 DOI: 10.3217/978-3-85125-603-1-09

CNN training using additionally training data extracted from frames of endoscopic videos

Georg Wimmer[1] and Michael Haefner[2] and Andreas Uhl[1]

Abstract— **Insufficient amounts of labeled training data poses a big problem in machine learning, especially for medical applications where medical image data sets are usually quite small. In this work we propose a method to increase the amount of labeled endoscopic image data in order to improve the classification accuracy of automated diagnosis systems for the classification of colonic polyps. Starting from a small colonic polyp endoscopic image database, we increase the number of images by tracking the content shown in the images through the endoscopic videos and by extracting patches from frames of the videos that show the same content as in the images of the database, but under different viewing conditions. By means of our proposed method we are able to increase the amount of labeled image data by factor 40, without adding images of insufficient image quality or images without clearly visible features for the differentiation of colonic polyps. We will show that this increased amount of training images can drastically improve the performance of CNNs, which are state-of-the-art in the automated classification of colonic polyps.**

I. Introduction

Modern endoscopy devices are able to take images and videos from inside the colon which facilitates computer-assisted analysis of the acquired material with the goal of detecting and diagnosing abnormalities.

Usually, endoscopic image databases consist of image patches that are manually extracted from images routinely captured during endoscopy or from manually chosen frames of endoscopic videos. The image patches of the databases show regions of interest with clearly visible mucosa structures and/or geometric features that enable a differentiation between healthy and affected mucosa (in our case we differentiate between different classes of colonic polyps). The labels for those extracted image patches are provided by medical experts.

Although videos are routinely recorded during endoscopy, the video material cannot be used for the training of automated diagnosis systems since there are no labels given for the mucosal regions shown in the videos, except for the ones where images of frames were manually extracted and labeled. Furthermore, for large parts of the video the image quality is insufficient to enable a classification of the shown mucosal regions.

In this work we propose a method that generates additional labeled image data with sufficient image quality by tracking the regions shown in the manually extracted patches (with given label information) throughout the video. By automatically extracting image patches of those regions from endoscopic video frames we generate additional image data with given label information. Those new image patches show the same regions as shown in the original, manually extracted image patches, but under different viewing conditions (different scales and viewpoints), with different image qualities and potentially also with different imaging modalities (image enhancement technologies like e.g. i-Scan modes can be switched on and off during endoscopy). The final step of our proposed method filters out all image patches with insufficient image quality. Contrary to previous approaches assessing the informativeness of frames in colonoscopic videos [1], [2], we do not only focus on image blur as quality measures but also on the visibility of mucosal texture structures and discard all images without clearly visible texture structures.

To test if our approach to increase the number of training images is suited for automated diagnosis systems, we apply them to the old (manually extracted) and to the new, enlarged database and compare the classification results of the two databases. More specifically, we train convolutional neural networks (CNNs) using both databases and compare their classification accuracy.

Convolutional neural networks are state-of-the-art in the automated diagnosis of colonic polyps and outperform handcrafted image representations as shown in [3], [4]. Generally, thousands or millions of images are used and required as data corpus to achieve well generalizing deep architectures. In endoscopic image classification however, the available amount of data usable as training corpus is often much more limited to a few hundreds or thousands of images or even less. By means of our proposed method to increase the amount of labeled training data we aim to overcome this issue and train nets that perform better and are less overfitted to the training data.

This work presents two contributions:

- We propose a method that fully automatically generates labeled endoscopic image data. The additional image data is extracted from frames of endoscopic images and those images with insufficient image quality are discarded. To the best of our knowledge, this has not been done before in literature.
- We train CNNs on the old, original database and the new, enlarged database and compare their results.

This work was supported by the Austrian Science Fund, TRP Project 206.

[1] G. Wimmer and A. Uhl are with the University of Salzburg, Department of Computer Sciences, Jakob Haringerstrasse 2, 5020 Salzburg, Austria {gwimmer, uhl}@cosy.sbg.ac.at

[2] M. Haefner is with the St. Elisabeth Hospital, Landstraßer Hauptstraße 4a, A-1030 Vienna, Austria

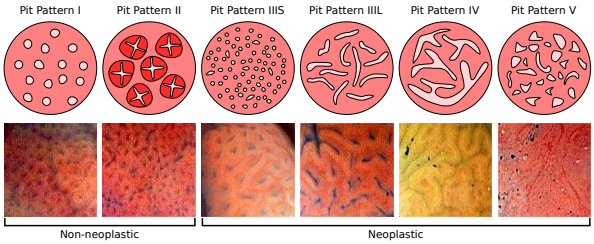


Fig. 1. The 6 pit pattern types along with exemplar images and their assigned classes in case of a two class (non-neoplastic vs neoplastic) differentiation

II. Colonic Polyps

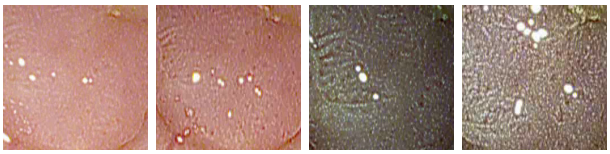
Colonic polyps are a rather frequent finding and are known to either develop into cancer or to be precursors of colon cancer. Colonic polyps are usually divided into hyperplastic, adenomatous and malignant polyps. In order to determine a diagnosis based on the visual appearance of colonic polyps, the pit pattern classification scheme was proposed by [5]. A pit pattern refers to the shape of a pit, the opening of a colorectal crypt. The various pit pattern types and exemplar (HM-endoscopic) images of the classes are presented in Fig 1. The pit pattern classification scheme differentiates among six types. Type I (normal mucosa) and II (hyperplastic polyps) are characteristics of non-neoplastic lesions, type III-S, III-L and IV are typical for adenomatous polyps and type V is strongly suggestive to malignant cancer.

In this work we use the two-classes classification scheme differentiating between non-neoplastic and neoplastic lesions. This classification scheme is quite relevant in clinical practice as indicated in [6].

Our original colonic polyp image database consists of manually extracted patches from frames of HD colonoscopic videos with high image quality. The patches are recorded using either white light (WL) endoscopy or the i-Scan technology. The i-Scan (Pentax) image processing technology [7] is a digital contrast enhancement method which consists of combinations of surface enhancement, contrast enhancement and tone enhancement. The three i-Scan modes operate as follows:

1) i-Scan1 augments pit pattern and surface details, providing assistance to the detection of dysplastic areas. This mode enhances light-to-dark contrast by obtaining luminance intensity data for each pixel and adjusting it to accentuate mucosal surfaces.
2) i-Scan2 expands on i-Scan1 by adjusting the surface and contrast enhancement settings and adding tone enhancement attributes to the image. i-Scan2 assists by intensifying boundaries, margins, surface architecture and difficult-to-discern polyps.
3) i-Scan3 is similar to i-Scan2, with increased illumination and emphasis on the visualization of vascular features. This mode accentuates pattern and vascular architecture.

In Fig. 2 we see an image showing an adenomatous polyp with WL endoscopy (a) and i-Scan (b,c,d)



(a) WL (b) i-Scan 1 (c) i-Scan 2 (d) i-Scan 3

Fig. 2. Images of a polyp using WL endoscopy and different i-Scan modes.

The i-Scan modes and WL can be switched on and off by the endoscopist during colonoscopy. Most of the videos contain sequences with all 4 imaging modalities (WL i-Scan 1,2,3).

The high definition (HD) colonic polyp image database was acquired by extracting patches of size $256 \times 256 \times 3$ from frames of HD-endoscopic (Pentax HiLINE HD+ 90i Colonoscope) videos. The database consists of patches gathered with 4 different imaging modalities (three different i-Scan modes (modes 1,2,3) and WL endoscopy). The database consists of 478 image patches (144 images showing non-neoplastic mucosa and 301 images showing neoplastic polyps) from 84 patients.

III. Frames from Endoscopic Videos

In this section we propose our unsupervised method to extract high quality image patches with label information from frames of endoscopic videos.

A. Movement Estimation

In endoscopic videos, the video capture device (the endoscope) is moving through the colon. So contrary to movies, its alone the video capture device that is moving and not the objects that are shown in the video. Since the movements are often very fast, it is quite hard to track the position of the regions shown in the endoscopic video. The rapid movements of the endoscope cause movement blur and fast changes of the distances from the camera to the mucosal wall (which causes sharp transitions from well focused sections of the video to sections of the video that are completely out of focus and hence quite blurry). Furthermore, there are no hard edges in the mucosal images and big parts of the videos are recorded out of focus because the distance from the endoscope to the mucosal wall is very often too high and sometimes also too low to be in the optimal focus range of the camera. That means big parts of the videos appear blurry. Additionally, the imaging modalities (WL, i-Scan 1,2,3) can change from one frame to the other in the videos.

This all makes it quite difficult to reliably track objects in the videos. Furthermore, major parts of the video do not exhibit enough image quality to be used to effectively differentiate between different types of polyps.

For the patches of our endoscopic image database, label information were provided by medical experts. As already mentioned in the introduction, we want to automatically generate additional image patches with label information. By tracking the region shown in the original image patch we are able to take more images of the considered region of interest.

So, our first task is to track the content shown in a 256 x 256 sized patch throughout the video (forward and backward through the video).

Since we are facing highly complex motion (e.g. position-variant transformations and parallax effects) in endoscopic videos, simple motion models are not sufficient to describe the motions between successive HD endoscopy video frames.

In this work we use the optical flow estimation by Black and Anandan [8], which is part of the implementation available for the work in [9]. This method is quite versatile when it comes to the estimation of arbitrary complex motions between images. This is mainly due to the fact that optical flow methods allow to estimate local motion, while simpler methods usually work well only with global motion.

We use the following notations to describe our proposed method: f_0 denotes the frame from which an image patch from our endoscopic image database was manually extracted, $p_0(\vec{x_0})$ denotes the manually extracted patch from frame f_0 and $\vec{x_0} = (x_0, y_0)$ denotes its position inside frame f_0 (the coordinate of the middle point of the patch). f_i denotes the i-th frame starting from f_0 (either forwards or backwards through the video) and $p_i(\vec{x_i})$ denotes the patch extracted from f_i, where x_i is the tracked position of the region shown in $p_0(\vec{x_0})$. The optical flow estimation is applied from frame (f_{i-1}) to frame (f_i) and not from f_0 to f_i because of the distinctly more accurate movement estimations in our experiments for estimating the movement from frame to frame. Movement estimation is applied to gray scale versions of the frames and image patches.

Although optical flow estimation usually works quite fine to track the content shown in a patch from frame to frame, it can fail in case of extreme motions, extreme image blur, and changing imaging modalities. Furthermore, errors in the movement estimation would add up the longer we track a region through the video.

To avoid any errors of the optical flow estimations, we apply a correlation based movement estimation as backup if the image patch $p_{i-1}(x_{i-1})$ is too different to the subsequent patch $p_i(x_i)$. More specifically, if the correlation coefficient between the patch $p_{i-1}(\vec{x}_{i-1})$ and patch $p_i(\vec{x}_i)$ (where $\vec{x}_i$ is the position tracked from $p_{i-1}(\vec{x}_{i-1})$ by means of the optical flow estimation) is beneath a threshold $C_t = 0.8$ (the correlation coefficient can range from +1 (for two identical patches) to -1 (for a patch and its inverse version)), then the position $\vec{x}_i$ of the patch p_i is re-evaluated by selecting the position that leads to the highest correlation coefficient with patch $p_{i-1}(\vec{x}_{i-1})$:

$$\vec{x_i} = (x_i, y_i) = \max_{(x,y)}(corr(p_{i-1}(x_{i-1}, y_{i-1}), p_i(x,y))), \quad \text{with}$$

$$corr(p_{i-1}(x_{i-1}, y_{i-1}), p_i(x,y)) = corr(q^{i-1}, q^i) = \frac{\sum_{m=1}^{256}\sum_{m=1}^{256}(q_{mn}^{i-1} - \overline{q}^{i-1})(q_{mn}^{i} - \overline{q}^{i})}{\sqrt{(\sum_{m=1}^{256}\sum_{m=1}^{256}(q_{mn}^{i-1} - \overline{q}^{i-1})^2)(\sum_{m=1}^{256}\sum_{m=1}^{256}(q_{mn}^{i} - \overline{q}^{i})^2)}},$$

where q_{mn}^i denotes the gray value of the pixel in the image patch q^i with position (m,n) and $\overline{q}^i$ denotes the medium gray value over all pixels in q^i.

This way of tracking is more time consuming than the optical flow estimation but it is also more accurate in the presence of strong blur and/or high camera movement.

Furthermore, we have four stop conditions to avoid errors in tracking the content of the patches. If one of the following stop conditions applies, then we stop tracking the position of the content in the patches any further since the risk of incorrect movement estimations becomes too high:

1) We stop if the estimated movement from one frame to the next one exceeds 50 pixels (Euclidean distance $d(x_{i-1}, x_i) > 50$).
2) We stop if two successive patches p_{i-1} and p_i differ too strongly. For this, we resize both patches to size 32×32. If the difference between the gray values of the downsized patches exceeds 10 in average, then the patches are considered as too different. This stop condition is applied to detect cuts in the video or to stop tracking if the content shown in the video changes too fast to enable a reliable tracking.
3) If the tracked position x_i of the patch p_i is so far outside of the frame f_i that the patch exceeds the border of the frame.
4) We stop at latest at the 300th iteration (maximal 300 iterations forward through the video and maximal 300 iterations backward through the video). Since the frame rate of the videos is 25 frames per second, this corresponds to tracking the content shown in the original patch p_0 for at most 12 seconds backwards and forwards through the video starting from frame f_0. We set that maximum number of iterations ($i \leq 300$) to avoid that small individual errors in the movement estimation sum up to a more significant error in the movement estimation.

We ensured by manual inspection of the automatically extracted patches, that all additionally extracted patches show the same content as shown in the original patch (p_0) (but with different viewpoints and scales). Hence, the label information of each original patch also applies to all patches p_i originated from the original patch p_0.

B. Image Quality Control

In order to differentiate between different types of polyps, the polyps and their pit-pattern structure have to be clearly visible. As already mentioned before, major parts of the video does not exhibit enough image quality to enable a differentiation between different types of polyps.

To ensure that only those automatically extracted image patches are further used to train automated diagnosis systems that enable a correct diagnosis, some criteria were determined to differentiate between images with high enough image quality and those images that are discarded because of limited image quality. The image quality tests are applied to grayscale versions of the (originally RGB) image patches. The following threshold values to differentiate between informative and non-informative patches were set so that the qual-

ity assessment of the automatically extracted image patches widely corresponds with the authors subjective opinion.

1) Reflections: If the number of overexposed pixels (gray value > 240) in a newly extracted patch exceeds 3500 (that is about one of 19 pixels), then the image patch is classified as uninformative.
2) Darkness: If the number of underexposed pixels (gray value < 45) in a newly extracted patch exceeds 4000 (that is about one of 15 pixels), then the patch is classified as uninformative.
3) Blur and visibility of texture structures: The image patch is subdivided into 10×10 pixel regions and standard deviations are computed for each of those regions. The highest (the top 20%) and the lowest (the bottom 20%) standard deviations are omitted and the mean value of the remaining standard deviations ($\overline{std}$) is computed as quality measure. The highest standard deviations are omitted because those outliers would heavily influence the mean value and since reflections can cause high standard deviations in the 10x10 pixel regions. The lowest standard deviations are omitted since it is not necessary that texture structures are clearly visible everywhere in the patch. It is sufficient if most parts of a patch are informative but it does not pose a problem if small parts of a patch are recorded out of focus. If $\overline{std}(p_i) < 5$, then the patch is classified as uninformative.
4) Blur and pit pattern structure: Our quality measure to rate the visibility of texture structures like the pit pattern structure is based on difference of Gaussians (DoG). We apply DoG by subtracting a Gaussian blurred image ($\sigma = 1$, filter size 5×5) from a stronger Gaussian blurred image ($\sigma = 3$ and filter size 9×9). Then each pixel value of the DoG image is replaced by its absolute value. The resulting non-negative DoG image of a grayscale endoscopic image patch highlights mucosal structures like the pit-pattern structure. Similar to the standard deviations and because of the same reasons, the highest 10% and the lowest 10% of the DoG values are omitted and the mean value ($\overline{DoG}$) is computed of the remaining DoG values. If $\overline{DoG}(p_i) < 2$, then the patch is classified as uninformative.
5) Comparison to the reference image patch p_0: Frames recorded with the i-Scan imaging modality show clearly more contrast and a better visability of mucosal structures than those frames that were recorded with traditional white light endoscopy. Furthermore, images showing healthy mucosa usually show less contrast than those images showing adenomatous polyps. So the quality measurements do not only respond to the quality of the frames but also to the used imaging modality and the shown content. There are even some original image patches that do not fulfill all of the before mentioned criteria (most of them show healthy mucosa and were captured using WL endoscopy). So we introduce an additional quality measure that balances the image patch quality with reference to the quality of the original image patch p_0. If an image patch p_i does not fulfill one of the before mentioned quality thresholds, but if the considered quality measure of the image patch p_i is at most 5% worse than the quality measure of the original patch p_0, then the image patch p_i is still classified as informative. We do not want to throw away patches that are only very slightly worse in terms of image quality than the reference patch p_0. On the other hand, if an image patch p_i is clearly more blurry and if the mucosal texture structures are clearly less visible than for the original image patch ($\overline{DoG}(p_i) < 0.6 \times \overline{DoG}(p_0)$ or $\overline{std}(p_i) < 0.6 \times \overline{std}(p_0)$), then the image patch p_i is classified as non-informative, even if it fulfills all before mentioned criteria.
6) Movement: If the estimated movement $d(x_{i-1}, x_i) > 15$, then the patch p_i is classified as uninformative. Image patches with higher movement almost always suffer from movement blur.
7) Duplicity: If $corr(p_i, p_{i-1}) > 0.95$, then p_i is classified as uninformative. If there is hardly any difference between two patches than we only use one of them. This step will be later motivated in Section IV.

In Figure 3 we show some examples of informative and non-informative patches. On the left side we see examples that were accepted as informative image patches and on the right side we see examples that originate from the same original patches as the image patches to their left and that were discarded because of insufficient image quality.

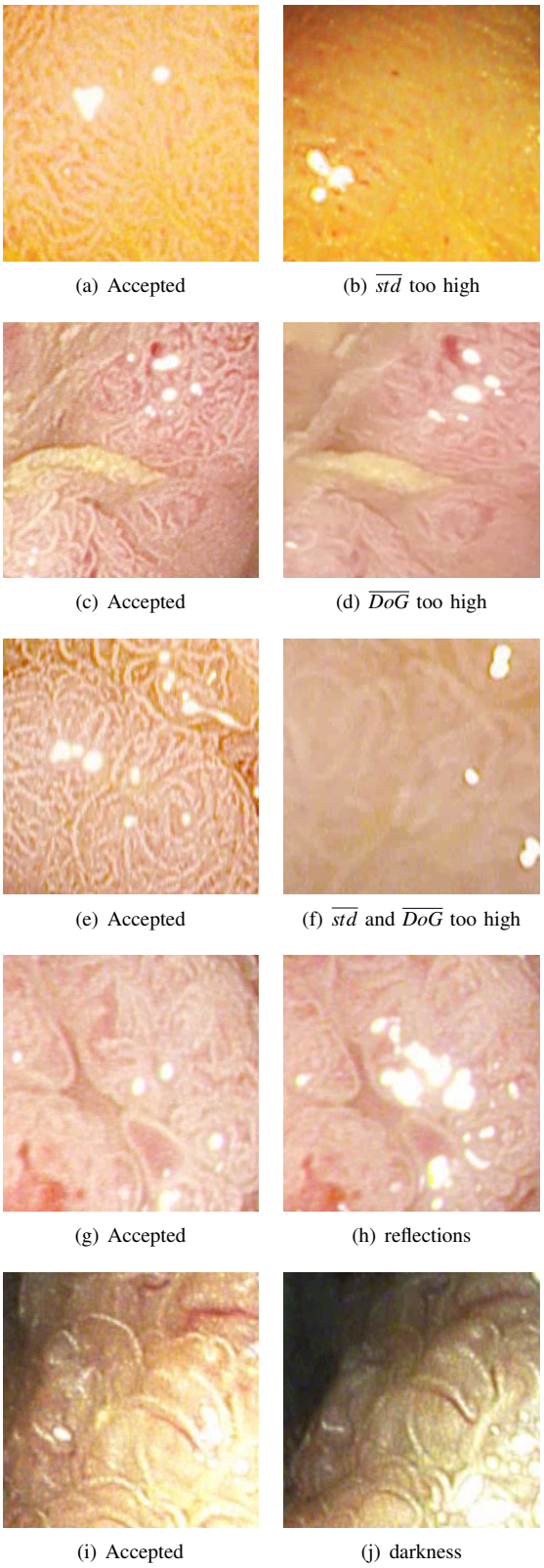
The original HD colonic polyp image data base consists of 478 image patches. The enlarged version of the database using our proposed method includes 18969 image patches. So in average, about 39 image patches were generated from one patch of the original database. In case of 40 original patches, no additionally patches with sufficient image quality could be generated. The maximum number of generated image patches from one original patch is 270. The standard deviation over the generated patches per original frame is 43, so there are huge variations in the number of additionally generated image patches per original patch.

IV. CNN TRAINING

This section gives the implementation details for CNN training and the description of the employed nets.

We employ two nets in this work, the VGG-f net [10] and the VGG-16 network [11]. The VGG-f net consists of five convolutional layers and three fully connected layers with a final SoftMax classifier. The VGG-16 net consists of 13 convolutional layers subdivided in 5 convolutional blocks (where each of the 2-3 convolutional layers inside of a block have the same number and sizes of filters) and three fully connected layers with a final SoftMax classifier.

The two nets are trained from scratch and we randomly initialize the coefficients of the layers based on [12]. The last fully connected layer is acting as soft-max classifier and



(a) Accepted (b) $\overline{std}$ too high

(c) Accepted (d) $\overline{DoG}$ too high

(e) Accepted (f) $\overline{std}$ and $\overline{DoG}$ too high

(g) Accepted (h) reflections

(i) Accepted (j) darkness

Fig. 3. Examples of informative (left) and uninformative (right) patches originating from the same original patch.

computes the training loss (log-loss). Stochastic gradient descent (SGD) with weight decay ($\lambda = 0.0005$) and momentum ($\mu = 0.9$) is used for the training of the models.

As already mentioned in the introduction, CNN training is applied to two different image databases. We use the original colonic polyp image database and the new enlarged database for training the CNNs.

Training is performed on batches of 128 images each, which are for each iteration randomly chosen from the training data and subsequently augmented (see Section V).

In case of the original database, the 128 images are for each iteration randomly chosen from the training data.

In case of the enlarged database, the image patches that originate from one original image patch are quite similar and we need to consider that fact for the selection of training images per batch. If we would randomly choose images for training like for the original database, then the images of some patients would be used very often for training (those with a lot of automatically extracted image patches), whereas images of other patients (those with a low number of automatically extracted image patches) would be used much less for training. On the other hand, if we would randomly choose the patients and then randomly choose one image per patient for training, then we would not really profit from the high amount of additional image data since the chance that a specific image is chosen for training from a patient with a high number of automatically extracted patches is very low (compared to an image of a patient with a low number of automatically extracted patches). So we decided to first randomly choose one of the original patches, whereat the probability of those original patches varies with the amount of images that originate from them. More specifically, the probability to choose one original patch is multiplied by factor f_n with $f_n = \sqrt[3]{n}$, where n is the number of patches originating from one original patch (including the original patch). After one original patch is chosen, we randomly choose an image patch that is originating from the considered original patch (including the original patch itself), where each image patch has the same chance to be chosen. So for example, if the number of image patches originating from original patch A is 100 ($n(A) = 100$) and 1 for patch B ($n(B) = 1$, only the original patch itself), then the probability of choosing any image belonging to patch A for training is $\sqrt[3]{100} = 4.64$ times higher than choosing the image belonging to patch B. On the other hand, the probability that the one image of patch B is chosen for training is higher by factor $100/4.64 = 21.54$ than the probability that one specific image of patch A is chosen. This approach to select the training images was the reason to discard image patches that are very similar to other image patches (the duplicity criteria in Section 3).

V. Experimental Setup

Our employed nets require input image sizes of $224 \times 224 \times 3$. The image data is normalized by subtracting the mean image of the training portion. We then linearly scale each image within $[-1, 1]$.

We use data augmentation to increase the number of images for training and validation. Augmentation is applied

CNN architecture	Training Database	
	Original Database	Enlarged Database
VGG-16	76.2 (6.6)	86.9(7.2)
VGG-f	83.9(5.7)	84.2(3.8)

TABLE I

MEAN ACCURACIES OVER THE 10 FOLDS AND THE STANDARD DEVIATIONS (IN BRACKETS) FOR THE TWO NETS ON BOTH DATABASES

to the batches of images extracted for training. The augmentation is based on cropping one sub-image (224×224 pixels) from each image patch with randomly chosen position. Subsequently, the sub-image is randomly rotated (0°, 90°, 180° or 270°) and randomly either flipped or not flipped around the horizontal axis. Validation is performed using a majority voting over five crops from the validation image using the upper left, upper right, lower left, lower right and center part.

We perform 10-fold cross-validation to achieve a stable estimation of the generalization error, where each of the 10 subsamples of an image database consists of the images from about 10% of the patients of a database. All images of one patient are in one subsample and each subsample consists of about 10% of the total images of a database (at least in case of the original database there are about 10% of the images in one subsample, in case of the enlarged database the number of images per subsample can vary depending on the number of additionally extracted patches per patient). All nets are trained using the training portion of our data corpus (9 of the 10 subsamples). The final validation is performed on the left-out part. That means for each database and for each of the two network architectures, ten different nets are trained, one for each of the 10 folds. To ensure the highest possible comparability between the results of the two databases, we used the same folds for both databases. That means the training data corpus of a fold consists of image data from the same patients for both databases (whereat the patients contain distinctly more images in case of the enlarged database). For both databases, validation is performed on the validation data corpus of the original database for each fold. That means validation is always performed on the same images for both databases (to have comparable results), whereas training is always performed on a much bigger data corpus in case of the enlarged database, but from images of the same patients as for the original database.

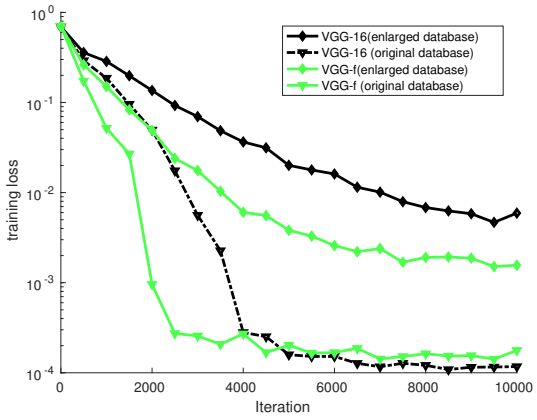
In our experiments, we compute the overall classification rate (OCR) for each fold and report the mean OCR over all 10 folds with the respective standard deviation.

A. Results and Discussion

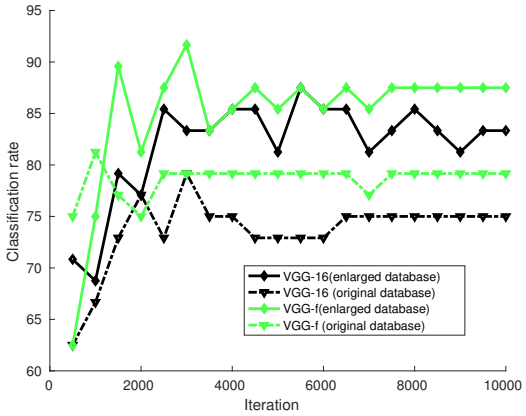
The results of the experiments using our two nets trained on the original colonic polyp database as well as trained on the enlarged version of the database are presented in Table I.

As we can see in Table I, the VGG-16 net clearly profits from additional training data (86.9% vs 76.2%). The VGG-f net on the other hand did not really profit from the additional training data. The results only increased by 0.3% using the enlarged database. A possible reasons for the different outcomes of the two nets is that the VGG-16 net has much more layers and parameters to be learned. The original database was not large enough to properly train this big net. The much smaller VGG-f net on the other side was not able to profit that much from the additional training data. The overall quality of the automatically extracted image patches is slightly worse compared to the original patches. So we guess that the difference in the quality of the training images (from the enlarged database) compared to the evaluation images (from the original database) is the reason for the only very small improvement of the results for the VGG-f net using the enlarged image database for training.

In Figure 4 we see the the training losses and the validation accuracies during training (fold 1 of 10) for the two nets on both databases. We can observe that for both net



(a) loss



(b) accuracy

Fig. 4. Comparison of the training losses and the validation accuracies during training on the two databases.

architectures, the training loss decreases much faster on the smaller original database and reaches much lower levels at the end of training as for training on the enlarged database. The validation accuracies stagnate earlier for training on the original database as for training on the enlarged database.

This all indicates that the nets are more overfitted to the training data corpus in case of the smaller original database.

VI. Conclusion

In this work we presented an approach to increase the amount of labeled image data by a fully automated system that extracts image patches of endoscopic video frames from mucosal regions where label information is available. In that way we were able to increase the number of images by factor 40 and hence distinctly increase the amount of training data for automated diagnosis systems. Care was taken that only images with sufficient image quality and clearly visible mucosal texture structures were extracted. We showed that the increased number of training images can drastically improve the performance of CNNs. The mean accuracy of the VGG-16 net increased from about 76 % using the original database to nearly 87% using the enlarged image database for training. The results of the much smaller VGG-f net did only slightly improve using the additional training data. We furthermore showed that the increased number of training images reduces the overfitting to the training data corpus.

References

[1] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang, "Convolutional neural networks for medical image analysis: Full training or fine tuning?" *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1299–1312, May 2016.

[2] J. Oh, S. Hwang, J. Lee, W. Tavanapong, J. Wong, and P. C. de Groen, "Informative frame classification for endoscopy video," *Medical Image Analysis*, vol. 11, no. 2, pp. 110 – 127, 2007. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S136184150600079X

[3] M. H. Eduardo Ribeiro, Andreas Uhl, "Colonic polyp classification with convolutional neural networks," in *Proceedings of the 29th IEEE International Symposium on Computer-Based Medical Systems (CBMS'16)*, June 2016, pp. 253–258.

[4] E. Ribeiro, A. Uhl, G. Wimmer, and M. Häfner, "Exploring deep learning and transfer learning for colonic polyp classification," *Computational and Mathematical Methods in Medicine*, vol. 2016, p. Article ID 6584725, 2016.

[5] S.-E. Kudo, S. Hirota, T. Nakajima, S. Hosobe, H. Kusaka, T. Kobayashi, M. Himori, and A. Yagyuu, "Colorectal tumours and pit pattern," *Journal of Clinical Pathology*, vol. 47, pp. 880–885, 1994.

[6] S. Kato, K.-I. Fu, Y. Sano, T. Fujii, Y. Saito, T. Matsuda, I. Koba, S. Yoshida, and T. Fujimori, "Magnifying colonoscopy as a non-biopsy technique for differential diagnosis of non-neoplastic and neoplastic lesions," *World Journal of Gastroenterology*, vol. 12, no. 9, pp. 1416–1420, March 2006.

[7] S. Kodashima and M. Fujishiro, "Novel image-enhanced endoscopy with i-scan technology," *World Journal of Gastroenterology*, vol. 16, no. 9, pp. 1043–1049, 2010.

[8] M. Black and P. Anandan, "The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields," *Computer Vision and Image Understanding*, vol. 63, pp. 75–104, 1996.

[9] S. Deqing, S. Roth, and M. Black, "Secrets of optical flow estimation and their principles," in *Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'10)*, 2010, pp. 2432–2439.

[10] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," in *British Machine Vision Conference, BMVC 2014, Nottingham, UK, September 1-5*, 2014.

[11] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[12] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *In Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS10)*, vol. 9, 2010, pp. 249–256.