Gastrointestinal Endoscopy* 72, 127–135.
- Hirai, K., Kanazawa, Y., Sagawa, R., Yagi, Y., 2011. Endoscopic image matching for reconstructing the 3-D structure of the intestines. *Medical Imaging Technology* 29, 36–46.
- Hirata, M., Tanaka, S., Oka, S., Kaneko, I., Yoshida, S., Yoshihara, M., Chayama, K., 2007a. Evaluation of microvessels in colorectal tumors by narrow band imaging magnification. *Gastrointestinal Endoscopy* 66, 945–952.
- Hirata, M., Tanaka, S., Oka, S., Kaneko, I., Yoshida, S., Yoshihara, M., Chayama, K., 2007b. Magnifying endoscopy with narrow band imaging for diagnosis of colorectal tumors. *Gastrointestinal Endoscopy* 65, 988–995.
- Hofmann, T., 1999. Probabilistic latent semantic analysis. In: *15th Uncertainty in Artificial Intelligence*, pp. 289–296.
- Hsu, C.W., Chang, C.C., Lin, C.J., 2003. A practical guide to support vector classification. <<http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf>>.
- Hsu, C.W., Lin, C.J., 2002. A comparison of methods for multiclass support vector machines. *IEEE Transactions on Neural Networks* 13, 415–425.
- Huang, T.K., Weng, R.C., Lin, C.J., 2006. Generalized Bradley–Terry models and multi-class probability estimates. *Journal of Machine Learning Research* 4.
- Ignjatovic, A., East, J.E., Guenther, T., Hoare, J., Morris, J., Ragunath, K., Shonde, A., Simmons, J., Suzuki, N., Thomas-Gibson, S., Saunders, B., 2011. What is the most reliable imaging modality for small colonic polyp characterization? Study of white-light, autofluorescence, and narrow-band imaging. *Endoscopy* 43, 94–99.
- Ikematsu, H., Matsuda, T., Emura, F., Saito, Y., Uraoka, T., Fu, K.I., Kaneko, K., Ochiai, A., Fujimori, T., Sano, Y., 2010. Efficacy of capillary pattern type IIIA/IIIB by magnifying narrow band imaging for estimating depth of invasion of early colorectal neoplasms. *BMC Gastroenterology* 10, 33.
- Imai, Y., Kudo, S., Tsuruta, O., Fujii, T., Hayashi, S., Tanaka, S., Terai, T., 2001. Problems and clinical significance of v type pit pattern diagnosis: report on round-table consensus meeting. *Early Colorectal Cancer* 5, 595–613.
- Jenkinson, F., Steele, R., 2010. Colorectal cancer screening – methodology. *The Surgeon* 8, 164–171.
- Joachims, T., 1998. Text categorization with support vector machines: learning with many relevant features. In: *Nedellec, C., Rouveiro, C. (Eds.), Machine Learning: ECML-98, Lecture Notes in Computer Science*, vol. 1398. Springer, Berlin/Heidelberg, pp. 137–142.
- John C. Platt, Nello Cristianini, J.S.T., 2000. Large margin dags for multiclass classification. In: *Advances in Neural Information Processing Systems 12 (NIPS1999)*, pp. 547–553.
- John Shawe-Taylor, N.C., 2000. *An Introduction to Support Vector Machines: And Other Kernel-based Learning Methods*. Cambridge University Press, Cambridge, New York.
- Johnson, C., MacCarty, R.L., Welch, T.J., Wilson, L.A., Harmsen, W.S., Ilstrup, D.M., Ahlquist, D.A., 2004. Comparison of the relative sensitivity of ct colonography and double-contrast barium enema for screen detection of colorectal polyps. *Clinical Gastroenterology and Hepatology* 2, 314–321.
- Joutou, T., Yanai, K., 2009. A food image recognition system with multiple kernel learning. In: *16th IEEE International Conference on Image Processing (ICIP)*, 2009, pp. 285–288.
- Jurie, F., Triggs, B., 2005. Creating efficient codebooks for visual recognition. In: *Tenth IEEE International Conference on Computer Vision, 2005. ICCV 2005*, pp. 604–610.
- Kanao, H., 2008. Clinical significance of type VI pit pattern subclassification in determining the depth of invasion of colorectal neoplasms. *World Journal of Gastroenterology* 14, 211–217.
- Kanao, H., Tanaka, S., Oka, S., Hirata, M., Yoshida, S., Chayama, K., 2009. Narrow-band imaging magnification predicts the histology and invasion depth of colorectal tumors. *Gastrointestinal Endoscopy* 69, 631–636.
- Karkanis, S., Iakovidis, D., Maroulis, D., Karras, D., Tzivras, M., 2003. Computer-aided tumor detection in endoscopic video using color wavelet features. *IEEE Transactions on Information Technology in Biomedicine* 7, 141–152.
- Karl, J., Wild, N., Tacke, M., Andres, H., Garczarek, U., Rollinger, W., Zolg, W., 2008. Improved diagnosis of colorectal cancer using a combination of fecal occult blood and novel fecal protein markers. *Clinical Gastroenterology and Hepatology* 6, 1122–1128.
- Ke, Y., Sukthankar, R., 2004. PCA-SIFT: a more distinctive representation for local image descriptors. In: *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004*, pp. II-506–II-513.
- Kiesslich, R., Goetz, M., Vieth, M., Galle, P.R., Neurath, M.F., 2007. Technology insight: confocal laser endoscopy for in vivo diagnosis of colorectal cancer. *Nature Reviews Clinical Oncology* 4, 480–490.
- Koenderink, J., 1984. The structure of images. *Biological Cybernetics* 50, 363–370.
- Kosaka, T., 1975. Fundamental study on the diminutive polyps of the colon by mucosal stain and dissecting microscope. *Journal of Coloproctology* 28, 218–228.
- Kudo, S., Hirota, S., Nakajima, T., Hosobe, S., Kusaka, H., Kobayashi, T., Himori, M., Yagyu, A., 1994. Colorectal tumours and pit pattern. *Journal of Clinical Pathology* 47, 880–885.
- Kudo, S., Tamura, S., Nakajima, T., Yamano, H., Kusaka, H., Watanabe, H., 1996. Diagnosis of colorectal tumorous lesions by magnifying endoscopy. *Gastrointestinal Endoscopy* 44, 8–14.
- Kwitt, R., Uhl, A., 2007a. Modeling the marginal distributions of complex wavelet coefficient magnitudes for the classification of Zoom-Endoscopy images. In: *Proc. of ICCV2007*, IEEE, pp. 1–8.
- Kwitt, R., Uhl, A., 2007b. Multi-directional multi-resolution transforms for zoom-endoscopy image classification. In: *Kurzynski, M., Puchala, E., Wozniak, M., Zolnierok, A. (Eds.), Computer Recognition Systems 2, Advances in Soft Computing*, vol. 45. Springer, Berlin/Heidelberg, pp. 35–43.
- Kwitt, R., Uhl, A., 2008. Color eigen-subband features for endoscopy image classification. In: *IEEE International Conference on Acoustics, Speech and Signal Processing, 2008. ICASSP 2008*, pp. 589–592.
- Kwitt, R., Uhl, A., Häfner, M., Gangl, A., Wrba, F., Vecsei, A., 2010. Predicting the histology of colorectal lesions in a probabilistic framework. In: *Proc. of CVPR2010 Workshop on Mathematical Methods in Biomedical Image Analysis (MMBIA2010)*, IEEE, pp. 103–110.
- Lazebnik, S., Schmid, C., Ponce, J., 2006. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 2169–2178.
- Lee, Y., Lin, Y., Wahba, G., 2001. *Multicategory Support Vector Machines*. Technical Report. Department of Statistics, University of Madison.
- Lin, M., Wong, K., Ng, W.L., Shon, I.H., Morgan, M., 2011. Positron emission tomography and colorectal cancer. *Critical Reviews in Oncology/Hematology* 77, 30–47.

- Lindeberg, T., 1994. Scale-space theory: a basic tool for analyzing structures at different scales. *Journal of Applied Statistics* 21, 225–270.
- Lodhi, H., Saunders, C., Shawe-Taylor, J., Cristianini, N., Watkins, C., 2002. Text classification using string kernels. *J. Mach. Learn. Res.* 2, 419–444.
- Lowe, D.G., 1999. Object recognition from local scale-invariant features. *IEEE International Conference on Computer Vision*, vol. 2, p. 1150.
- Lowe, D.G., 2004. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision* 60, 91–110.
- Machida, H., Sano, Y., Hamamoto, Y., Muto, M., Kozu, T., Tajiri, H., Yoshida, S., 2004. Narrow-Band imaging in the diagnosis of colorectal mucosal lesions: a pilot study. *Endoscopy* 36, 1094–1098.
- Maji, S., Berg, A., Malik, J., 2008. Classification using intersection kernel support vector machines is efficient. In: *IEEE Conference on Computer Vision and Pattern Recognition*, 2008. CVPR 2008, pp. 1–8.
- Maroulis, D., Iakovidis, D., Karkanis, S., Karras, D., 2003. CoLD: a versatile detection system for colorectal lesions in endoscopy video-frames. *Computer Methods and Programs in Biomedicine* 70, 151–166.
- Matsumoto, A., Tanaka, S., Oba, S., Kanao, H., Oka, S., Yoshihara, M., Chayama, K., 2010. Outcome of endoscopic submucosal dissection for colorectal tumors accompanied by fibrosis. *Scandinavian Journal of Gastroenterology* 45, 1329–1337.
- Matsushima, C., Yamauchi, Y., Yamashita, T., Fujiyoshi, H., 2010. Object detection using relational binarized HOG feature and binary selection by real adaboost. In: *Proc. of the 13th Meeting on Image Recognition and Understanding (MIRU2010)*, pp. 18–25.
- Mayinger, B., Oezturk, Y., Stolte, M., Faller, G., Benninger, J., Schwab, D., Maiss, J., Hahn, E.G., Muehldorfer, S., 2006. Evaluation of sensitivity and inter- and intra-observer variability in the detection of intestinal metaplasia and dysplasia in Barrett's esophagus with enhanced magnification endoscopy. *Scandinavian Journal of Gastroenterology* 41, 349–356.
- Meining, A., Rosch, T., Kiesslich, R., Muders, M., Sax, F., Heldwein, W., 2004. Inter- and intra-observer variability of magnification chromoendoscopy for detecting specialized intestinal metaplasia at the gastroesophageal junction. *Endoscopy* 36, 160–164.
- Mikolajczyk, K., Schmid, C., 2003. A performance evaluation of local descriptors. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2003. Proceedings 2003, pp. II-257–II-263.
- Mikolajczyk, K., Schmid, C., 2004. Scale & affine invariant interest point detectors. *International Journal of Computer Vision* 60, 63–86.
- Mikolajczyk, K., Schmid, C., 2005. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27, 1615–1630.
- Milgram, J., Cheriet, M., Sabourin, R., 2006. One against one or one against all: Which one is better for handwriting recognition with SVMs? In: *Proc. of 10th International Workshop on Frontiers in Handwriting Recognition*.
- Ministry of Health, Labour and Welfare, Japan, 2009. Vital Statistics in Japan – The latest trends. <<http://www.mhlw.go.jp/english/database/db-hw/vs01.html>>.
- Müller, H., Kalpathy-Cramer, J., Kahn, C., Hatt, W., Bedrick, S., Hersh, W., 2009. Overview of the ImageCLEFmed 2008 medical image retrieval task. In: Peters, C., Deselaers, T., Ferro, N., Gonzalo, J., Jones, G., Kurimo, M., Mandl, T., Penas, A., Petras, V. (Eds.), *Evaluating systems for multilingual and multimodal information access*, Lecture Notes in Computer Science, vol. 5706. Springer, Berlin/Heidelberg, pp. 512–522.
- Müller, H., Michoux, N., Bandon, D., Geissbuhler, A., 2004. A review of content-based image retrieval systems in medical applications – clinical benefits and future directions. *International Journal of Medical Informatics* 73, 1–23.
- Nakayama, H., Harada, T., Kuniyoshi, Y., 2010. Global gaussian approach for scene categorization using information geometry. In: *2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2336–2343.
- National Cancer Institute, US National Institutes of Health, 2010. Colon and Rectal Cancer. <<http://www.cancer.gov/cancertopics/types/colon-and-rectal>>.
- Nister, D., Stewenius, H., 2006. Scalable recognition with a vocabulary tree. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 2161–2168.
- Nowak, E., Jurie, F., Triggs, B., 2006. Sampling strategies for bag-of-features image classification. In: *Leonardis, A., Bischof, H., Pinz, A. (Eds.), Computer Vision - ECCV-2006, Lecture Notes in Computer Science*, vol. 3954. Springer, Berlin/Heidelberg, pp. 490–503.
- Oba, S., Tanaka, S., Oka, S., Kanao, H., Yoshida, S., Shimamoto, F., Chayama, K., 2010. Characterization of colorectal tumors using narrow-band imaging magnification: combined diagnosis with both pit pattern and microvessel features. *Scandinavian Journal of Gastroenterology* 45, 1084–1092.
- Oba, S., Tanaka, S., Sano, Y., Oka, S., Chayama, K., 2011. Current status of narrow-band imaging magnifying colonoscopy for colorectal neoplasia in Japan. *Digestion* 83, 167–172.
- Oh, J., Hwang, S., Lee, J., Tavanapong, W., Wong, J., de Groen, P.C., 2007. Informative frame classification for endoscopy video. *Medical Image Analysis* 11, 110–127.
- Onji, K., Yoshida, S., Tanaka, S., Kawase, R., Takemura, Y., Oka, S., Tamaki, T., Raytchev, B., Kaneda, K., Yoshihara, M., Chayama, K., 2011. Quantitative analysis of colorectal lesions observed on magnified endoscopy images. *Journal of Gastroenterology* 46, 1382–1390, <http://dx.doi.org/10.1007/s00535-011-0459-x>.
- Oto, A., 2002. Virtual endoscopy. *European Journal of Radiology* 42, 231–239.
- Padhani, A.R., 1999. Advances in imaging of colorectal cancer. *Critical Reviews in Oncology/Hematology* 30, 189–199.
- Panossian, A.M., Raimondo, M., Wolfen, H.C., 2011. State of the art in the endoscopic imaging and ablation of Barrett's esophagus. *Digestive and Liver Disease* 43, 365–373.
- Perronnin, F., 2008. Universal and adapted vocabularies for generic visual categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30, 1243–1256.
- Perronnin, F., Dance, C., Csukca, G., Bressan, M., 2006. Adapted vocabularies for generic visual categorization. In: *Leonardis, A., Bischof, H., Pinz, A. (Eds.), Computer Vision - ECCV-2006, Lecture Notes in Computer Science*, vol. 3954. Springer, Berlin/Heidelberg, pp. 464–475.
- Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A., 2007. Object retrieval with large vocabularies and fast spatial matching. In: *IEEE Conference on Computer Vision and Pattern Recognition*, 2007. CVPR '07, pp. 1–8.
- Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A., 2008. Lost in quantization: improving particular object retrieval in large scale image databases. In: *IEEE Conference on Computer Vision and Pattern Recognition*, 2008. CVPR 2008, pp. 1–8.
- Qiu, B., Yanai, K., 2008. Objects over the world. In: *Huang, Y.M., Xu, C., Cheng, K.S., Yang, J.F., Swamy, M., Li, S., Ding, J.W. (Eds.), Advances in Multimedia Information Processing - PCM 2008, Lecture Notes in Computer Science*, vol. 5353. Springer, Berlin/Heidelberg, pp. 296–305.
- Quelhas, P., Odobez, J.M., 2007. Multi-level local descriptor quantization for bag-of-features image representation. In: *Proceedings of the 6th ACM international conference on Image and Video Retrieval*, ACM, New York, NY, USA, pp. 242–249.
- Raghavendra, M., Hewett, D.G., Rex, D.K., 2010. Differentiating adenomas from hyperplastic colorectal polyps: narrow-band imaging can be learned in 20 minutes. *Gastrointestinal Endoscopy* 72, 572–576.
- Rex, D.K., Kahi, C., O'Brien, M., Levin, T., Pohl, H., Rastogi, A., Burgart, L., Imperiale, T., Ladabaum, U., Cohen, J., 2011. The American society for gastrointestinal endoscopy PIVI (Preservation and incorporation of valuable endoscopic innovations) on real-time endoscopic assessment of the histology of diminutive colorectal polyps. *Gastrointestinal Endoscopy* 73, 419–422.
- Saito, S., Tajiri, H., Ohya, T., Nikami, T., Aihara, H., Ikegami, M., 2011. Imaging by magnifying endoscopy with NBI implicates the remnant capillary network as an indication for endoscopic resection in early colon cancer. *International Journal of Surgical Oncology* 2011, 1–10.
- Saito, Y., Uraoka, T., Matsuda, T., Emura, F., Ikehara, H., Mashimo, Y., Kikuchi, T., Fu, K.I., Sano, Y., Saito, D., 2007. Endoscopic treatment of large superficial colorectal tumors: a case series of 200 endoscopic submucosal dissections (with video). *Gastrointestinal Endoscopy* 66, 966–973.
- Sanford, K.W., McPherson, R.A., 2009. Fecal occult blood testing. *Clinics in Laboratory Medicine* 29, 523–541.
- Sano, Y., Horimatsu, T., Fu, K.I., Katagiri, A., Muto, M., Ishikawa, H., 2006. Magnifying observation of microvascular architecture of colorectal lesions using a narrow-band imaging system. *Digestive Endoscopy* 18, S44–S51.
- Sano, Y., Ikematsu, H., Fu, K.I., Emura, F., Katagiri, A., Horimatsu, T., Kaneko, K., Soetikno, R., Yoshida, S., 2009. Meshed capillary vessels by use of narrow-band imaging for differential diagnosis of small colorectal polyps. *Gastrointestinal Endoscopy* 69, 278–283.
- Schmid, C., Mohr, R., 1997. Local grayvalue invariants for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19, 530–535.
- Schölkopf, B., Smola, A.J., 2002. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT Press, Cambridge, Mass. [u.a.].
- Shin, L.K., Poulos, P., Jeffrey, R.B., 2010. MR colonography and MR enterography. *Gastrointestinal Endoscopy Clinics of North America* 20, 323–346.
- Shotton, J., Johnson, M., Cipolla, R., 2008. Semantic texton forests for image categorization and segmentation. In: *IEEE Conference on Computer Vision and Pattern Recognition*, 2008. CVPR 2008, pp. 1–8.
- Sivic, J., Zisserman, A., 2003. Video Google: a text retrieval approach to object matching in videos. In: *Ninth IEEE International Conference on Computer Vision*, 2003. Proceedings, pp. 1470–1477.
- Sonnenburg, S., Rätsch, G., Schäfer, C., Schölkopf, B., 2006. Large scale multiple kernel learning. *Journal of Machine Learning Research*, 1531–1565.
- Stehle, T., Auer, R., Gross, S., Behrens, A., Wulff, J., Aach, T., Winograd, R., Trautwein, C., Tischendorf, J., 2009. Classification of colon polyps in NBI endoscopy using vascularization features. In: *Proc. of Medical Imaging 2009: Computer-Aided Diagnosis*, SPIE, pp. 72602S–72602S-12.
- Steinwart, I., Christmann, A., 2008. *Support Vector Machines*, first ed. Springer, New York.
- Sundaram, P., Zomorodian, A., Beaulieu, C., Napel, S., 2008. Colon polyp detection using smoothed shape operators: preliminary results. *Medical Image Analysis* 12, 99–119.
- Swain, M.J., Ballard, D.H., 1991. Color indexing. *International Journal of Computer Vision* 7, 11–32, <http://dx.doi.org/10.1007/BF00130487>.
- Takemura, Y., Yoshida, S., Tanaka, S., Kawase, R., Onji, K., Oka, S., Tamaki, T., Raytchev, B., Kaneda, K., Yoshihara, M., Chayama, K., 2012. Computer-aided system for predicting the histology of colorectal tumors by using narrow-band imaging magnifying colonoscopy (with video). *Gastrointestinal Endoscopy* 75, 179–185.
- Takemura, Y., Yoshida, S., Tanaka, S., Onji, K., Oka, S., Tamaki, T., Kaneda, K., Yoshihara, M., Chayama, K., 2010. Quantitative analysis and development of a computer-aided system for identification of regular pit patterns of colorectal lesions. *Gastrointestinal Endoscopy* 72, 1047–1051.
- Tamai, N., Sakamoto, T., Nakajima, T., Matsuda, T., Saito, Y., Tajiri, H., Koyama, R., Kido, S., 2011. Su1566 computer-assisted automatic identification system for colorectal narrow band imaging (NBI) classification. *Gastrointestinal Endoscopy* 73, AB306–AB307.

- Tamaki, T., Yoshimuta, J., Takeda, T., Raytchev, B., Kaneda, K., Yoshida, S., Takemura, Y., Tanaka, S., 2011. A system for colorectal tumor classification in magnifying endoscopic NBI images. In: Kimmel, R., Klette, R., Sugimoto, A. (Eds.), *Computer Vision – ACCV 2010*, Lecture Notes in Computer Science, vol. 6493. Springer, Berlin/Heidelberg, pp. 452–463.
- Tamegai, Y., 2007. Endoscopic submucosal dissection (ESD) for large colorectal tumors comparing with endoscopic piecemeal mucosal resection (EPMR). *Gastrointestinal Endoscopy* 65, AB275.
- Tanaka, S., Kaltenbach, T., Chayama, K., Soetikno, R., 2006. High-magnification colonoscopy (with videos). *Gastrointestinal Endoscopy* 64, 604–613.
- Tischendorf, J., Gross, S., Winograd, R., Hecker, H., Auer, R., Behrens, A., Trautwein, C., Aach, T., Stehle, T., 2010. Computer-aided classification of colorectal polyps based on vascular patterns: a pilot study. *Endoscopy* 42, 203–207.
- Tong, S., Koller, D., 2002. Support vector machine active learning with applications to text classification. *J. Mach. Learn. Res.* 2, 45–66.
- Tosun, A.B., Kandemir, M., Sokmensuer, C., Gunduz-Demir, C., 2009. Object-oriented texture analysis for the unsupervised segmentation of biopsy images for cancer detection. *Pattern Recognition* 42, 1104–1112.
- Tuytelaars, T., 2010. Dense interest points. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2281–2288.
- Tuytelaars, T., Mikolajczyk, K., 2007. Local invariant feature detectors: a survey. *Foundations and Trends® in Computer Graphics and Vision* 3, 177–280.
- Tweedle, E., Rooney, P., Watson, A., 2007. Screening for rectal cancer? will it improve cure rates? *Clinical Oncology* 19, 639–648.
- Vapnik, V.N., 1998. *Statistical Learning Theory*. Wiley, New York [u.a.].
- Vedaldi, A., Fulkerson, B., 2008. VLFeat: an open and portable library of computer vision algorithms. <<http://www.vlfeat.org/>>.
- Vedaldi, A., Zisserman, A., 2012. Efficient additive kernels via explicit feature maps. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34, 480–492.
- Wada, Y., Kudo, S., Kashida, H., Ikehara, N., Inoue, H., Yamamura, F., Ohtsuka, K., Hamatani, S., 2009. Diagnosis of colorectal lesions with the magnifying narrow-band imaging system. *Gastrointestinal Endoscopy* 70, 522–531.
- Watanabe, T., Itabashi, M., Shimada, Y., Tanaka, S., Ito, Y., Ajioka, Y., Hamaguchi, T., Hyodo, I., Igarashi, M., Ishida, H., Ishiguro, M., Kanemitsu, Y., Kokudo, N., Muro, K., Ochiai, A., Oguchi, M., Ohkura, Y., Saito, Y., Sakai, Y., Ueno, H., Yoshino, T., Fujimori, T., Koinuma, N., Morita, T., Nishimura, G., Sakata, Y., Takahashi, K., Takiuchi, H., Tsuruta, O., Yamaguchi, T., Yoshida, M., Yamaguchi, N., Kotake, K., Sugihara, K., for Cancer of the Colon, J.S., Rectum, 2012. Japanese Society for Cancer of the Colon and Rectum (JSCCR) guidelines 2010 for the treatment of colorectal cancer. *International Journal of Clinical Oncology* 17, 1–29. <http://dx.doi.org/10.1007/s10147-011-0315-2>.
- Weston, J., Watkins, C., 1999. Support vector machines for multi-class pattern recognition. In: *Proc. of 7th European Symposium on Artificial Neural Networks (ESANN1999)*, pp. 219–224.
- Winn, J., Criminisi, A., Minka, T., 2005. Object categorization by learned universal visual dictionary. In: *Tenth IEEE International Conference on Computer Vision, 2005. ICCV 2005*, pp. 1800–1807.
- Wu, C., SiftGPU: A GPU implementation of scale invariant feature transform (SIFT). <http://cs.unc.edu/ccwu/siftgpu/>.
- Yanai, K., Qiu, B., 2009. Mining cultural differences from a large number of geotagged photos. In: *Proceedings of the 18th International Conference on World Wide Web*. ACM, New York, NY, USA, pp. 1173–1174.
- Ye, B.D., Byeon, J.S., Kwon, S., Kim, B., Myung, S.J., Yang, S.K., Kim, J.H., 2008. Clinical course of submucosal colorectal cancers treated by endoscopic mucosal resection. *Gastrointestinal Endoscopy* 67, AB312.
- Yoshida, H., Dachman, A.H., 2004. Computer-aided diagnosis for CT colonography. *Seminars in Ultrasound, CT, and MRI* 25, 419–431.
- Zhang, J., Marszałek, M., Lazebnik, S., Schmid, C., 2007. Local features and kernels for classification of texture and object categories: a comprehensive study. *International Journal of Computer Vision* 73, 213–238.
- Zhou, X., Zhuang, X., Yan, S., Chang, S.F., Hasegawa-Johnson, M., Huang, T.S., 2008. SIFT-Bag kernel for video event analysis. In: *Proceeding of the 16th ACM International Conference on Multimedia*. ACM, New York, NY, USA, pp. 229–238.