Medical Image Analysis 16 (2012) 1415–1422

Contents lists available at SciVerse ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Endoscopic image analysis in semantic space

R. Kwitt [a,*], N. Vasconcelos [b], N. Rasiwasia [b], A. Uhl [c], B. Davis [a], M. Häfner [d], F. Wrba [e]

[a] Kitware Inc., Chapel Hill, NC, USA
[b] Statistical Visual Computing Laboratory, UC San Diego, USA
[c] Department of Computer Sciences, University of Salzburg, Austria
[d] Department for Internal Medicine, St. Elisabeth Hospital, Vienna, Austria
[e] Department of Clinical Pathology, Medical University of Vienna, Austria

ARTICLE INFO

Article history:
Available online 29 May 2012

Keywords:
Image understanding
Pit pattern analysis
Semantic modeling

ABSTRACT

A novel approach to the design of a semantic, low-dimensional, encoding for endoscopic imagery is proposed. This encoding is based on recent advances in scene recognition, where semantic modeling of image content has gained considerable attention over the last decade. While the semantics of scenes are mainly comprised of environmental concepts such as *vegetation*, *mountains* or *sky*, the semantics of endoscopic imagery are medically relevant visual elements, such as polyps, special surface patterns, or vascular structures. The proposed semantic encoding differs from the representations commonly used in endoscopic image analysis (for medical decision support) in that it establishes a *semantic space*, where each coordinate axis has a clear human interpretation. It is also shown to establish a connection to Riemannian geometry, which enables principled solutions to a number of problems that arise in both physician training and clinical practice. This connection is exploited by leveraging results from information geometry to solve problems such as (1) recognition of important semantic concepts, (2) semantically-focused image browsing, and (3) estimation of the *average-case* semantic encoding for a collection of images that share a medically relevant visual detail. The approach can provide physicians with an easily interpretable, semantic encoding of visual content, upon which further decisions, or operations, can be naturally carried out. This is contrary to the prevalent practice in endoscopic image analysis for medical decision support, where image content is primarily captured by discriminative, high-dimensional, appearance features, which possess discriminative power but lack human interpretability.

1. Motivation

Over the past decade, there has been increased research interest in decision-support systems for endoscopic imagery. In the context of routine examinations of the colon, an important task is to perform *pit pattern* discrimination. This is usually guided by the Kudo criteria (Kudo et al., 1994), based on the observation of a strong correlation between the visual appearance of the highly-magnified mucosa and the visual appearance of dissected specimen under the microscope. Pit pattern analysis not only facilitates *in vivo* predictions of the histology but represents a valuable guideline for treatment strategies in general. The Kudo criteria discriminates between five pit pattern types I–V, where type III is subdivided into III-S and III-L. Types I and II are usually characteristic of *non-neoplastic* lesions, types III and IV indicate adenomnatous polyps and type V is highly indicative for invasive carcinoma. Apart from incidental image structures, such as colon folds, the pit patterns are the predominant concepts upon which histological predictions are made. While images where one particular pit pattern type is prevalent are fairly rare, mixtures of pit patterns are quite commonly found in practice. The development of decision-support systems for endoscopic imagery is desirable for several reasons.

First, routine examinations often involve unnecessary biopsies or polyp resections, both because physicians are under serious time pressure and because the standard protocol dictates the use of biopsies in cases of uncertainty. This is controversial, since resecting metaplastic lesions is time-consuming and the removal of invasive cancer can be hazardous.

Second, the interpretation of the acquired image material can be difficult, due to high variability in image appearance depending on the type of imaging equipment. Novel modalities, such as high-magnification endoscopy, narrow band imaging (NBI) or confocal laser endomicroscopy (CLE) all highlight different mucosal structures; CLE even provides an *in vivo* view of deeper tissue layers at a microscopic scale. A critical problem is that visual criteria for assessing the malignant potential of colorectal lesions are still under extensive clinical evaluation and substantial experience

* Corresponding author. Address: Kitware Inc., 101 East Weaver St, Carrboro, NC 27510, USA. Tel.: +1 9199235575.
E-mail address: roland.kwitt@kitware.com (R. Kwitt).

1361-8415/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.media.2012.04.010

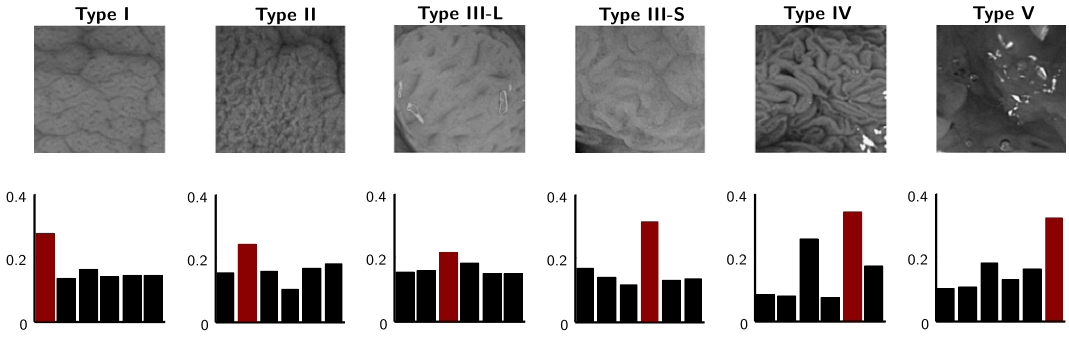

Fig. 1. Endoscopy images of the colon mucosa (top row), taken by a high-magnification endoscope, showing typical mucosal structures (*pit patterns*). The bottom row shows the semantic encoding proposed in this work. The height of each bar indicates the probability that a particular type of visual structure is present in the image. The red bar reports to the type of structure that is actually present. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

(Tung et al., 2001) is usually required to achieve good results under these criteria.

Third, decision-support systems can be a helpful aid to the training of future physicians. Due to differences among endoscopic imaging modalities and endoscope brands, it is advisable to train the physician on data from the very device to be used in practice. However, the learning of Kudo's pit pattern classification requires experienced physicians to go through the time-consuming selection of images representative of the different pit pattern types. This is a tedious process, which becomes unmanageable for large-scale datasets.

For all these reasons, there has been increasing interest in decision-support systems for endoscopic imagery over the last decade. This effort has been predominantly directed to the use of automated image content analysis techniques in the prediction of histopathological results (e.g. André et al., 2009, 2010; Tischendorf et al., 2010; Kwitt et al., 2011; Häfner et al., 2012). It has led to a plethora of approaches that first compute a collection of localized appearance features and then input these features to a discriminant classifier, usually a support vector machine. From a purely technical point of view, this problem description is similar to scene recognition problems in the computer vision literature, with the difference that invariance properties of the image representation, such as invariance to rotation or translation, are considered more important in the medical field. A relevant research trend in computer vision is to replace the inference of scene labels from appearance descriptors alone by more abstract, intermediate-level, representations (Fei-Fei and Perona, 2005; Lazebnik et al., 2006; Boureau et al., 2010; Rasiwasia et al., 2006; Rasiwasia and Vasconcelos, 2008; Dixit et al., 2011). The prevalent approach to scene classification is to learn a codebook of so called *visual words*, from a large corpus of appearance descriptors, and represent each image as an histogram—known as the bag-of-words (BoW) histogram—of codeword indices. These mid-level representations are input to a discriminant classifier for scene label prediction.

In the context of pit-pattern classification, this classification architecture could, in principle, be used to produce a class label, such as *neoplastic* or *non-neoplastic*, to be presented to a physician. However, while BoW histograms have state-of-the-art recognition rates for both medical and computer vision applications, they are not generally amenable to human interpretation. This is due to the facts that they (1) are high-dimensional, and (2) define a space whose coordinate axes lack semantic interpretation. This lack of interpretability raises a number of difficulties to the clinical deployment of the resulting decision-support systems. First, while the resulting predictions are valuable, it is not uncommon for the medical community to reject *black-box* solutions that do not provide interpretable information on how these predictions were reached. Second, the lack of insight on the factors that determine the predicted image labels severely compromise their usefulness for physician training. Third, it has been recently argued that a more semantically-focused mid-level representation is conducive to better recognition results (cf. Schwaninger et al., 2006; Rasiwasia and Vasconcelos, 2008). Several works have, in fact, shown that an image representation which captures the occurrence probabilities of predefined semantic concepts is not only competitive with BoW, but computationally more efficient due to its lower dimensionality. Since the semantic concepts can be chosen so as to be interpretable by physicians, the approach is also conducive to a wider acceptability by the medical community. For example, (André et al., 2012) demonstrated that low-dimensional semantic encodings are highly beneficial to the interpretation of CLE imagery.

The goal of this work is to establish a semantic encoding of endoscopic imagery, so as to produce systems for automated malignancy assessment of colorectal lesions of greater flexibility than those possible with existing approaches. We demonstrate the benefits of the proposed encoding on image material obtained during routine examinations of the colon mucosa. The imaging modality is high-magnification chromo-endoscopy, which offers a level of visual detail suitable for the categorization of mucosal surface structures into different *pit pattern* types. Some typical images are shown in the top row of Fig. 1. The aforementioned shortcomings of previous approaches are addressed by adapting a recent method (Rasiwasia and Vasconcelos, 2008) from the scene recognition literature to the inference of semantic encodings for endoscopic imagery. Some examples of these encodings are shown in the bottom row of Fig. 1. While the general principle is well established in the computer vision literature, we demonstrate that it is a principled solution for a number of important applications in the domain of endoscopic image analysis. The first is the automated assessment of the malignant potential of colorectal lesions, where the proposed semantic encoding is shown to enable state-of-the-art image classification with substantially increased human interpretability of classifier predictions. The second is a tool to browse endoscopic image databases by typicality of particular pit patterns, allowing trainees in gastroenterology to find *most-representative* cases for each pit pattern class. The third is a strategy to determine images which represent the *average*-case for a particular pit pattern type. This enables physicians to keep track of what they typically see in clinical practice. A preliminary version of this work appeared in Kwitt et al. (2011).

2. The design of the semantic space

We start by introducing some notation. We denote by $\mathscr{D} = \{(I_i, \boldsymbol{c}_i)\},\ i = 1, \ldots, D$ a corpus of D image-caption tuples $(I_i, \boldsymbol{c}_i)$,

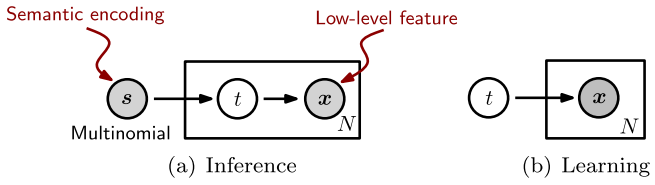

Fig. 2. Image formation models for inference and learning.

where image I_i is augmented by a binary caption vector $\boldsymbol{c}_i$. Captions are drawn from a dictionary $\mathscr{T} = \{t_1, \ldots, t_C\}$ of C semantic concepts (e.g., the pit pattern types). It is assumed that the database $\mathscr{D}$ is *weakly-labeled* in the sense that while $c_{ij} = 1$ signifies the presence of the jth semantic concept in image i, $c_{ij} = 0$ does not necessarily imply its absence. This situation is common in medical image analysis where often only the most dominant, or most medically relevant, concept is annotated. In this particular case, caption vectors contain only one non-zero entry: the class label for the image. It is further assumed that each image is represented by a collection of N low-level features $\boldsymbol{x}_i \in \mathscr{X}$, i.e. $I_i = \{\boldsymbol{x}_1^i, \ldots, \boldsymbol{x}_N^i\}$, computed from N localized image patches $P_j^i, j = 1, \ldots, N$. These patches can be evenly distributed across the image, or obtained with any other sampling strategy. $\mathscr{X} \subset \mathbb{R}^d$ is a low-level feature space, e.g., the space of SIFT (Lowe, 2004) descriptors.

The generative model for the proposed semantic encoding is shown in Fig. 2(a). Visual features $\boldsymbol{x}_i$ are independently drawn from concepts t, and concepts are drawn from a multinomial random variable with parameter vector $\boldsymbol{s} \in [0,1]^C$. Given an image I, the mutinomial parameters in $\boldsymbol{s}$ are inferred from $\{\boldsymbol{x}_i\}_{i=1}^N$ as follows (the image index j is omitted for brevity). First, the concept of largest posterior probability is found per $\boldsymbol{x}_i$, i.e. $t_i^* = q_b(\boldsymbol{x}_i)$ with

$$q_b(\boldsymbol{x}_i) = \arg\max_{t \in \mathscr{T}} P_{T|\boldsymbol{X}}(t|\boldsymbol{x}_i) = \arg\max_{t \in \mathscr{T}} \frac{P_{\boldsymbol{X}|T}(\boldsymbol{x}_i|t)}{\sum_w P_{\boldsymbol{X}|T}(\boldsymbol{x}_i|w)}. \quad (1)$$

This assumes equal prior probability for all concepts, but could be easily extended for a non-uniform prior. The mapping $q_b : \mathscr{X} \to \mathscr{T}$ *quantizes* features into concepts in a Bayesian, *minimum probability-of-error*, fashion. The concept occurrences of I are then summarized in a concept occurrence vector $(o_1, \ldots, o_C)'$, where $o_t = |\{i : t_i^* = t\}|$ is the number of occurrences of concept t in image I. Finally, an MAP estimate of $\boldsymbol{s}$, under the assumption of a Dirichlet prior of parameter α, is computed with

$$\hat{\boldsymbol{s}} = \left(\frac{o_1 + \alpha - 1}{\sum_w (o_w + \alpha - 1)}, \ldots, \frac{o_C + \alpha - 1}{\sum_w (o_w + \alpha - 1)} \right)'. \quad (2)$$

Note that α acts as a regularization parameter. In the terminology of Rasiwasia and Vasconcelos (2008), $\hat{\boldsymbol{s}}$ is denoted the *semantic multinomial (SMN)* of image I. This establishes a mapping $\Pi : \mathscr{X}^N \to \mathbb{P}^{C-1}, I \mapsto \boldsymbol{s}$ from an image represented in feature space $\mathscr{X}^N$ to an image represented as a point on the *semantic (probability) simplex* $\mathbb{P}^{C-1}$. If the boundaries of the simplex have zero probability (a constraint that can be enforced by the Dirichlet regularizer) the simplex is a Riemannian manifold when endowed with the Fisher information metric $\mathscr{I}$ (cf. Lebanon, 2005). Since we refer to $\mathbb{P}^{C-1}$ as the semantic (probability) simplex, $(\mathbb{P}^{C-1}, \mathscr{I})$ is denoted the *semantic manifold*. It will later be shown that information geometry provides a rich set of tools for performing various operations on this manifold.

Learning of the Π mapping requires estimates of the concept-conditional distributions $P_{\boldsymbol{X}|T}(\boldsymbol{x}|t)$ from the available weakly-labeled image data. Since the concept label of *each* visual feature is not known, this is done with resort to multiple instance learning (Maron, 1998), based on the image formation model of Fig. 2(b). The visual features extracted from all images labeled with concept t are pooled into dataset $\mathscr{D}_t = \left\{ \boldsymbol{x}_i^j | \boldsymbol{c}_t^j = 1 \right\}$, which is then used to estimate $P_{\boldsymbol{X}|T}(\boldsymbol{x}|t)$. The intuition is that visual features representative of the semantic concept are more likely to occur in the training set and dominate the probability estimates. In multiple instance learning terminology, $\mathscr{D}_t$ is the *bag of positive examples* for concept t. Fig. 3 shows a schematic illustration of the SMN representation for a toy three-concept problem.

2.1. Implementation

The proposed implementation of semantic encoding relies on Gaussian mixture models to estimate the concept-conditional probability densities $P_{\boldsymbol{X}|T}(\boldsymbol{x}|t)$. The mixture parameters are estimated with the EM algorithm (Dempster et al., 1977), initialized by k-means++ (Arthur and Vassilvitskii, 2007), and the covariance matrices restricted to diagonal form. The low-level appearance representation is based on SIFT descriptors[1] (using 4×4 grid cells with 8 orientation histogram bins), due to the prevalence and success of SIFT in a wide variety of computer vision applications. In our implementation, SIFT descriptors are computed on an evenly-spaced 8×8 pixel grid. Previous studies (Fei-Fei and Perona, 2005, 2009) have shown that this dense-SIFT representation has good recognition performance.

3. Analysis of endoscopic imagery in semantic space

The semantic image encoding of Section 2 was applied to three application scenarios of potential interest for endoscopic image analysis: (1) assessment of the malignant potential of colorectal lesions by recognizing non-neoplastic and neoplastic lesions, (2) semantically-focused browsing for *most-representative* cases and (3) determination of *average*-case image representatives per semantic concept.

3.1. Data material

Our data material are colonoscopy images (either 624×533 or 586×502 pixel), acquired throughout 2005–2009 in the Department of Gastroenterology and Hepathology of the Medical University of Vienna. All images were captured with a high-magnification Olympus Evis Exera CF-Q160ZI/L endoscope, using a magnification factor of up to $150\times$. The original dataset consists of 327 images of a total of 40 patients. To obtain a larger sample size, 256×256 pixel regions were manually extracted (with minimum overlap) to obtain a final dataset of 716 images. All images were converted to grayscale for further processing. Examination of the lesions was performed after dye-spraying with indigo-carmine, a routine procedure to enhance structural details. Biopsies were taken of lesions classified as pit pattern types I, II and V, since I and II need not be removed and type V cannot be removed, as explained in Section 1. Lesions of type III-S, III-L and IV were removed endoscopically. Table 1 lists the number of patients and images for a dataset split into non-neoplastic (i.e. types I, II) and neoplastic lesions (types III–V).

3.2. Recognizing non-neoplastic/neoplastic lesions

As a first application scenario, we demonstrate a method to classify the images into non-neoplastic and neoplastic lesions, given the proposed semantic encoding by SMNs. It is safe to claim that this is the most well-studied application scenario in the endoscopic image analysis literature. Many specifically-tailored low-level appearance features have been proposed recently, mainly considering the problem from a pure pattern classification

[1] LEAR implementation: http://lear.inrialpes.fr/people/dorko/downloads.html.

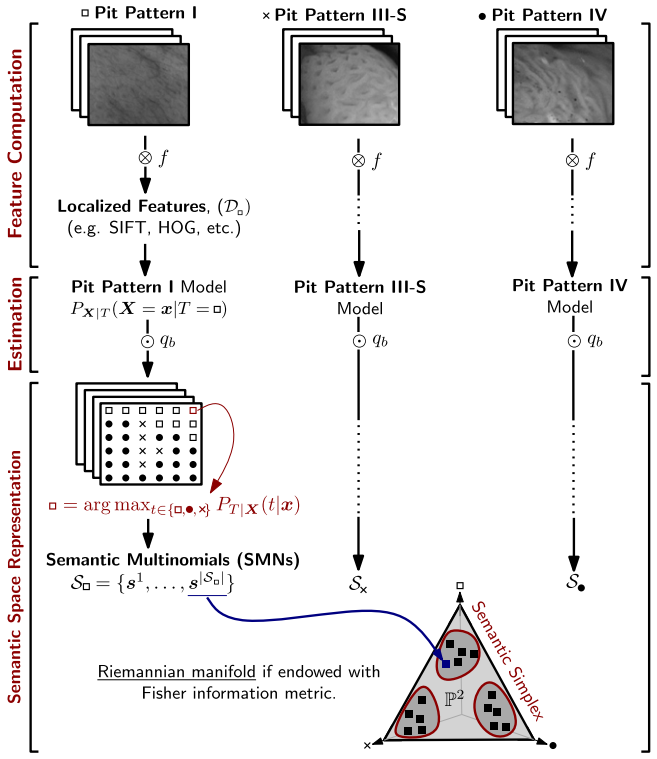

Fig. 3. Semantic encoding of images as points on the semantic simplex.

Table 1
Number of images and patients for non-neoplastic and neoplastic lesions.

	Non-neoplastic	Neoplastic	Total
Number of images	198	518	716
Number of patients	14	26	40

perspective. While the discriminant classifier at the end of the pipeline is often optimized for a particular feature type, our approach is based on generic SMNs. This leads to a natural classification method that is independent of the underlying appearance-level representation.

Although it would be possible to train a support vector machine with an RBF kernel on the semantic representation, this would not respect the structure of the underlying Riemannian manifold. In fact, a RBF kernel based on the Euclidean distance would correspond to assuming that the SMNs reside in flat Euclidean space. This would ignore the fact that the SMNs are parameters of multinomial distributions, and thus represent points on the multinomial manifold. Better performance can usually be obtained by adapting the similarity measure to the structure of the manifold. For a Riemannian manifold the natural similarity measure is the associated geodesic distance. Although geodesics tend to be difficult to compute—and rarely have closed-form solution—this is not the case for the semantic manifold. In this case, it is possible to exploit the well-known isomorphism

$$F : \mathbb{P}^{C-1} \rightarrow \mathbb{S}^{C-1}, \quad \boldsymbol{s} \mapsto 2\sqrt{\boldsymbol{s}} \tag{3}$$

between the Riemannian manifolds $(\mathbb{P}^{C-1}, \mathscr{I})$ and $(\mathbb{S}^{C-1}, \delta)$, where $\mathbb{S}^{C-1}$ is the $(C-1)$ sphere (of radius two) and δ the Euclidean metric inherited when embedding $\mathbb{S}^{C-1}$ in $\mathbb{R}^C$. Under this isometry, the geodesic distance between $\boldsymbol{s}^i$ and $\boldsymbol{s}^j$ reduces to the great-circle distance between $F(\boldsymbol{s}^i)$ and $F(\boldsymbol{s}^j)$, i.e.,

$$d_{\mathscr{I}}(\boldsymbol{s}^i, \boldsymbol{s}^j) = d_\delta(F(\boldsymbol{s}^i), F(\boldsymbol{s}^j)) = 2\arccos\left(\langle\sqrt{\boldsymbol{s}^i}, \sqrt{\boldsymbol{s}^j}\rangle\right). \tag{4}$$

This provides a closed-form solution for computing distances between SMNs on the semantic manifold. It is also possible to prove (see Appendix A) that the kernel defined by the negative of this geodesic distance

$$k(\boldsymbol{s}^i, \boldsymbol{s}^j) := -d_{\mathscr{I}}(\boldsymbol{s}^i, \boldsymbol{s}^j) \tag{5}$$

satisfies all the requirements of a conditionally positive-definite (cpd) kernel, see (Schölkopf, 2000). This is interesting because cpd kernels can be used in the standard SVM architecture and share many of the closure properties of positive-definite kernels (Schölkopf and Smola, 2001). These properties enable the use of weighted sums of kernels or even a spatial pyramid variant of (5), as proposed in Grauman and Darrell (2005) or Lazebnik et al. (2006).

In summary, the use of the kernel of (5) within a SVM classifier is a principled approach to the combination of (1) a semantic space image representation based on low-level appearance features with (2) a state-of-the-art kernel-based discriminant classifier that respects the structure of the semantic space.

3.2.1. Experiments

A quantitative evaluation of the proposed classification strategy was performed with a *leave-one-patient-out protocol* recently adopted by many works (André et al., 2011; Häfner et al., 2012; Kwitt et al., 2011). This is in contrast to previous studies that have primarily followed a simple *leave-one-sample-out* evaluation protocol (Kwitt et al., 2010; Häfner et al., 2010; André et al., 2009). Leave-one-patient-out is more restrictive in the sense that *all* images from one patient are left out during SVM training. In a leave-one-sample-out protocol, only *one* image (over the whole collection) is left out per cross-validation run. When there are several images from the same patient, this can bias the reported classification rates. In addition, leave-one-patient-out further implies that there is no bias with respect to 256×256 pixel regions coming from the same image as well. For those reasons, the comparison of classification rates is restricted to recently published results, on the same database, that follow the leave-one-patient-out protocol.

The only parameter of the proposed classifier that requires tuning is the SVM cost factor C, which we optimize on the training data in each leave-one-patient-out iteration using ten linearly spaced values of $\log C \in [-2,4]$. We note that the proposed kernel is advantageous in the sense that it requires no tuning of kernel parameters such as the bandwidth of RBF kernels. Table 2 lists the average recognition rate (i.e., accuracy) for non-neoplastic vs. neoplastic lesion classification, together with sensitivity and specificity values. Sensitivity is defined as the total number of correctly classified images showing neoplastic lesions divided by the total number of images showing neoplastic lesions. The definition of specificity follows accordingly. The proposed approach is compared to a recent study of Häfner et al. (2012) which implements the *Opponent-Color Local-Binary-Pattern (OCLBP)* features of Mäenpää et al. (2002), the *Joint Color Multiscale LBP (JC-MB-LBP)* features of Häfner et al. (2009) as well as a new feature called *Local Color Vector Pattern (LCVP)*. Note that all the rates reported for

Table 2
Comparison of leave-one-patient-out classification results to various state-of-the-art methods, on the database used in this work. In case there is *no* statistically significant difference (at 5% significance) in the class predictions with respect to the top approach (bold), the classification accuracies are underlined.

Approach	Accuracy	Sens.	Spec.	Dim.
Proposed	82.0	95.0	48.0	6
LCVP (Häfner et al., 2012)	<u>79.6</u>	94.5	40.4	256
OCLBP (Mäenpää et al., 2002)	<u>70.0</u>	84.6	31.9	6912
JC-MB-LBP (Häfner et al., 2009)	**82.7**	93.2	55.1	196608

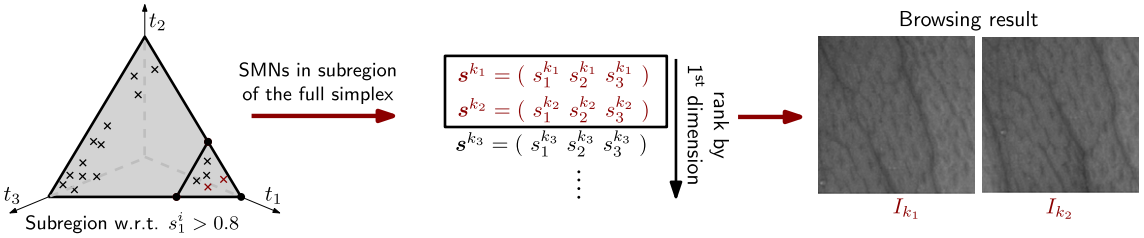

Fig. 4. Identifying the images, represented by SMNs, which are *most-characteristic* for concept t_1 (i.e. pit pattern type I).

these approaches were obtained from color (RGB) images, while we only use grayscale image information.[2] To assess whether an approach produces statistically significant class predictions with respect to the top result (bold), we employ a McNemar test at 5% significance. Rejection of the null-hypothesis (i.e., *no* statistically significant difference) is indicated by an underlined classification accuracy in Table 2.

We emphasize that, in contrast to the previous studies, we have made no effort to find the optimal appearance features for pit patterns. The experiment is instead designed to demonstrate that classification is possible with a much lower-dimensional representation (cf. last column of Table 2) that *can* be interpreted by humans. In fact, the proposed approach is not a direct competitor to previously published works, since we pursue a somewhat orthogonal research direction: to facilitate *image understanding* of endoscopic imagery at a semantic level, and not so much to derive new features. The proposed approach is not restricted to SIFT features, and any future improvements in the development of features that capture medically relevant visual structures (e.g. pit patterns) can be used to design improved semantic spaces.

3.3. Browsing endoscopic imagery by semantic information

Providing future gastroenterologists with a tool to browse endoscopic imagery by *typicality* of particular semantic concepts is a core motivation for the use of semantic image representations. The proposed representation addresses this problem very naturally. To identify the images most-characteristic of concept t_i (e.g. pit pattern III-L), it suffices to identify the subregion of the semantic simplex whose SMNs represent images where the t_ith concept is prominent with probability p. This can be trivially done by selecting the images for which $s_i > p$, $p \in [0,1]$. Fig. 4 illustrates this idea for $s_1 > 0.8$. Sorting the SMNs in the selected region along the ith dimension produces a list of the most-characteristic (i.e., top-ranked) images for concept t_i.

3.3.1. Experiments

To evaluate the proposed strategy for semantically-focused browsing, we first perform a visual comparison of the browsing results to *textbook* illustrations of the different pit pattern types, shown in the top row of Fig. 5. The SMNs were sorted along the dimension corresponding to each concept, and the K top-ranked images were extracted. To establish a realistic clinical scenario we ensured that the extracted images do *not* belong to the same patient. We call this *patient pruning* of the result set. Fig. 5 shows the images obtained when browsing for the $K = 5$ top-ranked images of each pit pattern. Images that remain after the patient pruning step are highlighted. Since the database images are not uniformly distributed over patients, the pruning step did not produce an equal number of browsing results per concept. Nevertheless, a comparison to the textbook illustrations reveals the desired correspondences, e.g., the characteristic gyrus-like structures of pit pattern IV, the round pits of pit pattern I, and the complete loss of structure for pit pattern V. Also presented is an incorrect browsing result for pit pattern III-L, depicting an image of type IV. This is a typical error due to the difficulty of distinguishing types III-L and IV, which have similar structural elements.

In addition to the visual inspection of the results in Fig. 5, we conducted a more objective evaluation using the ground-truth caption vectors of each image. This was based on the *average error rate* of the system when browsing the K top-ranked images per concept. A leave-one-patient-out protocol was used: (1) the patient's images were removed from the database, (2) SMNs were estimated from the remaining images, (3) the K top-ranked images per concept were extracted (now using the whole database) and (4) the patient pruning step was performed. The average error rate was then calculated as the percentage of images (averaged over all leave-one-patient-out runs) in the final browsing result of concept t_i which do not match the corresponding ground-truth caption vectors (i.e., zero entry at the ith position). Fig. 6 shows the average error rate as a function of K. At the operating point $K = 10$, $\approx$10% of the images were misclassified in the final browsing result. This rate is higher than that reported in our preliminary study (Kwitt et al., 2011) mainly because we now use generic SIFT descriptors, instead of the more texture-tailored DCT coefficient vectors of Kwitt et al. (2011).

3.4. Estimation of average-case representatives

The final application scenario considered in this work is to derive a semantic encoding representative of the *average-case* image within a given image collection (e.g., grouped by pit pattern). While Section 3.3 focused on the *corners* of the semantic simplex, i.e., the very characteristic cases, we now focus on the average-case images. Fig. 7 illustrates the difference between searching for the *most-characteristic* image for a particular concept versus searching for its *average-case* representative. Both have a clear merit: the first is of value for training purposes while the second facilitates visualization of what a physician typically sees in clinical practice.

As in the previous application scenarios, it is possible to draw on resources from information geometry to tackle the problem in a principled way. Rather than computing the arithmetic mean of a collection of SMNs, which would not respect the structure of the semantic manifold, we compute its Frechét mean. The Frechét mean of a collection of N points $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_N$ on a general (connected) Riemannian manifold $(\mathscr{M}, g)$ is defined as

$$\boldsymbol{\mu} = \arg\min_{\boldsymbol{a} \in \mathscr{M}} \sum_{i=1}^{N} w_i d_g^2(\boldsymbol{a}, \boldsymbol{a}_i) \tag{6}$$

where g denotes the Riemannian metric and d_g denotes the geodesic distance among two points, induced by g. When $\mathscr{M}$ is the Euclidean space, i.e., $d_g(\boldsymbol{a}, \boldsymbol{a}_i) = \|\boldsymbol{a} - \boldsymbol{a}_i\|$, $\boldsymbol{\mu}$ reduces to the arithmetic mean. In our application, where $d_g := d_{\mathscr{S}}$ is the geodesic distance on the semantic manifold, the Frechét mean has no closed-form. For this

[2] The results reported by Häfner et al. (2012) are slightly higher when using LAB.

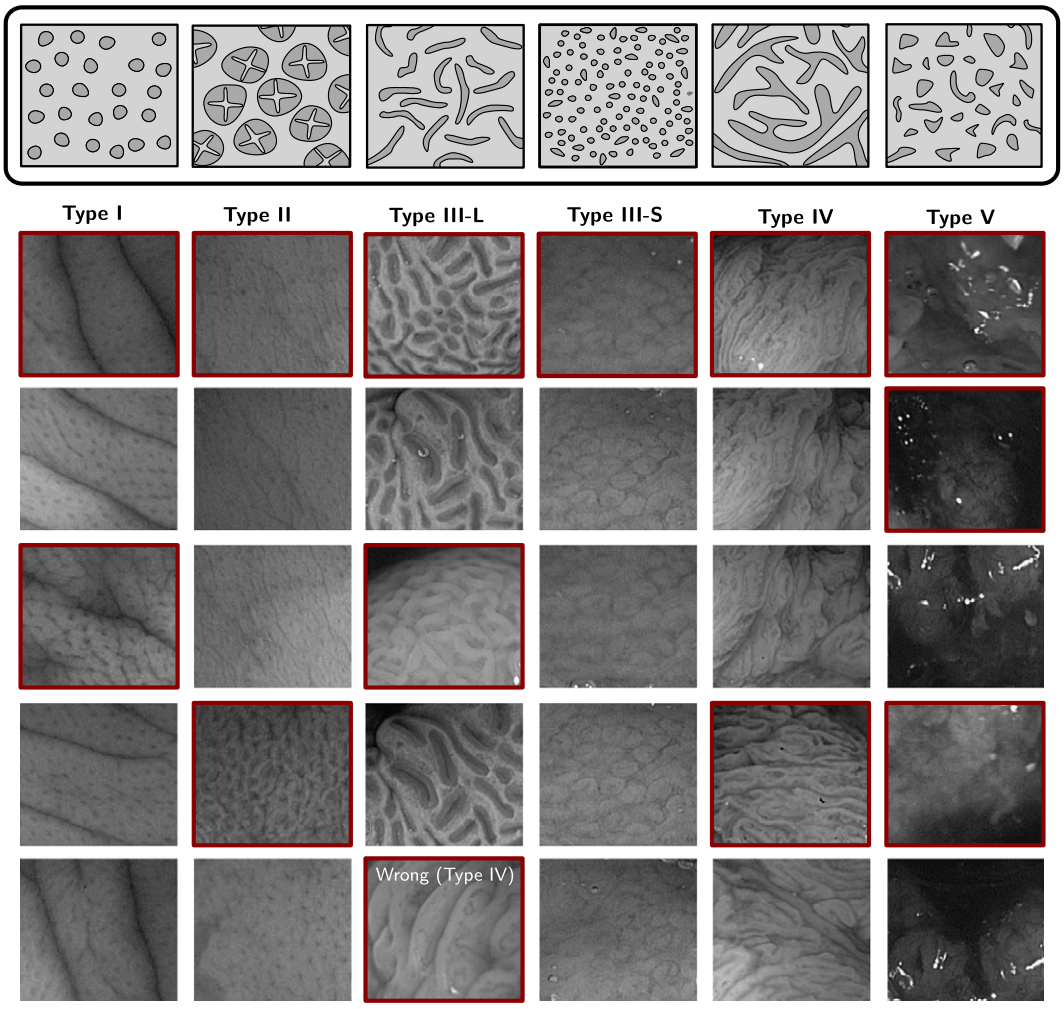

Fig. 5. Browsing result for querying the top $K = 5$ *most-characteristic* images per pit pattern. The subset of all images that remains after patient pruning is highlighted. The top row shows a schematic *textbook* visualization of the Kudo criteria for pit pattern discrimination (cf. Kudo et al., 1994): Type I is characterized by normal, round pits, type II by asteroid, stellar or papillary pits, type III-L by tubular or round pits (usually larger than type I), type III-S by tubular or round pits (usually smaller than type I), type IV by dendritic or gyrus-like pits and type V by irregular arrangements, or a compete loss of structure.

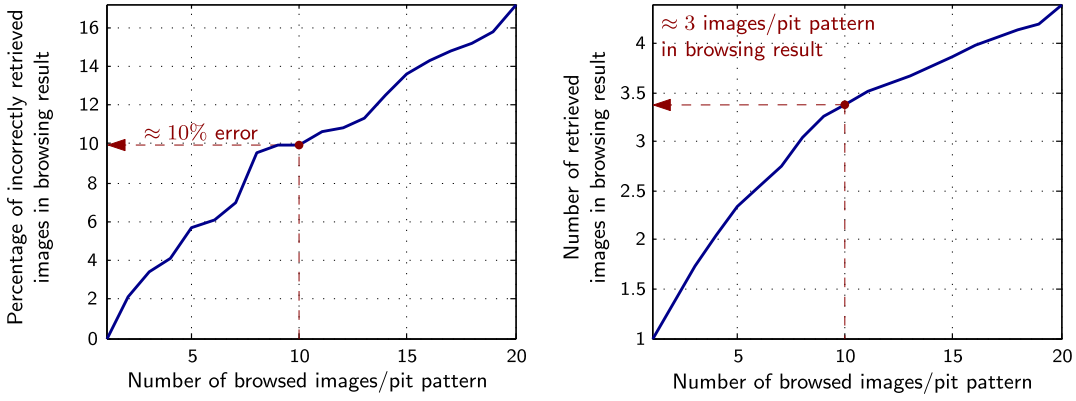

Fig. 6. Percentage of incorrectly retrieved images in the browsing result (i.e., incorrect pit pattern) when browsing for the $k = 1, \ldots, 20$ most-characteristic images per pit pattern (left); average number of images (from different patients) in that browsing result (right).

reason, we employ a gradient descent approach outlined by Pennec (2006). Under this method, the Frechét mean $\boldsymbol{\mu}_{k+1}$ at iteration $k+1$ is

$$\boldsymbol{\mu}_{k+1} = \exp_{\boldsymbol{\mu}_k}\left[\frac{1}{N}\sum_{i=1}^{N}\log_{\boldsymbol{\mu}_k}(\boldsymbol{a}_i)\right]. \tag{7}$$

$\boldsymbol{\mu}_0$ is a suitable initial value (e.g., chosen randomly from $\boldsymbol{a}_i$, $i = 1, \ldots, N$) and $\exp_{\boldsymbol{x}}$ and $\log_{\boldsymbol{x}}$ denote the Riemannian exponential and log map, respectively. Let $T_{\boldsymbol{a}}\mathcal{M}$ be the tangent plane at point $\boldsymbol{a}$, $\boldsymbol{v} \in T_{\boldsymbol{a}}\mathcal{M}$ the tangent vector, and $\gamma : [0, 1] \to \mathcal{M}$ the geodesic starting at $\boldsymbol{a}$ with velocity $\boldsymbol{v}$. The Riemannian exponential map $\exp_{\boldsymbol{a}} : T_{\boldsymbol{a}}\mathcal{M} \to \mathcal{M}, \boldsymbol{v} \mapsto \gamma(1)$ maps the tangent vector $\boldsymbol{v}$ to the end

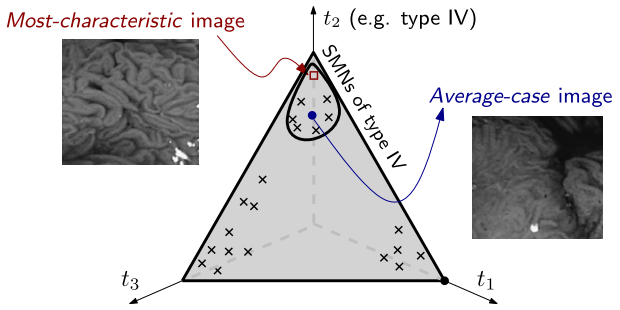

Fig. 7. Illustration of the difference between the *most-characteristic* image of a pit pattern type (here, type IV) and its *average-case*.

of the geodesic. The exponential map is a local diffeomorphism in a neighborhood $\mathcal{N}(\boldsymbol{a})$ of $\boldsymbol{a}$. Given that $\mathcal{N}(\boldsymbol{a})$ is the largest such neighborhood, the inverse mapping $\mathcal{N}(\boldsymbol{a}) \rightarrow T_{\boldsymbol{a}}\mathcal{M}$ is denoted the Riemannian log map $\log_{\boldsymbol{a}}$. Hence, (7) can by be interpreted as mapping points of the manifold onto the tangent plane at the current Frechét mean estimate, taking the expected value and performing the inverse mapping to the manifold.

To compute the Fréchet mean from the SMNs corresponding to images labeled with a particular concept, we can again exploit the isometry between $(\mathbb{P}^{C-1}, \mathscr{I})$ and $(\mathbb{S}^{C-1}, \delta)$ (cf. Section 3.2). The *SMN Frechét mean* is then computed as

$$\boldsymbol{\mu}_{k+1} = \exp_{\boldsymbol{\mu}_k}\left[\frac{1}{N}\sum_{i=1}^{N}\log_{\boldsymbol{\mu}_k}\frac{F(\boldsymbol{s}^i)}{\|F(\boldsymbol{s}^i)\|}\right]. \tag{8}$$

On the unit-sphere, the Riemannian exponential and log map are given by

$$\log_{\boldsymbol{x}}(\boldsymbol{y}) = \frac{\arccos(\langle \boldsymbol{x}, \boldsymbol{y}\rangle)}{\sqrt{1-\langle \boldsymbol{x}, \boldsymbol{y}\rangle^2}}(\boldsymbol{y} - \langle \boldsymbol{x}, \boldsymbol{y}\rangle \boldsymbol{x}) \tag{9}$$

$$\exp_{\boldsymbol{x}}(\boldsymbol{y}) = \cos(\|\boldsymbol{y}\|)\boldsymbol{x} + \sin(\|\boldsymbol{y}\|)\ \|\boldsymbol{y}\|^{-1}\boldsymbol{y} \tag{10}$$

Since $\boldsymbol{\mu}_k$ resides on the unit-sphere, it is necessary to project the Frechét mean back onto the simplex $\mathbb{P}^{C-1}$ to obtain the *average-case* SMN encoding $\bar{\boldsymbol{s}}$ as $\bar{\boldsymbol{s}} = \boldsymbol{\mu}_k^2$.

3.4.1. Experiments

The Frechét mean computation was used in combination with the geodesic distance to determine the *average-case* image per concept. The strategy is as follows: given the Frechét mean $\bar{\boldsymbol{s}}_t$ from the SMNs of the images labeled with concept $t \in \mathscr{T}$, it is possible to find the image I_{r_t} closest to the average-case, with respect to the geodesic distance on the semantic manifold, i.e.

$$r_t = \arg\min_{\boldsymbol{s}_i \in \mathscr{D}_t} d_{\mathscr{I}}(\bar{\boldsymbol{s}}, \boldsymbol{s}_i). \tag{11}$$

Fig. 8 shows the images closest to the average-case (upper part, top row) in terms of the Frechét mean as well as the Frechét mean of the corresponding SMNs (upper part, bottom row) itself, for each category. For comparison, the figure also shows the most-characteristic samples (lower part, top row) and corresponding SMNs (lower part, bottom row). When compared to the most-characteristic images, the introductory graphic of Fig. 1, or the browsing result of Fig. 5, the average-case images appear less typical of each pit pattern type. This reflects the observation that pit patterns rarely occur in a pure form, and mixtures of pit patterns are a very common occurrence in clinical practice.

4. Discussion

In this work, we have proposed a new approach for endoscopic image analysis. It was argued that focusing on discriminative appearance features to predict histological findings leads to *black-box* decision-support systems which might lack acceptance by the medical community. As a possible solution to this problem, we adopted a recent semantically-focused scene-recognition approach from computer vision to establish a semantic encoding of endoscopic images. Based on this encoding, and the induced semantic space, we demonstrated that classification, semantically-focused browsing, and the computation of class representatives can be implemented in a principled manner by drawing on results from information geometry. Experiments on a collection of high-magnification colonoscopy images have shown that classification in semantic space achieves accuracies similar to highly-optimized, appearance-feature based approaches with significantly lower feature space dimensionality. Moreover, the proposed strategy for browsing images by semantic typicality has been shown

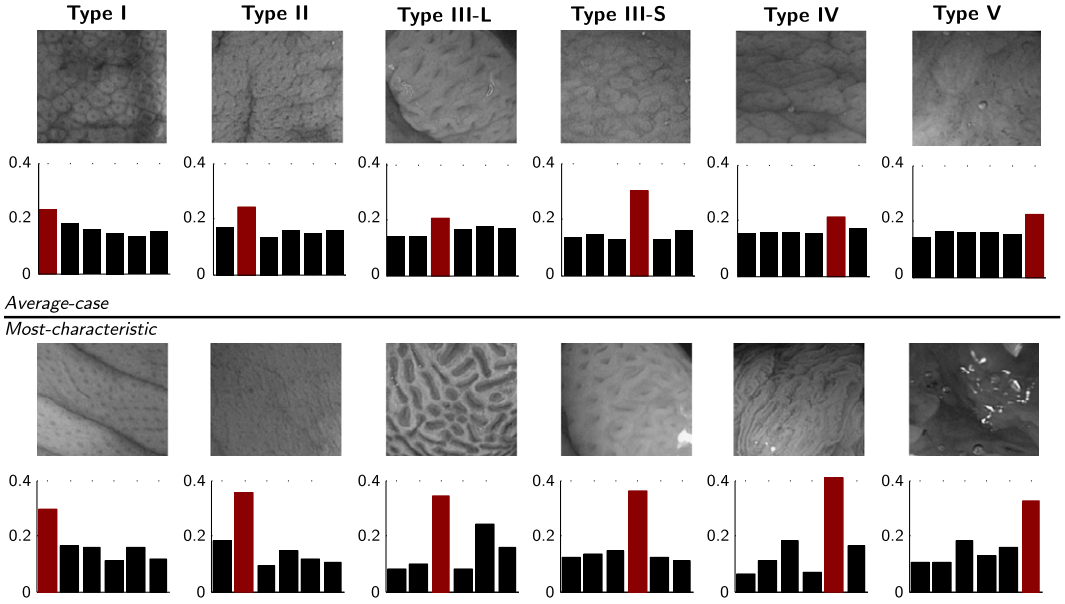

Fig. 8. Top: *average-case* image representatives per pit pattern type and corresponding Fréchet means. Bottom: *most-representative* images and corresponding SMNs.

capable of retrieving the most-characteristic images per concept with small browsing error. Finally, an analysis of the average-case images per concept revealed that the average-case may in fact be harder to interpret or categorize due to less distinctive structures compared to the textbook illustrations. In the future, we intend to develop a *difficulty* measure, similar to (André et al., 2010), for classifying endoscopic imagery, possibly based on some sort of average geodesic distance to the most-representative samples.

Acknowledgments

This work is sponsored, in part, by NIH/NIBIB sponsored *National Alliance of Medical Image Computing* (NA-MIC, PI: Kikinis, 1U54EB005149-01, http://www.na-mic.org), NIH/NCI sponsored *Image Registration for Ultrasound-Based Neurosurgical Navigation* (NeuralNav/TubeTK, PI: Aylward, Wells, 1R01CA138419–01, http://public.kitware.com/Wiki/NeuralNav) and NSF sponsored CCF-0830535 (PI: Vasconcelos).

Appendix A. Conditional positive-definiteness of Eq. (5)

We provide a thorough proof that the kernel defined in (5) is conditionally positive-definite (cpd).

Theorem 1 (*Power Series of Dot-Product Kernels (Smola et al., 2000, 2001)*). *Let $k : \mathbb{S}^{d-1} \times \mathbb{S}^{d-1} \to \mathbb{R}$ denote a dot-product kernel on the unit sphere in a d-dimensional Hilbert space. The kernel is positive definite (pd) if and only if there is a function $f : \mathbb{R} \to \mathbb{R}$ such that $k(x,y) = f(\langle x,y\rangle)$ and the coefficients of the Taylor series of f are non-negative, i.e.*

$$f(t) = \sum_{n=0}^{\infty} c_n t^n \text{ with } c_n \geqslant 0. \tag{A.1}$$

According to this theorem, it suffices to check the coefficients of the Taylor series expansion of f for non-negativity to proof that k is a pd kernel. Note, that by *dot-product* we refer to the standard scalar product in $\mathbb{R}^d$.

Proposition 1. *The semantic kernel*

$$k_s(\boldsymbol{x},\boldsymbol{y}) = -2\arccos(\langle\sqrt{\boldsymbol{x}},\sqrt{\boldsymbol{y}}\rangle) \quad \text{with } \boldsymbol{x},\boldsymbol{y} \in [0,1)^d \tag{A.2}$$

and $\|x\|_1 = 1$, $\|y\|_1 = 1$ is conditionally positive definite (cpd).

Proof. First, we note that for any $x \in [0,1)^d$ with $\|x\|_1 = 1$, the square-root $\sqrt{\boldsymbol{x}}$ (component-wise) resides on the unit sphere $\mathbb{S}^{d-1}$, embedded in d-dimensional Euclidean space $\mathbb{R}^d$. Given that $g{:}[0,1] \to [0,\pi]$ is defined as

$$g(t) = \pi - 2\arccos(t), \tag{A.3}$$

we can write the semantic kernel of (A.2) as

$$k_s(\boldsymbol{x},\boldsymbol{y}) = g(\langle\sqrt{\boldsymbol{x}},\sqrt{\boldsymbol{y}}\rangle) - \pi. \tag{A.4}$$

Our first step is to show that the dot-product kernel $k_g(\boldsymbol{x},\boldsymbol{y}) := g(\langle\sqrt{\boldsymbol{x}},\sqrt{\boldsymbol{y}}\rangle)$ induced by g satisfies the condition of Theorem 1 and is thus pd. Hence, we first build the Taylor series expansion of arccos (x) around 0, i.e.

$$\arccos(x) = \frac{\pi}{2} - \sum_{n=0}^{\infty} \frac{(2n)!}{2^{2n}(n!)^2}\frac{1}{2n+1}x^{2n+1} \quad \forall\, x : |x| < 1. \tag{A.5}$$

The convergence condition $|x| < 1$ is satisfied, as $|\langle x,y\rangle| < 1$ for x, $y \in [0,1)^d$. Next, we can write $g(t)$ as

$$g(t) = \sum_{n=0}^{\infty} c_n t^{2n+1} \text{ with } c_n = \frac{(2n)!}{2^{2n}(n!)^2}\frac{2}{2n+1} \tag{A.6}$$

and we immediately see $\forall n$: $c_n \geqslant 0$. It follows that the dot-product kernel k_g induced by g is pd. To show that the semantic kernel is cpd, we proceed as follows: we know from Schölkopf and Smola (2001) that each pd kernel is also cpd and that each real constant is cpd. According to the closure properties of pd kernels, which can be carried over to the class of cpd kernels, the sum of two cpd kernels is cpd as well. The fact that we can write the semantic kernel as

$$k_s(\boldsymbol{x},\boldsymbol{y}) = \underbrace{k_g(\boldsymbol{x},\boldsymbol{y})}_{cpd} + \underbrace{(-\pi)}_{cpd} \tag{A.7}$$

completes the proof. □

References

André, B., Buchner, T.V.A., Shahid, M., Wallace, M., Ayache, N., 2010. An image retrieval approach to setup difficulty levels in training systems for endomicroscopy diagnosis. In: MICCAI.
André, B., Buchner, T.V.A., Wallace, M., Ayache, N., 2011. A smart atlas for endomicroscopy using automated video retrieval. Med. Image Anal. 15, 460–476.
André, B., Vercauteren, T., Ayache, N., 2010. Endomicroscopic video retrieval using mosaicing and visual words. In: ISBI.
André, B., Vercauteren, T., Buchner, A., M. Wallace, M., Ayache, N., 2012. Learning semantic and visual similarity for endomicroscopy video retrieval. IEEE Trans. Med. Imag. 31, 1276–1288.
André, B., Vercauteren, T., Perchant, A., Buchner, A.M., Wallace, M.B., Ayache, N., 2009. Endomicroscopic image retrieval and classification using invariant visual features. In: ISBI.
Arthur, D., Vassilvitskii, S., 2007. k-means++: the advantages of careful seeding. In: SODA.
Boureau, Y.L., Bach, F., LeCun, Y., Ponce, J., 2010. Learning mid-level features for recognition. In: CVPR.
Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Stat. Soc. – Ser. B 39, 1–38.
Dixit, M., Rasiwasia, N., Vasconcelos, N., 2011. Adapted gaussian mixtures for image classification. In: CVPR.
Fei-Fei, L., Perona, P., 2005. A bayesian hierarchical model for learning natural scene categories. In: CVPR.
Grauman, K., Darrell, T., 2005. Pyramid match kernels: Discriminative classification with sets of image features. In: ICCV.
Häfner, M., Liedlgruber, M., Uhl, A., Gangl, M., Vecsei, A., Wrba, F., 2009. Pit pattern classin+ncation using extended local binary patterns. In: ITAB.
Häfner, M., Liedlgruber, M., Uhl, A., Vecsei, A., Wrba, F., 2012. Color treatment in endoscopic image classification using multi-scale local color vector patterns. Med. Image Anal. 16, 75–86.
Häfner, M., Liedlgruber, M., Uhl, A., Wrba, F., Vécsei, A., Gangl, A., 2010. Endoscopic image classification using edge-based features. In: ICPR.
Kudo, S., Hirota, S., Nakajima, T., Hosobe, S., Kusaka, H., Kobayashi, T., Himori, M., Yagyuu, A., 1994. Colorectal tumours and pit pattern. J. Clin. Pathol. 47, 880–885.
Kwitt, R., Rasiwasia, N., Vasconcelos, N., Uhl, A., Häfner, M., Wrba, F., 2011. Learning pit pattern concepts for gastroenterological training. In: MICCAI.
Kwitt, R., Uhl, A., Häfner, M., Gangl, A., Wrba, F., Vécsei, A., 2010. Predicting the histology of colorectal lesions in a probabilistic framework. In: MMBIA.
Lazebnik, S., Schmid, C., Ponce, J., 2006. Beyond bags of features: Spatial pyramid matching for recognizing scene categories. In: CVPR.
Lebanon, G., 2005. Riemannian Geometry and Statistical Machine Learning. Ph.D. thesis, Carnegie Mellon University.
Lowe, D., 2004. Distinctive image features from scale-invariant keypoints. IJCV 60, 91–110.
Mäenpää, T., Pietikäinen, T., Viertola, M., 2002. Separating color and ppattern information for color texture discrimination. In: ICPR.
Maron, O., 1998. Multiple-instance learning for natural scene classification. In: ICML.
Pennec, X., 2006. Intrinsic statistics on riemannian manifolds: basic tools for geometric measurements. J. Math. Imag. Vis. 25, 127–154.
Rasiwasia, N., Moreno, P., Vasconcelos, N., 2006. Query by semantic example. In: ACM CIVR.
Rasiwasia, N., Vasconcelos, N., 2008. Scene classification with low-dimensional semantic spaces and weak supervision. In: CVPR.
Schölkopf, B., 2000. The kernel trick for distances. In: NIPS.
Schölkopf, B., Smola, A., 2001. Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press, Cambridge, MA, USA.
Schwaninger, A., Vogel, J., Hofer, F., Schiele, B., 2006. A psychophysically plausible model for typicality ranking of natural scenes. ACM Trans. Appl. Percept. 3, 333–353.
Smola, A., Ovari, Z., Williamson, R., 2000. Regularization with dot-product kernels. In: NIPS.
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.
Tung, S.Y., Wu, C.S., Su, M.Y., 2001. Magnifying colonoscopy in differentiating neoplastic from nonneoplastic lesions. Am. J. Gastroenterol. 96, 2628–2632.