Learning Pit Pattern Concepts for Gastroenterological Training

Roland Kwitt[1,⋆], Nikhil Rasiwasia[2], Nuno Vasconcelos[2], Andreas Uhl[1], Michael Häfner[3], and Friedrich Wrba[4]

[1] Dept. of Computer Sciences, Univ. of Salzburg, Austria
[2] Dept. of Electrical and Computer Engineering, Univ. of California, San Diego, USA
[3] Dept. of Internal Medicine, St. Elisabeth Hospital, Vienna, Austria
[4] Dept. of Clinical Pathology, Vienna Medical Univ., Austria

Abstract. In this article, we propose an approach to learn the characteristics of colonic mucosal surface structures, the so called *pit patterns*, commonly observed during high-magnification colonoscopy. Since the discrimination of the pit pattern types usually requires an experienced physician, an interesting question is whether we can automatically find a collection of images which most typically show a particular pit pattern characteristic. This is of considerable practical interest, since it is imperative for gastroenterological training to have a representative image set for the textbook descriptions of the pit patterns. Our approach exploits recent research on semantic image retrieval and annotation. This facilitates to learn a semantic space for the pit pattern concepts which eventually leads to a very natural formulation of our task.

1 Motivation

Over the past few years there has been considerable research in computer-based systems to guide *in vivo* assessment of colorectal polyps, using endoscopic imaging. This research is motivated by the prevalence of colorectal cancer, one of the three most commonly diagnosed forms of cancer in the US, and its high mortality rate. Following the concept of the *adenoma-carcinoma sequence* [11], colorectal cancer predominantly develops from adenomatous polyps, although adenomas do not inevitably become cancerous. In fact, the resection of colorectal adenomas reduces the incidence of colorectal cancer. In this context, it is safe to say that the ultimate objective of image analysis is to distinguish *neoplastic* from *non-neoplastic* lesions, although finer grained discriminations are obviously possible. While early approaches (e.g. [5]) to computer-assisted dignity assessment were based on visual data from conventional white-light endoscopes, research has shifted towards novel imaging modalities. These include narrow band imaging (NBI, e.g. [9]), high-magnification chromo-endoscopy (HMCE e.g. [4]), and probe-based confocal laser endomicroscopy (e.g. [1]). The emergence of these

⋆ This work is partially funded by the Austrian Science Fund (FWF) under Project No. L366-N15.

G. Fichtinger, A. Martel, and T. Peters (Eds.): MICCAI 2011, Part III, LNCS 6893, pp. 280–287, 2011.

novel imaging modalities has made it challenging for gastroenterologists to interpret the acquired imagery. In order to prevent serious mistakes (e.g. perforation of the colon, etc.), substantial experience with the particular imaging modality and the highlighted tissue structures is still necessary, especially in situations where the physician's assessment differs from that of the decision support system (which is still an uncommon tool in clinical practice).

In this work, we tackle the problem of preparing prospective gastroenterologists for clinical practice with the novel imaging modalities. We argue that, during gastroenterological training, it is imperative to have (i) access to a database of labeled images from the prospective imaging modality and (ii) possibility to browse through images depicting the *textbook description* of a particular structure. In the absence of a computer vision system to assemble these images, an experienced gastroenterologist will typically have to work through a vast image repository, to sort out the most relevant training examples. We propose a computer vision solution to this problem, based on recent advances in semantic image retrieval [8]. This is a formulation of image database search, where images are mapped onto a *semantic space* of image *concepts*.

While, in computer vision, concepts are usually cars or buildings, the idea can be applied to the pit pattern classes commonly used in the medical literature for prediction of histopathological results (cf. [3]). Unlike [8], we are not interested in the strict task of retrieval by semantic example. Instead, our work is directed to the semantic browsing scenario. This is the scenario where gastroentrologists are able to browse the image space efficiently by focusing on regions where particular concepts, i.e. pit patterns, are most prominent. We propose a system that enables this type of *semantically focused browsing*. Although our approach is generic, we demonstrate its applicability in the context of HMCE and Kudo's pit pattern analysis scheme [6].

The technical details of pit pattern browsing are given in the following section. Section 3 is devoted to experimental results and Sect. 4 presents our conclusions.

2 Learning the Pit Pattern Concepts

The starting point of our approach is a database of endoscopy images $\mathcal{D} = \{I_1, \ldots, I_{|\mathcal{D}|}\}$ and a collection of concepts $\{w_1, \ldots, w_C\}$, i.e. the pit pattern types. We require that each database image is augmented by a binary *caption* vector $\boldsymbol{c}_y \in \{0,1\}^C$, where $c_y^j = 1$ signifies that the j-th concept is present in image I_y. This is termed a *weakly* labeled set of images, since $c_y^j = 0$ does not necessarily mean that the j-th concept is not present. We further don't know which image region contains the annotated concept (i.e. no prior segmentation available). In fact, weak labeling is carried to the extreme, since the caption vectors only contain one non-zero entry for the prominent concept. This follows from the fact that the medical labeling procedure is based on reconciliation with histopathological ground-truth. For example, if the laboratory results indicate a normal gland and the gastroenterologist has the visual impression of a pit pattern type I, then the image is labeled with that type. However, a pit pattern of type II

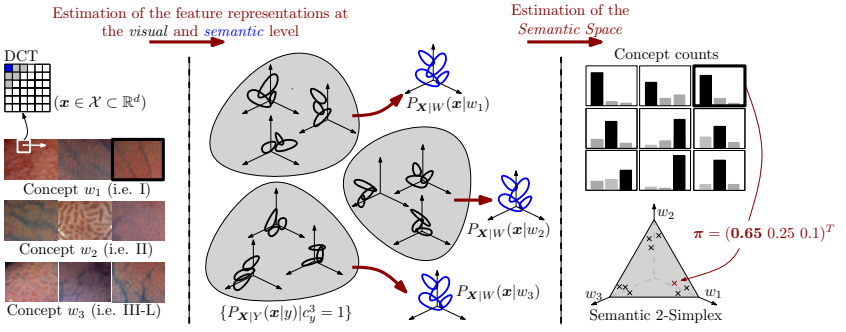


Fig. 1. Processing pipeline for learning three exemplary pit pattern *concepts* $\{w_1, w_2, w_3\}$ (e.g. I, II, III-L). First, we decompose each image into a collection of *localized* features. Then, we estimate (i) the *semantic-level* feature representations $\{P_{\boldsymbol{X}|W}(\boldsymbol{x}|w_i)\}_{i=1}^3$ from the *visual-level* feature representations and (ii) the mapping from feature space to *semantic space*, i.e. the semantic 2-simplex embedded in $\mathbb{R}^3$.

(or any other type) can also be visible in some areas of the image. While this labeling strategy guarantees that the annotated pit pattern is visible to some extent, at least to the experienced gastroenterologist, many of the images do not convey the textbook description [6] corresponding to the labeled pit pattern.

2.1 Image Representation at the Visual and Semantic Level

The first stage for learning the image concepts is similar to previous studies (e.g. [7]), where automated dignity assessment was the primary objective. Each image I in the database is represented by a collection of *localized* features $I = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ drawn independently from a random vector $\boldsymbol{X}$, defined in some feature space $\mathcal{X} \subset \mathbb{R}^d$. This stage is illustrated in the leftmost part of Fig. 1. Defining a random variable Y (with realizations in $\{1, \ldots, |\mathcal{D}|\}$) such that $Y = y$ when features are drawn from image I_y, the probability of image I at the *visual* level is

$$P_{\boldsymbol{X}|Y}(I|y) = \prod_{j=1}^{N} P_{\boldsymbol{X}|Y}(\boldsymbol{x}_j|y) \ . \tag{1}$$

The density (i.e. the generative model) $P_{\boldsymbol{X}|Y}(\boldsymbol{x}|y)$ for image I_y is estimated by a K_V-component multivariate Gaussian mixture

$$P_{\boldsymbol{X}|Y}(\boldsymbol{x}|y) = \sum_{k=1}^{K_V} \gamma_y^k \mathcal{G}(\boldsymbol{x}; \boldsymbol{\mu}_y^k, \boldsymbol{\Sigma}_y^k) \text{ with } \sum_k \gamma_y^k = 1 \ , \tag{2}$$

based on the corresponding collection of features.

In contrast to [7], where captions are neglected, and images retrieved by visual similarity (query-by-visual-example), the captions now represent a key component of the system. Introducing a random variable W (with realizations in

$\{1, \ldots, C\}$) such that $W = i$ when features are drawn from concept w_i, induces a new collection of probability densities $\{P_{\boldsymbol{X}|W}(\boldsymbol{x}|w_i)\}_{i=1}^{C}$ on $\mathcal{X}$. These densities are denoted the feature representations at *semantic* level. Assuming conditional independence of the features given concept membership, the *concept-conditional* probability of image I at the *semantic level* is

$$P_{\boldsymbol{X}|W}(I|w) = \prod_{j=1}^{N} P_{\boldsymbol{X}|W}(\boldsymbol{x}_j|w) \ . \tag{3}$$

Similar to density estimation at the visual level, we use multivariate Gaussian mixtures with K_S components to estimate $P_{\boldsymbol{X}|W}(\boldsymbol{x}|w)$, i.e.

$$P_{\boldsymbol{X}|W}(\boldsymbol{x}|w) = \sum_{l=1}^{K_S} \alpha_w^l \mathcal{G}(\boldsymbol{x}; \boldsymbol{\nu}_w^l, \boldsymbol{\Phi}_w^l) \text{ with } \sum_l \alpha_w^l = 1 \ . \tag{4}$$

Modeling the densities at the visual *and* semantic level by Gaussian mixtures has the convenient advantage that we can exploit the hierarchical mixture modeling approach of [10] to estimate the mixture parameters at the semantic level $\{\alpha_w^l, \boldsymbol{\nu}_w^l, \boldsymbol{\Phi}_w^l\}$ from the mixture parameters at the visual level $\{\lambda_y^k, \boldsymbol{\mu}_y^k, \boldsymbol{\Sigma}_y^k\}$. This step is visualized in the middle of Fig. 1, where the mixtures associated with several images in a class are summarized by the class' single semantic-level mixture. Note that the number of Gaussian components at semantic level ($C \times K_S$) is considerably smaller than the number of Gaussian components at visual level ($|\mathcal{D}| \times K_V$). Figure 1 illustrates the case where $K_V = 3$ and $K_S = 4$. The computational effort to estimate the semantic-level mixtures, using the method of [10], is also considerably smaller than that required for direct estimation of $P_{\boldsymbol{X}|W}(\boldsymbol{x}|w)$ based on the pooled features of all images annotated with concept w.

2.2 Learning the Semantic Space

The identification of images which most characteristically depict a particular concept requires a semantic image representation with explicit control over the concepts. In [8], the authors introduce the idea of an image as a point on a *semantic space*. The image is first modeled as a vector $I_y = (n_y^1, \ldots n_y^C)^T$ of concept counts (cf. top right part of Fig. 1), where n_y^k is the number of feature vectors in the y-th image drawn from the k-th concept. The concept count vectors are then modeled as realizations of a multinomial random variable $\boldsymbol{T}$. As illustrated on the bottom right of Fig. 1, the parameter vector $\boldsymbol{\pi}_y = (\pi_y^1, \ldots, \pi_y^C)^T$ of the multinomial distribution associated with the image is a point on the standard $(C-1)$-simplex, since $\sum_i \pi^i = 1$. This simplex is denoted the *semantic space*, and $\boldsymbol{\pi}_y$ the *semantic multinomial (SMN)* associated with image I_y.

The question is how to estimate the mapping based on a database of tuples $\{(I_y, \boldsymbol{c}_y)\}_{y=1,\ldots,|\mathcal{D}|}$. For that purpose, we employ a modification of the semantic multiclass labeling approach of [2] which implements the mapping based on an estimation of posterior concept probabilities, i.e.

$$\pi_y^w = P_{W|\boldsymbol{X}}(w|I_y) = \frac{P_{\boldsymbol{X}|W}(I_y|w) P_W(w)}{P_{\boldsymbol{X}}(I_y)} \ . \tag{5}$$

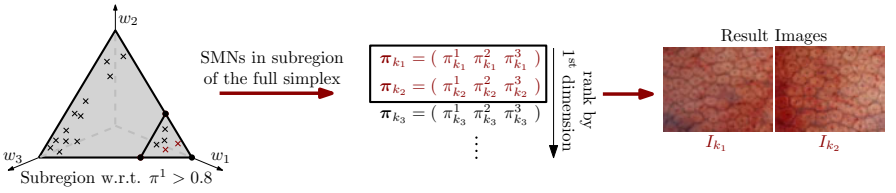


Fig. 2. Identifying the images, represented by SMNs, which most typically represent the concept w_1 (here pit pattern type I)

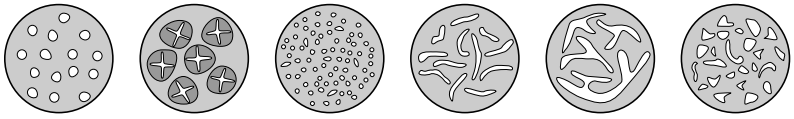
Although it is possible to directly estimate π_y^w by assuming a uniform concept prior $P_W(w)$ and estimating $P_{\boldsymbol{X}}(I_y)$ by $\sum_w P_{\boldsymbol{X}|W}(I_y|w)P_W(w)$, we choose an alternative approach to cover for numerical instabilities. The strategy is to compute $\log P_{W|\boldsymbol{X}}(w|\boldsymbol{x}_j)$ for *each* feature vector $\boldsymbol{x}_j$ separately, using (4), then determine the concept with the largest posterior probability per feature vector and eventually tally the occurrences of the *winning* concepts. This facilitates Maximum-Likelihood (ML) parameter estimation of $\boldsymbol{\pi}_y$ through $\forall w : \tilde{\pi}_y^w = n_y^w \cdot 1/N$. In [8], the authors further suggest regularization with a Dirichlet prior, which leads to a maximum-a-posteriori estimate

$$\hat{\pi}_y^w = \frac{\tilde{\pi}_y^w + \pi_0}{\sum_{i=1}^{C}(\tilde{\pi}_y^i + \pi_0)}, \tag{6}$$

where π_0 is a regularization parameter. The remaining question is how to identify the desired set of images from the semantic representation in the form of points on the semantic space. Given that we aim to identify the images most typical of the concept w_i (e.g. pit pattern III-L), we only need to navigate on the simplex. In fact, we can easily identify a subregion of the full simplex whose SMNs represent images where the w_i-th concept is prominent with probability t, by using $\pi^i > t, t \in [0, 1]$. Fig. 2 illustrates this idea for $\pi^1 > 0.8$. Sorting the SMNs in that region along the i-th dimension gives a list of the most representative (i.e. top-ranked) images for concept w_i.

3 Experimental Evaluation

We evaluate our proposed approach on a weakly labeled database of 716 HMCE images (magnified 150×) of size 256 × 256 pixel, captured by an Olympus Evis Exera CF-Q160ZI/L endoscope. The images stem from a total of 40 patients and the database contains only images where the histological ground truth is coherent with the annotated pit pattern type. The cardinalities of the image sets per concept are $\{124, 74, 124, 20, 276, 98\}$ for pit pattern types I, II, III-L, III-S, IV and V. A graphical and textual description of the pit pattern characteristics is provided in Fig. 3. Apart from the gastroenterologist's experience, these descriptions represent the *textbook* material for dignity assessment of colorectal lesions. The images are converted from RGB to YBR color space for further processing.



(a) I (b) II (c) III-S (d) III-L (e) IV (f) V

Fig. 3. Schematic illustration of the six pit pattern types according to [6]. The typical characteristics are (a) normal, round pit; (b) asteroid, stellar or papillary pit; (c) tubular or round pit (smaller than type I); (d) tubular or round pit (larger than type I); (e) dendritic or gyrus-like pit; (f) irregular arrangements, loss of structure.

As *localized* features, we use DCT coefficients (extracted in zigzag scan order) of 8×8 patches, obtained from a sliding window, moving by two pixel increments in both image dimensions (cf. Fig. 1). We extract the first 16 coefficients (including the DC coefficient) from the same patch across the color channels and arrange the coefficients in feature vectors according to a YBRYBRYBR... interleaving pattern. The Gaussian mixtures to model $P_{\boldsymbol{X}|Y}(\boldsymbol{x}|y)$ are fitted by the classic Expectation-Maximization (EM) algorithm. The number of mixture components at this level is set to $K_V = 8$ and we restrict the covariance matrices to diagonal form. At the semantic level, we set $K_S = 64$ and estimate the parameters using the hierarchical estimation approach of [10]. Regarding SMN estimation, we choose a regularization parameter $\pi_0 = 1/6$, although experiments show that the approach is not sensitive to this choice.

To demonstrate that we can actually identify images which most typically depict the textbook pit pattern descriptions, we sort the SMNs on the semantic simplex along the dimension corresponding to each concept and extract the K top-ranked images. We further ensure that the extracted images do *not* belong to the same patient in order to establish a realistic scenario. We refer to this step as *patient pruning* of the result set. Figure 4 shows the images after pruning the $K = 10$ top-ranked images per pit pattern concept. Due to the fact that the database images are not uniformly distributed over the patients, the pruning step has the effect that the cardinality of the final browsing result per concept is not equal. Nevertheless, a comparison to the illustrations and descriptions in Fig. 3 reveals the correspondences we were looking for: we observe the characteristic gyrus-like structures of pit pattern IV, the round pits of pit pattern I, or the complete loss of structure in case of pit pattern V for instance.

Besides visual inspection of the results in Fig. 4, we conduct a more objective evaluation by exploiting the ground-truth caption vectors for each image. In particular, we evaluate the *average error rate* of the system when browsing the K top-ranked images per concept. We perform a *leave-one-patient-out (LOPO)* evaluation, where we adhere to the following protocol per patient: (i) remove the patient's images from the database, (ii) estimate the SMNs based on the remaining images, (iii) extract the K top-ranked images (now using the whole database) per concept and (iv) perform the patient pruning step. The average error rate is then calculated as the percentage of images (averaged over all LOPO

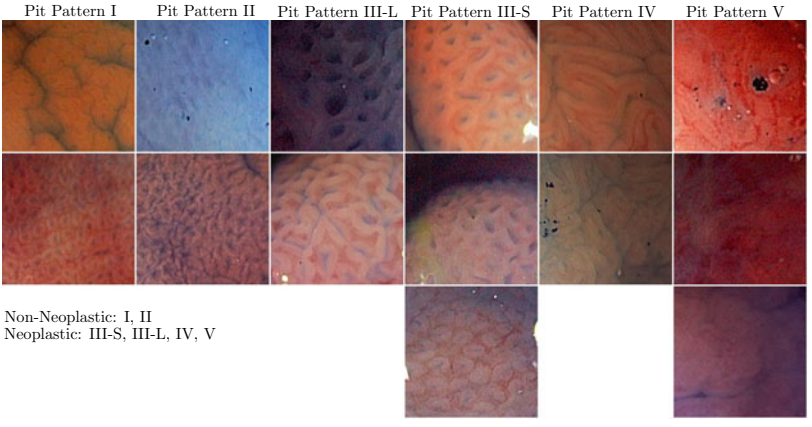


Fig. 4. Result of identifying the most representative images for each pit pattern concept at the operating point $K = 10$ (*with* patient pruning)

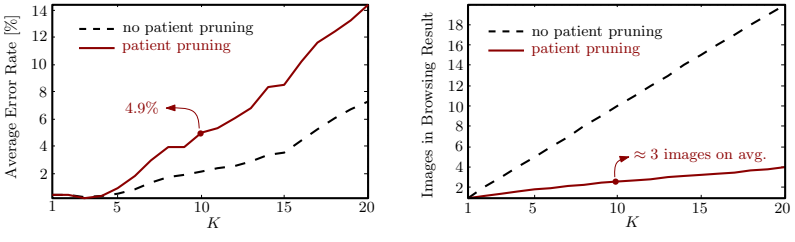


Fig. 5. Average error rate with respect to the ground-truth caption vectors, when browsing the K top-ranked images of each pit pattern type

runs) in the final browsing result of concept w_i which do not belong there according to the corresponding ground-truth caption vectors (i.e. zero entry at the i-th position). Figure 5 shows the average error rate in dependence of K *with* and *without* patient pruning (for comparative reasons). At the operating point $K = 10$ for instance, we obtain three images per concept on average at an error rate of 4.9%. This corresponds to ≈ 0.88 wrong images in the final browsing result.

4 Concluding Remarks

Motivated by the need to provide prospective gastroenterologists with a collection of images showing the most typical characteristic of a particular mucosal structure during endoscopy, we presented a generic approach to establish a semantic space on a database of weakly labeled HMCE images. To the best of our knowledge, introducing the notion of a semantic domain to that problem has not been done so far. On the basis of Kudo's pit pattern analysis scheme,

we demonstrated that browsing the semantic space in *interesting* regions in fact allows to isolate the most characteristic images for each pit pattern type.

References

1. André, B., Vercauteren, T., Perchant, A., Buchner, A.M., Wallace, M.B., Ayache, N.: Endomicroscopic image retrieval and classification using invariant visual features. In: Proceedings of the IEEE International Symposium on Biomedical Imaging: From Nano to Macro (ISBI 2009), Boston, MA, USA, pp. 346–349 (June 2009)
2. Carneiro, G., Vasconcelos, N.: Formulating semantic image annotation as a supervised learning problem. In: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR 2005), San Diego, CA, USA, pp. 163–168 (June 2005)
3. East, J.E., Suzuki, N., Saunders, B.P.: Comparison of magnified pit pattern interpretation with narrow band imaging versus chromoendoscopy for diminutive colonic polyps: A pilot study. Gastrointest. Endosc. 66(2), 310–316 (2007)
4. Häfner, M., Gangl, A., Kwitt, R., Uhl, A., Vécsei, A., Wrba, F.: 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.) MICCAI 2009. LNCS, vol. 5761, pp. 247–254. Springer, Heidelberg (2009)
5. Karkanis, S.A., Iakovidis, D., Karras, D., Maroulis, D.: Detection of lesions in endoscopic video using textural descriptors on wavelet domain supported by artificial neural network architectures. In: Proceedings of the IEEE International Conference on Image Processing (ICIP 2001), Thessaloniki, Greece, pp. 833–836 (October 2001)
6. Kudo, S., Hirota, S., Nakajima, T., Hosobe, S., Kusaka, H., Kobayashi, T., Himori, M., Yagyuu, A.: Colorectal tumours and pit pattern. J. Clin. Pathol. 47(10), 880–885 (1994)
7. Kwitt, R., Uhl, A., Häfner, M., Gangl, A., Wrba, F., Vécsei, A.: Predicting the histology of colorectal lesions in a probabilistic framework. In: Proceedings of the IEEE International Workshop on Mathematical Methods in Biomedical Image Analysis (MMBIA 2010), San Francisco, CA, USA, pp. 103–110 (June 2010)
8. Rasiwasia, N., Moreno, P., Vasconcelos, N.: Bridging the gap: Query by semantic example. IEEE Trans. Multimedia 9(5), 923–938 (2007)
9. Tischendorf, J.J.W., Gross, S., Winograd, R., Hecker, H., Auer, R., Behrens, A., Trautwein, C., Aach, T., Stehle, T.: Computer-aided classification of colorectal polyps based on vascular patterns: a pilot study. Endoscopy 42(3), 203–207 (2010)
10. Vasconcelos, N., Lippman, A.: Image indexing with mixture hierachies. In: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR 2001), Kauai, HI, USA, pp. 3–10 (December 2001)
11. Vogelstein, B., Fearon, E.R., Hamilton, S.R., Kern, S.E., Preisinger, A.C., Leppert, M., Nakamura, Y., White, R., Smits, A.M., Bos, J.L.: Genetic alterations during colorectal-tumor development. N. Engl. J. Med. 319(9), 525–532 (1988)