


Endoscopic Image Classification and Retrieval using Clustered Convolutional Features

Jamil Ahmad¹  · Khan Muhammad¹ · Mi Young Lee¹ · Sung Wook Baik¹

Received: 19 July 2017 / Accepted: 8 October 2017
© Springer Science+Business Media, LLC 2017

Abstract With the growing use of minimally invasive surgical procedures, endoscopic video archives are growing at a rapid pace. Efficient access to relevant content in such huge multimedia archives require compact and discriminative visual features for indexing and matching. In this paper, we present an effective method to represent images using salient convolutional features. Convolutional kernels from the first layer of a pre-trained convolutional neural network (CNN) are analyzed and clustered into multiple distinct groups, based on their sensitivity to colors and textures. Dominant features detected by each cluster are collected into a single, layout-preserving feature map using a spatial maximal activator pooling (SMAP) approach. A moving window based structured pooling method then captures spatial layout features and global shape information from the aggregated feature map to populate feature histograms. Finally, individual histograms for each cluster are combined into a single comprehensive feature histogram. Clustering convolutional feature space allow extraction of color and texture features of varying strengths. Further, the SMAP approach enable us to select dominant discriminative features. The proposed features are compact and capable of conveniently outperforming several existing features extraction approaches in retrieval and classification tasks on endoscopy images dataset.

Keywords Image retrieval · Features extraction · Convolution · Classification · Spatial pooling · Endoscopy

Introduction

With the advent of minimally invasive surgical procedures through endoscopy, the volume of images in multimedia databases at hospitals have grown to unprecedented levels [1]. This ever-growing multimedia big data makes it increasingly difficult to retrieve relevant information in an efficient and reliable manner [2]. In recent decades, content based image retrieval (CBIR) methods have been used to locate relevant images in large image collections, based on content similarity of a query image and images stored in the database [3]. This pair-wise image matching is the core of each CBIR system, which determines similarity between image pairs based on their features. Typically, images are represented using their contents by extracting features from colors, textures, shapes, and spatial layout. Several methods have been proposed to effectively model images as feature vectors. However, most of these methods rely on low-level features extraction which fail to model high level semantics in images, and are often plagued with high dimensionality, making indexing and matching processes highly inefficient [4]. These drawbacks often restrict their application to large scale data in real environments.

Traditionally, visual features were algorithmically extracted from images using their color or texture contents, spatial layout, shape features, or transform domain features [5]. Some of these methods include color structure descriptor (CSD) [6] which captures color features and their spatial layout into a compact histogram. A small window is convolved over the entire image and the corresponding histogram bins of the colors appearing in that window are incremented. Another

This article is part of the Topical Collection on *Image & Signal Processing*

✉ Sung Wook Baik
sbaik@sejong.ac.kr

¹ Digital Contents Research Institute, Sejong University, Seoul, Republic of Korea

approach known as color difference histogram (CDH) [7] encodes uniform color difference features by exploiting color and edge orientations. Liu et al. introduced multi-texton histogram (MTH) [8] which attempts to capture texture features by modeling images as collections of small texture elements called texton. Features were aggregated into a histograms which were used to represent color images. This work was extended by the same authors in [9], who presented micro-structure descriptor (MSD) which integrates low-level color and texture features using micro-structures and edge orientation features. In structure elements descriptor (SED) [10], the authors used several small patterns which were detected in images and pooled into histograms to represent images. Multi-scale local structure patterns (MS-LSP) histogram [11] approach used multi-scale versions of images and detected 20 distinct patterns at three different scales. The gathered features were weighted with spatial saliency map and collected in a feature histogram. Besides these approaches, powerful and robust features were also developed like scale invariant features transform (SIFT) [12] and speeded-up robust features (SURF) [13] to represent images. These features were later used in deriving representation schemes like bag-of-visual-words (BoVW) [14–16] for image retrieval systems. In the context of medical image representation and classification, Wang et al. [17] used dual-tree complex wavelet transform features with twin support vector machine to detect pathological brain diseases in MRI images. Zhang et al. solved the same problem using synthetic minority oversampling along with extreme learning machines [18]. In [19], endoscopic image classification was carried out with local binary patterns and neural network. Recently, Wang et al. [20] used pseudo Zernike moments as rotation invariant shape features to detect Alzheimer disease. All these methods analyze local regions in images and extract visual features to perform image retrieval or classification. The main problems associated with CBIR systems relying on hand-crafted features is ineffective content modeling, high feature dimensionality, and extraction of irrelevant and less useful features. Further, a majority of these methods construct global representations for images by combining local features without considering spatial layout information in an effective manner, which leads to low retrieval performance. In addition, the hand-crafted color and texture feature extraction methods only work with the dataset for which they are designed. It becomes difficult to generalize these features to be used with other types of images.

In recent years, hand-engineered features have been overshadowed by learned representations. Thanks to the availability of huge amounts of data, powerful computing facilities like GPGPUs, and intelligent algorithms like deep learning which can automatically learn features from raw data. Deep convolutional neural networks (CNN) [21, 22], deep denoising autoencoders (DAE) [23, 24], and Siamese CNNs [25] have performed significantly well in visual recognition

tasks including object recognition, image retrieval, and image segmentation [26, 27], etc. Due to their overwhelming performance, these hierarchical architectures, particularly CNNs have attracted significantly large research community, working in both computer vision, and non-vision domains. However, they are often used as black box and researchers are striving to understand their internal representation for utilizing them more effectively for vision-related tasks [28, 29]. Efforts are underway to further advance the performance of these methods, which require thorough understanding of these architectures.

In this paper, we study the convolution feature space using the kernels from the first layer of a pre-trained convolutional neural network (CNN) known as AlexNet [21] to represent, classify, and retrieve images from endoscopy image archives. We attempt to study the characteristics of these kernels and derive a compact and powerful image representation using clustered convolutional feature space approach. Further, we devise a simple method to pool convolutional features compactly, capturing spatial layout characteristics without increasing the feature dimensions. Through experimental evaluations, we show that the proposed method outperforms several existing state-of-the-art hand-engineered feature extraction methods on a challenging dataset.

The rest of the paper is organized as: “[Image representation in convolutional feature space](#)” Section presents the proposed method and illustrates the features extraction method using spatial pooling in clustered convolutional feature space. Experimental evaluations are carried out in “[Experiments and Results](#)” Section and the paper concludes in “[Conclusion and future work](#)” Section with advantages and limitations of the proposed scheme and also provides future research directions.

Image representation in convolutional feature space

Convolutional neural networks have been thoroughly investigated for image retrieval. Usually, the neuronal activations of the fully connected layers, or dimensionally reduced convolutional features are known to perform well in image retrieval systems [30–32]. Activation maps in the convolutional layers of deep CNNs contain lots of information and utilizing all the activation values as feature maps yield very high dimensional feature vectors. Further, the deeper layers learns discriminative features regarding images on which the network is trained. Each neuron in the deeper layers become sensitive to certain objects of their parts. Though these features are regarded as generic features, some images (for instance, endoscopy images) may not contain any object or part which could render the deeper features less useful. In such cases, the relatively shallower layers may be more suitable for features extraction which are more generic than the deeper layer features. In this

work, we studied the convolutional features from the first convolutional layer of AlexNet model pre-trained on ImageNet dataset [21]. This layer consists of 96 kernels of size 11×11 , and we believe that the 96 feature maps effectively model visual contents in images. We analyzed the convolutional feature space using two different characteristics of kernels including color-sensitivity and texture-sensitivity. A simple method has been devised to measure the color and texture sensitivity of convolution kernels using cross channel diversity, and channel-wise spatial diversity. Using these two characteristics of kernels, we attempted to cluster them into multiple groups. Each group of kernels is separately applied to the input image to generate feature maps. Spatial maximal activator pooling strategy [33] is then used to aggregate the feature maps into a single feature map. The contents of this feature map are then pooled using an effective structured pooling method. The derived representation is compact and possesses high discriminative capabilities. Further details of the various components in the proposed framework are provided in the subsequent sections.

Convolutional kernels sensitivity to colors & textures

Convolution kernels of the first layer in AlexNet are known to model basic color and texture features in images. Computing their sensitivity to colors and textures can help in understanding the types of features these kernels detect. In this regard, we present simple methods to measure the sensitivity of kernels to colors and textures using the following.

$$CS_i = \sum_{x=1}^m \sum_{y=1}^m \sigma(K_{x,y,R}, K_{x,y,G}, K_{x,y,B}) \quad (1)$$

where CS refers to color-sensitivity, calculated as the sum of standard deviations (σ) computed among the three color channels (R, G, B) for each coefficient of the kernel, and m refers to the width and height of the i^{th} kernel K . The standard deviation between the various channels at particular positions is a measure of the kernel's sensitivity to colors. Low score indicates little presence of color content embedded in the kernel, whereas higher scores mean that greater color content has been embedded in the particular kernel, making them sensitive to particular colors. In a similar manner, their sensitivity to texture can also be measured as:

$$TS_i = \sum_{x=1}^{m-1} \sum_{y=1}^m \sigma(K_{x,y}, K_{x+1,y}) + \sum_{x=1}^m \sum_{y=1}^{m-1} \sigma(K_{x,y}, K_{x,y+1}) \quad (2)$$

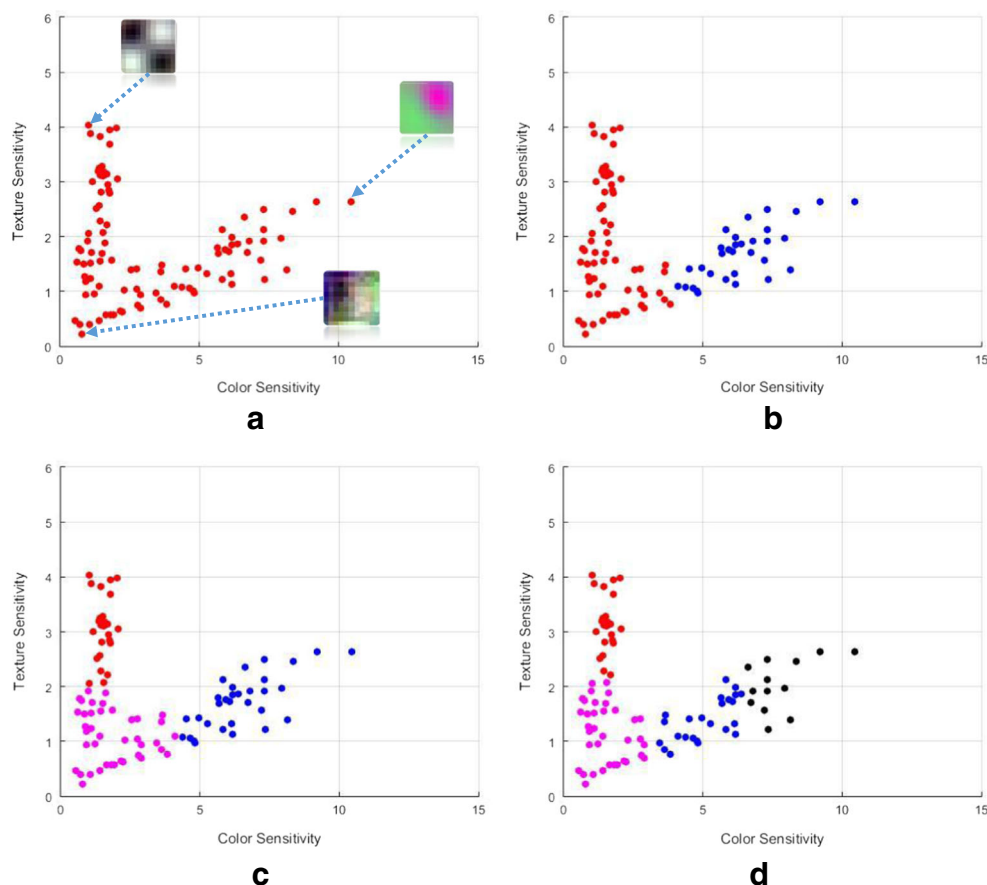
where TS indicates texture-sensitivity score computed as sum of standard deviations (σ) calculated between adjacent kernel coefficients in vertical and horizontal directions at all color channels. Procedure defined in (2) is applied on all color channels of the kernel. High standard deviation between

neighboring coefficients in the kernel at individual channels reflect their sensitivity to texture, as it refers to high textural content embedding. Though it is difficult to measure by visual inspection, the proposed scheme effectively represents them in the two-dimensional feature space defined over color and texture sensitivity.

Clustering kernels

Based on the color and texture sensitivity characteristics of various convolutional kernels, they can be represented in a two dimensional (2D) feature space. It will allow us to visualize them and then study their characteristics in a more effective manner. Fig. 1 represents the kernels in the 2D feature space with varying number of clusters. Fig. 1a shows the distribution of various kernels based on their CS and TS scores. It can be seen that some kernels have high color sensitivity, whereas, others have high texture sensitivity and very low color sensitivity. Using this information, we have clustered them into 2, 3, and 4 clusters as shown in Fig. 1b–d. By grouping them into two clusters, we get two sets of kernels. One set of kernels have very high texture sensitivity (indicated by red dots) and the other set have high color sensitivity as indicated in blue color in Fig. 1b. However, it can be noticed that the red group contains several kernels which have low texture sensitivity and high color sensitivity. Hence, two clusters do not effectively group them into meaningful sets. Similar is the case with 4 clusters as shown in Fig. 1d where the clusters look mixed up. For three clusters, we get three well defined sets of kernels, one of which include kernels having high texture sensitivity (red). The other group of kernels (magenta) consist of kernels with both low color and texture sensitivities. The third cluster (blue) consist of kernels having high color sensitivity and relatively less texture sensitivity. The three sets of kernels have been presented in Fig. 2. It is interesting to note the similarity in their characteristics in each individual cluster. For instance, the first set consist of 24 kernels having high TS scores with no colors. They resemble the Gabor filters often used for texture representation [34]. The second set consist of 41 kernels with low TS and CS scores and can be observed the presence of both color and texture content. Similarly, the last set consist of 31 kernels, which have high CS scores and the smooth rich color content can be clearly seen. Applying these kernels at once and then recording the maximum activations may not effectively capture salient features. Convoluting with individual sets of kernels will allow us to capture more fine-grained features for effective image representation. The clustered feature maps are then analyzed separately for distinctive features.

Fig. 1 Clusters in convolutional feature space, (a) convolutional feature space corresponding to color and texture sensitivity, (b) two clusters, (c) three clusters, (d) four clusters



Spatial maximal activator pooling

Feature maps effectively model visual contents in CNNs which are further analyzed by higher layers for global/contextual features. In the current scenario, we aim to utilize the feature maps for deriving a compact, discriminative representation. Each input image is resized to 64×64 spatial dimensions, which is then convolved with stride 1 and replicate padding to avoid reducing dimensions. We separately convolve the input image with each set of kernels to obtain separate sets of feature maps as shown in Fig. 3. Each set of feature map contain a lot of useful and

possibly redundant information about the input image. However, if we use all the values in the feature maps, we will get a very high dimensional representation. In order to effectively use the information in these feature maps, we used a spatial maximal activator pooling strategy where we constructed a single feature map known as spatial maximal activator (SMA) map from each set of feature maps. Instead of the activation value in the feature maps, we collected the information regarding the kernel generating maximum activation values at each pixel position across all feature maps as shown in Fig. 4. For instance, if we convolve the input with four kernels, we will

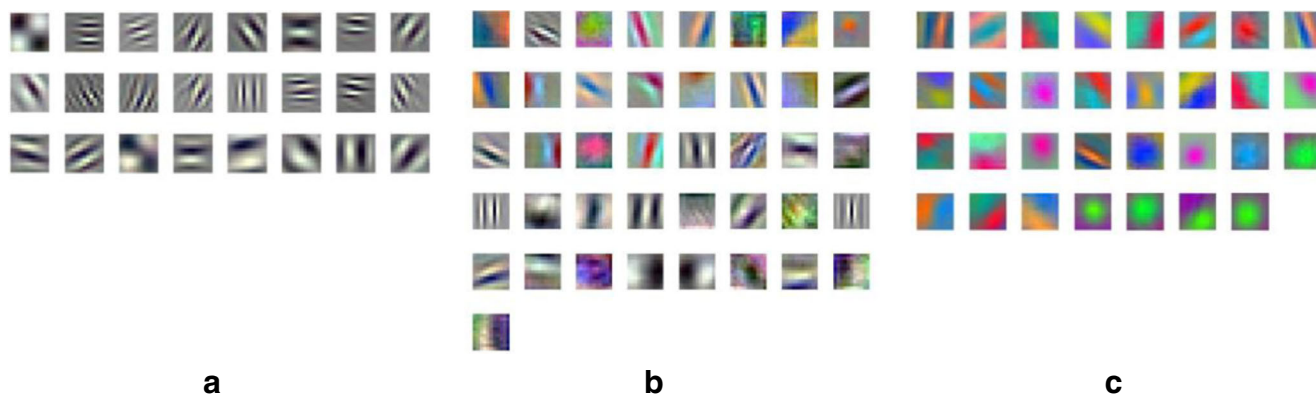


Fig. 2 Clustered Kernels (a) Cluster-1: High texture sensitivity (b) Cluster-2: Low color & texture sensitivity, (c) Cluster-3: High color sensitivity

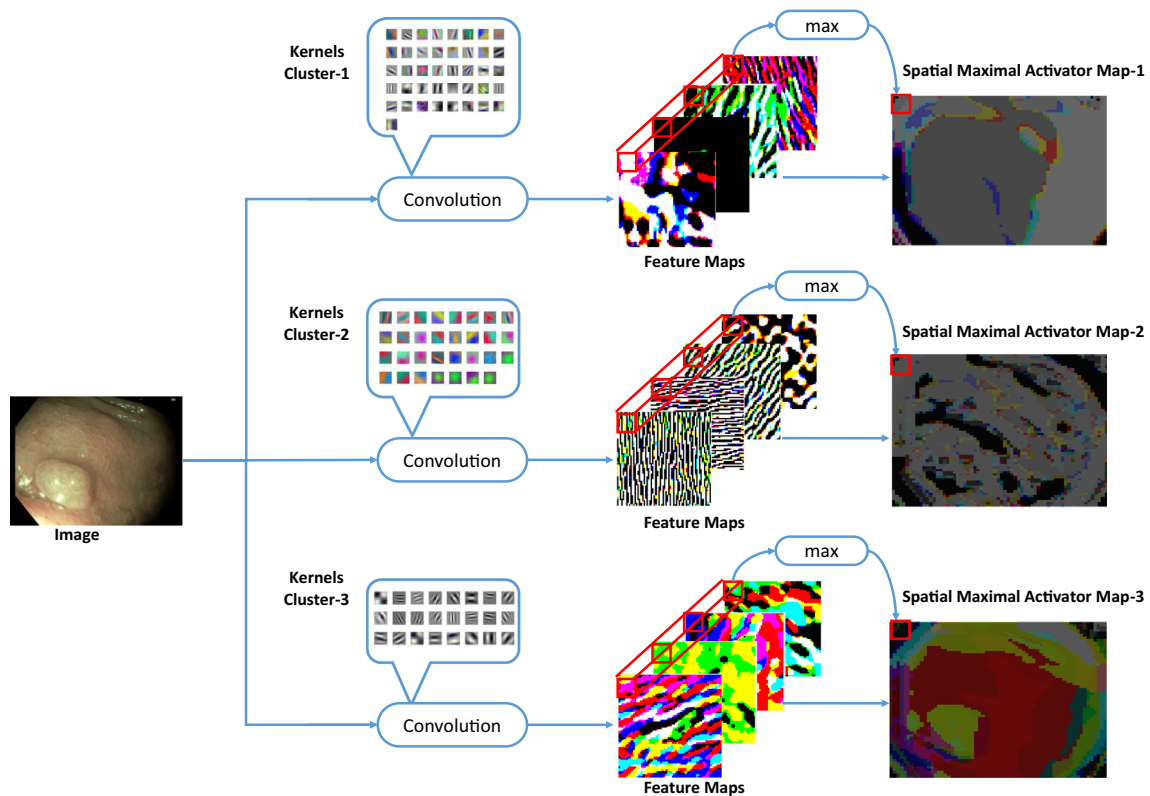


Fig. 3 Clustered convolutional feature space

get four feature maps. The maximal activator value for the red color channel is determined by identifying the kernel which produced maximum value at the first pixel position across red color channel in all feature maps. The same scheme is repeated for all positions and color channels in the SMA map. It conveniently selects the most prominent features while preserving their spatial layout information which can be analyzed further.

Structured pooling of features

The SMA map contain important information about the input image. However, it still is very high dimensional and

needs to be reduced. Further, we also need to capture the spatial layout of these features in order to form a discriminative global representation. Typical spatial pooling approaches collect localized features from non-overlapping or partially overlapping regions and then combine them to form a high dimensional feature vector, or they simply use max pooling or average pooling approaches. Contrary to these, we propose a structured pooling approach similar to the one described in color structure descriptor (CSD) [6] which captures layout information while keeping the feature dimension unchanged. However, instead of colors, we pool information regarding kernels from the SMA

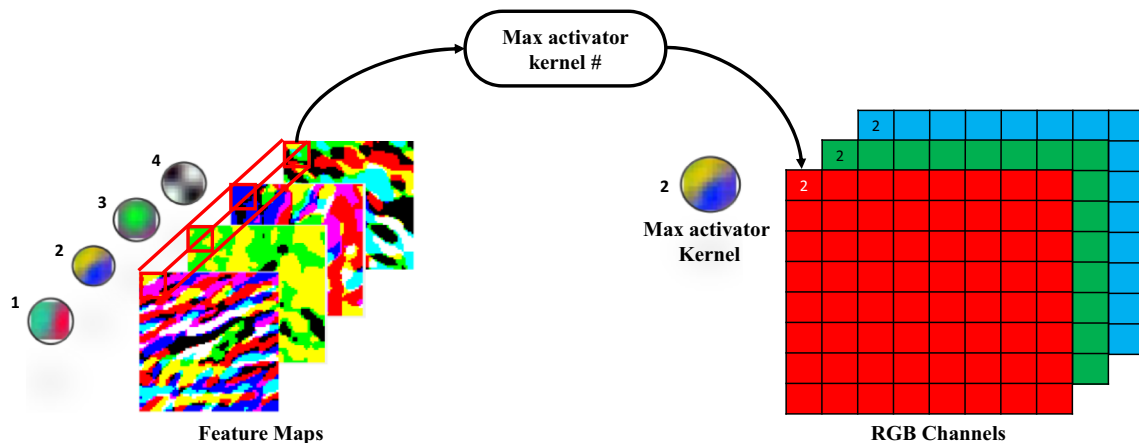
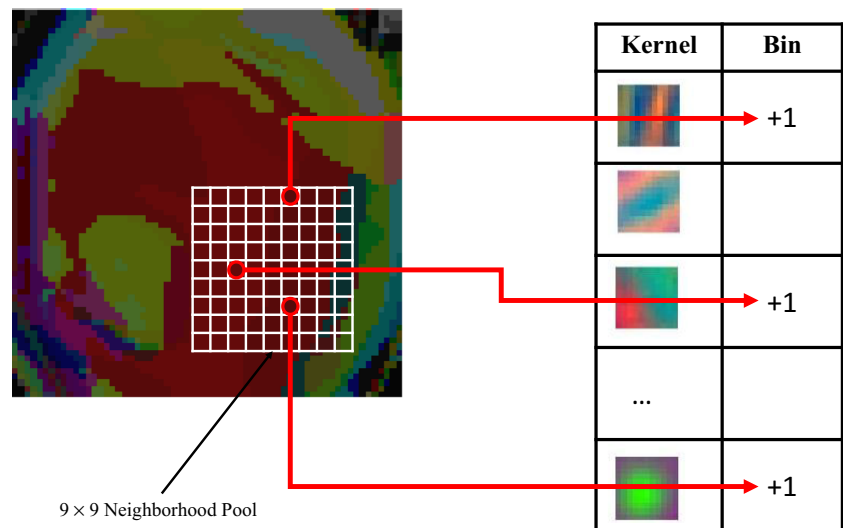


Fig. 4 Spatial Maximal Activator Pooling approach

Fig. 5 Structured Pooling of Convolutional Features



map into a feature histogram as shown in Fig. 5. A 9×9 window is convolved over the entire SMA map in order to capture layout structure characteristics of features. The various histogram bins corresponding to particular kernels which appear within this window are incremented. This process is repeated for every pixel position in all color channels. Kernel co-occurrences are effectively measured in this scheme as the kernels which appear together most of the time will get incremented a lot. In this way, a compact feature histogram is populated for each SMA map. The number of bins in this histogram is equal to the number of kernels in that particular cluster. This structured pooling offers several advantages over other similar approaches including, capturing of global shape information and feature co-occurrences, without increasing feature dimensionality.

Compact image representation

Spatial pooling of all the SMA maps will yield individual feature histograms containing features of colors and textures. These histograms are concatenated to obtain a final feature histogram which is used to represent images in the proposed CBIR system. The number of bins in the final histogram is equal to the number of kernels in the first convolutional layer of AlexNet model, i.e. 96. This 96 bin histogram contains information regarding the color and texture contents of images as well as their spatial relationship. The derivation of final feature histogram is depicted in Fig. 6. For pair-wise image matching, it is important to have discriminative capability in the extracted features. The proposed features contain sufficient discriminative ability which will eventually yield better performance at image retrieval.

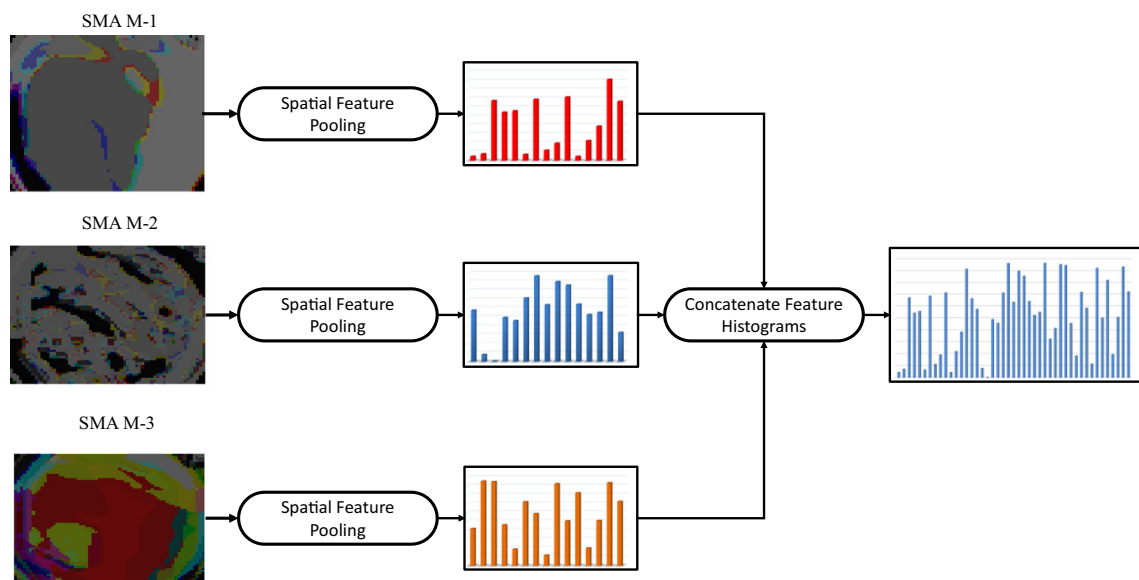


Fig. 6 Compact image representation with proposed features

Table 1 Classification performance comparison on Kvasir dataset

Method	Precision	Recall	F-measure	AUC
2 GF + Random Forest [35]	0.713	0.715	0.711	0.952
6 GF + Random Forest [35]	0.732	0.732	0.727	0.954
3 Layer CNN [35]	0.589	0.408	0.453	0.796
6 Layer CNN [35]	0.661	0.640	0.651	0.942
96 Features + Naïve Bayes	0.64	0.643	0.638	0.940
96 Features + Random Forest	0.673	0.673	0.672	0.948
96 Features + SVM (Proposed)	0.754	0.755	0.753	0.956

Experiments and results

This section presents the various experiments conducted to evaluate the various performance aspects of proposed scheme on two popular datasets. Details of datasets and discussions on results are provided in the following sections.

Datasets

We used an annotated endoscopy images dataset (Kvasir [35]) for evaluating image classification and retrieval performance. It consists of 4000 images, which have been annotated and verified by experienced endoscopists. The images have been grouped into 8 different classes based on anatomical landmarks, pathological findings, or endoscopic procedures. Each class consists of 500 images having different resolutions ranging from 720 × 576 to 1920 × 1072 pixels. Several experiments were designed and executed to assess the classification and retrieval performance based on the proposed features. All the experiments were conducted in MATLAB 2016. Details of the experiments and the results are provided in the subsequent sections.

Endoscopic image classification in kvasir dataset

The extracted features were used to represent endoscopy images to train various classifiers including Naïve Bayes,

Random Forest, and Support Vector Machine (SVM). A 10-fold cross validation strategy was used to perform the classification. During each experiment with individual classifiers, we randomly chose 90% of the data and used it for training the classifier. The remaining was used to test the classifier’s performance. This process was repeated 10 times, each time with a different training and test set. Results of the various classifiers based on the proposed 96 features (indicated as 96 F) and other approaches has been provided in Table 1. Results in the first four rows were taken from the work presented in [35], where they used 2 and 6 global features (GF) with Random Forest classifiers to classify endoscopy images in the Kvasir dataset. The other two methods used a 3 layer and a 6 layer CNN to perform the classification. We used the 96 features with Naïve Bayes and achieved 0.64 precision, 0.643 recall, and 0.638 F-measure, which are slightly better than the 3 layer CNN. Similarly, with Random Forest classifiers, we achieved 0.672 F-measure, which is better than the 6 layer CNN. We achieved the best classification results (0.753 F-measure) with SVM classifier using a linear kernel.

Table 2 lists per-class recognition rates using SVM classifier [36] in the form of a confusion matrix. Labels mentioned at the top of the table correspond to the predicted labels generated by the classifier, whereas the labels at the right of the table represent the ground truth. Each value in the diagonal of the table indicates the percentage accuracy. The other values represent the confusion or incorrect classification. In the category “Polyps”, 71.8% images were correctly classified by the proposed classification approach. However, 14% of these images were incorrectly classified as “ulcerative-colitis”, 8.6% were misclassified as “normal-cecum”, and small number of images were misclassified as other categories. Most of the images in “esophagitis” were correctly classified (68.8%) and a large portion of these images were mislabeled as “normal-z-line”. In the third category “dyed-lifted-polyps”, the classifier achieved 61.4% accuracy. However, 35.8% of these images were misclassified as “dyed-resection-margins” because of the similar appearance. Similar performance and misclassification was

Table 2 Per-category classification performance in Kvasir dataset

Predicted Class								Actual Class
polyps	esophagitis	dyed-lifted-polyps	normal-pylorus	ulcerative-colitis	dyed-resection-margins	normal-cecum	normal-z-line	
71.8	0.4	1	4.2	14	0	8.6	0	polyps
0	68.8	0	2.6	0	0.2	0	28.4	esophagitis
2	0	61.4	0	0.8	35.8	0	0	dyed-lifted-polyps
0.4	1.6	0	95	0.4	0	0	2.6	normal-pylorus
16	0	0.4	1.6	73.6	0.4	7.8	0.2	ulcerative-colitis
1	0	36	0	0	63	0	0	dyed-resection-margins
7	0	0	0	2.6	0	90.4	0	normal-cecum
0	18.6	0	1.6	0.2	0	0	79.6	normal-z-line

noticed in “dyed-resection-margins” category as well. “Normal Pylorus” images were classified with the highest accuracy of 95%. Images from the category “Ulcerative Colitis” were classified with 73.6% accuracy, with some degree of confusion with “Polyps” and “Normal-cecum”. The “normal-cecum” and “normal-z-line” categories were classified with accuracies 90.4% and 79.6%, respectively.

Endoscopic image retrieval in kvasir dataset

During these experiments, query images were randomly chosen from each category and top N images were retrieved using the proposed 96 features. Some visual retrieval results for four

different queries have been shown in Fig. 7 where the upper left most image is the query image and the remaining are the top 25 retrieved images. For each of these queries, our system was able to retrieve visually similar images at top ranks. Each of these queries, belong to distinct image categories, and the retrieval results indicate that the proposed features possess significant representational capability involving color, texture, local shape and layout features. To acquire quantitative performance scores, we executed different queries from each category and obtained retrieval scores in the form of precision for scopes ranging from 10 to 200. Scope represents the number of images retrieved during a particular query. Twenty images were randomly chosen from each category and top 200 images

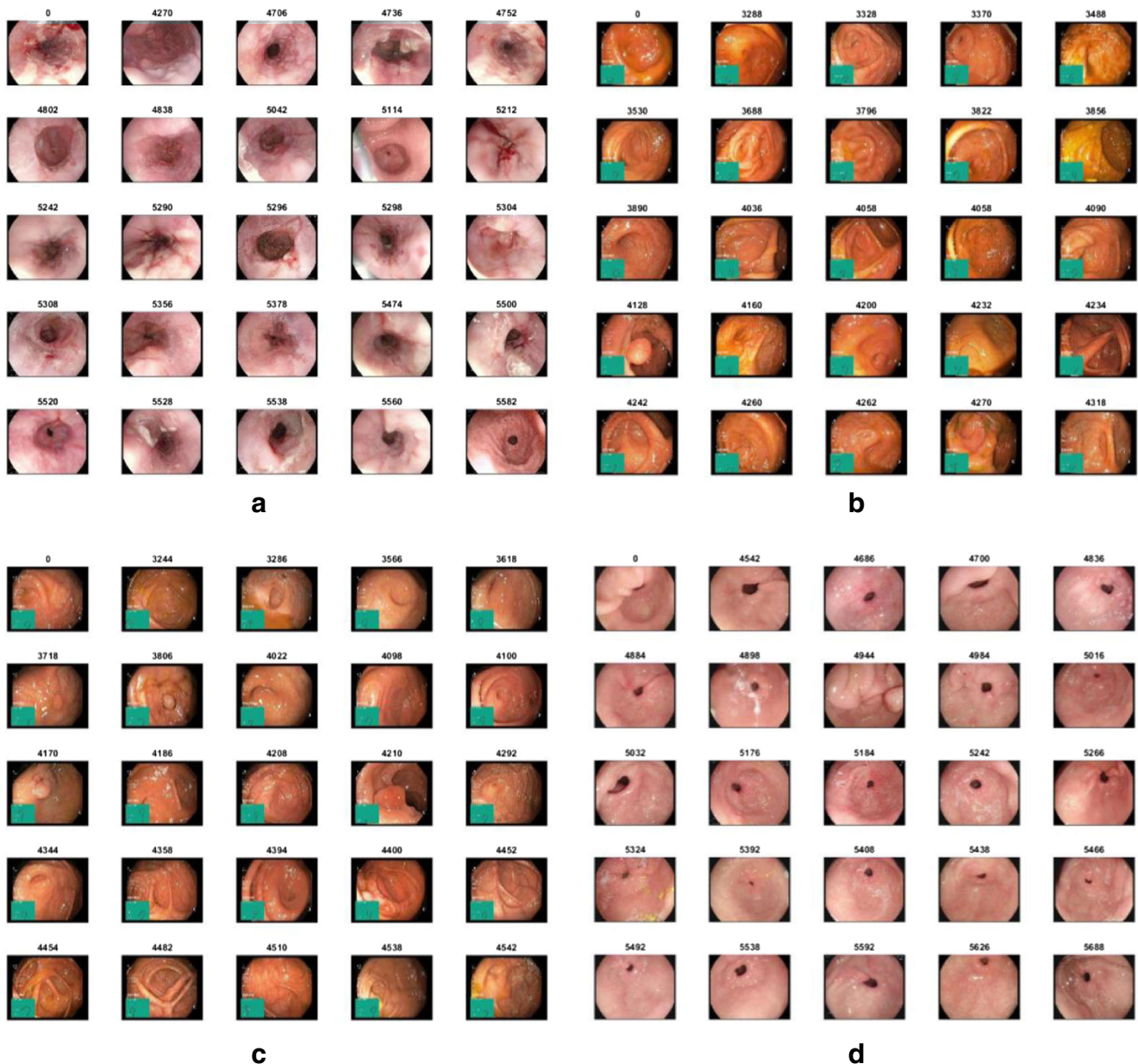


Fig. 7 Retrieval Results for Kvasir dataset

were retrieved. Each time, the precision scores were computed for various scope values. The final scores from each category were obtained after averaging the scores of twenty queries.

Figure 8 show quantitative results for various categories defined in the dataset. Figure 8a reports precision scores for various scope ranges in the three distinct categories defined as anatomical landmarks. Retrieval performance of the “normal-pylorus” is better than the other two categories at scopes below 150. Its performance is significantly higher than the other categories for scope up to 100. Its performance decreases at a regular pace after scope increases from 125. Images in “normal-cecum” category were retrieved with better scores for all scopes than the “normal-z-line” category, and showed relatively stable performance across all scopes. The retrieval performance for “normal-z-line” was relatively poor, especially when scope was increased from 50. Overall, precision scores for all these categories were above 80% for scopes up to 50. Figure 8b shows precision scores for three pathological categories including esophagitis, polyps, and ulcerative colitis. At scope 10, esophagitis and ulcerative colitis had similar

precision scores, however, the score for polyps was significantly lower. When the scope was increased, ulcerative-colitis showed a much drastic decrease in performance as compared to the other two categories, whose performance decreased gracefully. For scope up to 50, the precision scores for all these categories was above 60%. Figure 8c shows retrieval results for two remaining categories, which indicate that both these categories perform poorly when the scope is increased from 10. Although, in general, retrieval systems usually desire for better precision scores at low scope, retrieval performance in these categories need to be improved. Figure 8d shows the overall retrieval performance across all categories. For top-10 retrieval, the proposed features yield more than 82% precision. The performance degrades gracefully with increase in scope. At 200, it achieves around 61% precision.

Effect of number of kernel clusters

In this experiment, we tested the effects of varying the number of clusters on image retrieval performance. Four different

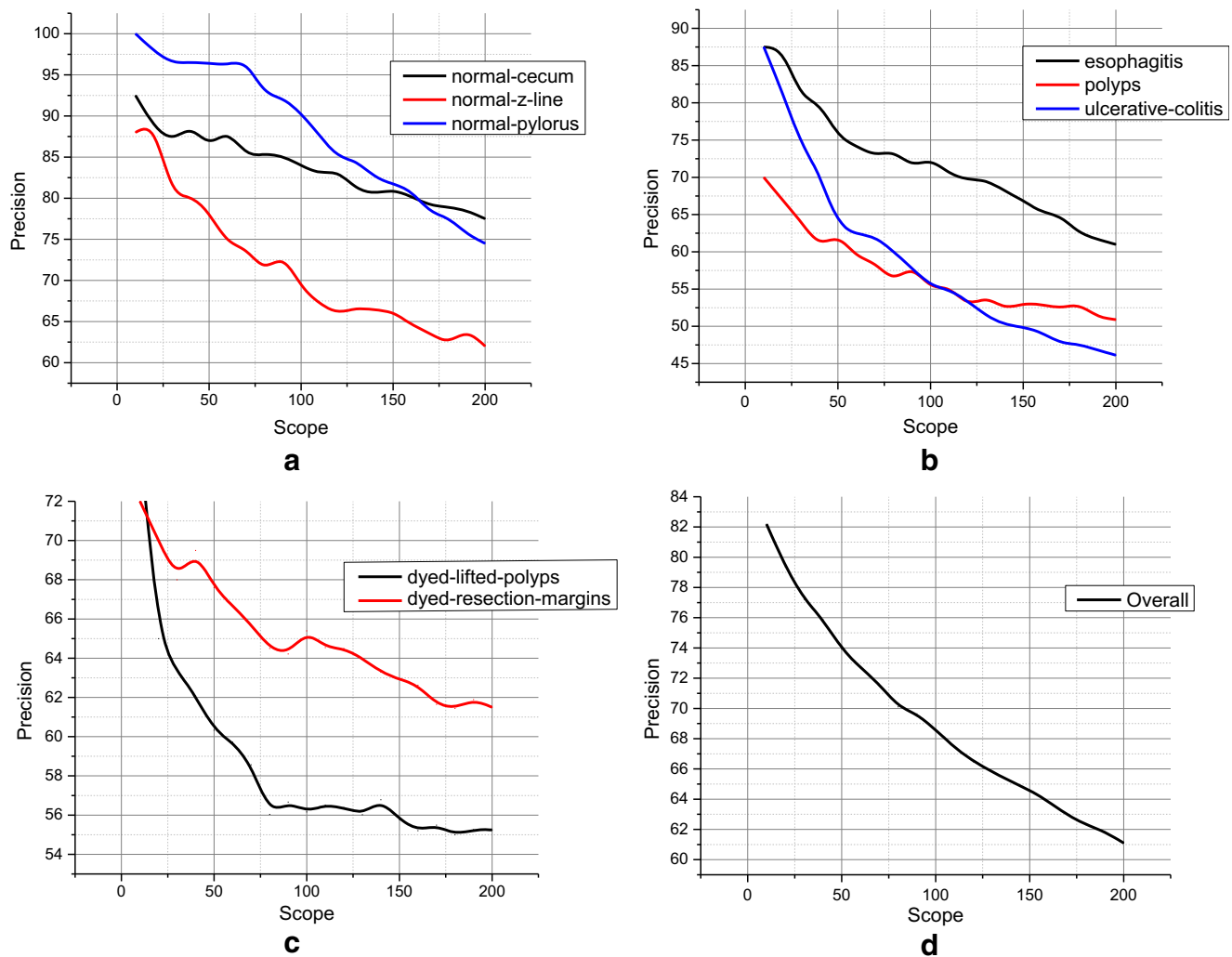


Fig. 8 Per category precision for varying scope in (a) anatomical landmarks, (b) pathological categories, (c) polyp removal, and (d) overall

experiments were conducted with different number of clusters. During the first experiment, we tested retrieval without any clustering of kernels. At this setting, the proposed method treats both texture and color features together and selects only prominent features irrespective of the fact that a point in image may contain both texture and color features of significant importance. Therefore, the retrieval performance with this setting is quite low, as can be seen in Fig. 9. By increasing the number of clusters to two, the retrieval performance got significantly improved for all scopes. This improvement is particularly attributed to the fact that color and texture features are now treated separately, and the texture content in images is effectively represented with this setting. The performance got further improved with three clusters because this setting provides a much better grouping of kernels as discussed in “Clustering Kernels” Section. Retrieval scores improved significantly at scope 10 and scope beyond 50. By further increasing the number of clusters, slight performance improvement was noticed. However, the overall performance with 3 clusters was the optimum. These results reveal the importance in clustering in the convolutional feature space. The first layer kernels detect basic color and texture features of generic nature. A particular kernel may be sensitive to colors or textures or both. Understanding their sensitivity to colors and textures would allow us to represent the features in more effective ways. With three clusters, we were able to extract dominant color and texture features with those kernels which have high sensitivities. The third cluster consisted of those kernels which are slightly sensitive to both. This separation also allows us to assign weights to features, depending on the requirements of the dataset. Further, we can also ignore certain kernels when particular features are absent in a dataset. For instance, we can ignore color sensitive kernels when dealing with grayscale images. Further analysis along these lines can reveal the effectiveness of proposed approach which will be performed in future.

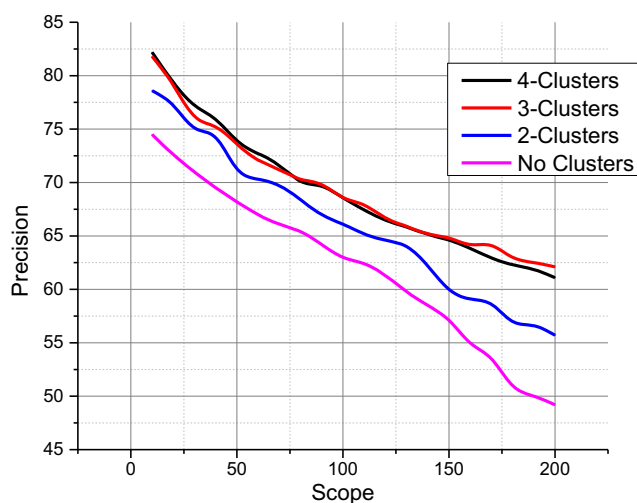


Fig. 9 Performance with varying number of kernel clusters

Effect of spatial pooling

The SMA maps contain both color and texture features at specific locations where they exist in the input image. An effective pooling strategy can accumulate these features into compact feature histograms. For this purpose, we experimented with different strategies to populate histograms, including max, average, and structured pooling. In max pooling, the maximum values across feature maps were gathered to form feature vector. Average pooling used average values of feature maps instead of maximum to populate the histogram. The proposed structured pooling used the window based approach to collect features from SMA map. One experiment was carried out to assess retrieval performance using a simple approach where features were pooled into the respective bins without regard to their spatial location from the SMA map. With this setting, the features lost some of its discriminative strength and achieved high degree of invariance. Feature histograms were invariant to geometric transformations, however, the retrieval score was relatively low. In the second experiment, we used the structured pooling strategy to form the global image representation. With this strategy the precision scores improved significantly as shown in Fig. 10. Similarly, feature map-wise max and average pooling schemes were separately tested which yielded lower precision scores than the SMAP based approaches, particularly with higher scopes. Average pooling offered better precision than max pooling scheme.

Retrieval performance comparison with existing feature extraction methods

Several existing features extraction methods were used to retrieve endoscopy images in Kvasir dataset. Randomly chosen query images were used to retrieve top 50 images for each

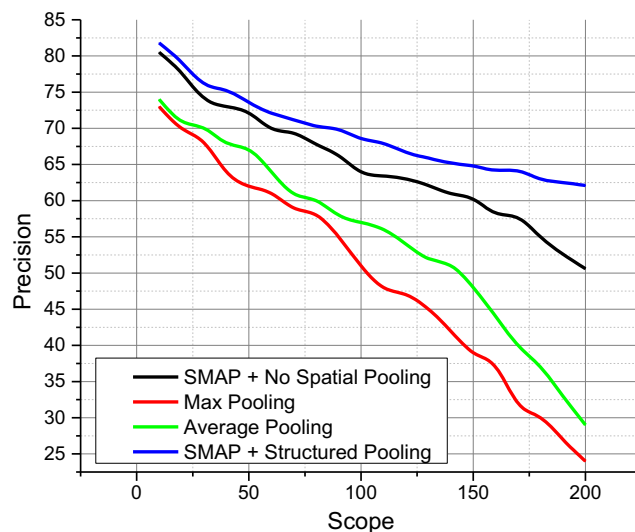


Fig. 10 Performance comparison with different pooling approaches

Table 3 Retrieval performance in Kvasir dataset

Method	Precision@50
MSD [9]	63.85
DSD [37]	64.11
MTH [8]	65.05
CDH [7]	68.48
SED [10]	71.64
MS-LSP [11]	72.16
Proposed	74.02

feature set. Precision scores computed at scope 50 have been reported in Table 3. Features extraction methods including SED and MS-LSP take into account local color and texture features, their scores were close to the proposed method. Other approaches like CDH, MTH, DSD, and MSD also take into account local color and texture features, however they exhibited less discriminatory capabilities for endoscopy images. The proposed method achieved 74.02 precision at scope 50, which shows the effectiveness of the method in representing endoscopy images.

Conclusion and future work

In this paper, we studied the convolutional kernels from the first convolutional layer of AlexNet model. We showed that the kernels can effectively model visual contents by capturing color and texture features. The characteristics of these kernels were studied using two measures, color-sensitivity and texture-sensitivity. These two measures were used to analyze the convolutional feature space created by these kernels in the deep CNN. We found that the kernels vary significantly in their sensitivities to colors and textures. Some of the kernels had significantly higher color-sensitivity whereas others have higher sensitivities to textures. Based on these observations, we clustered the convolutional feature space into three distinct clusters. Each of these clusters contain a number of kernels having similar characteristics. These individual sets of kernels were used to extract color and texture features separately from the input image and then aggregate those features into a single feature map called spatial maximal activator map. The features in these maps were collected into a histogram in such a way that their spatial layout information is also captured without increasing the dimensions of the features, using a structured pooling approach. Experiments revealed that the proposed method provides superior retrieval performance on endoscopy images dataset.

We also conducted experiments with other pooling approaches including map-wise max pooling and average pooling. However, their retrieval performance was significantly lower than the proposed SMAP based pooling. Though the first layer kernels are known to extract basic features,

combining them into a global representation is a key factor in transforming them into useful representations. The proposed clustered convolutional feature space along with spatial maximal activator pooling approach allowed us to capture salient color and texture features separately. Further, the structured pooling approach enabled us to construct a discriminative global representation. We strongly believe, that the proposed method can be used in representing medical images with high color and texture content, effectively in CBIR systems. The framework can be further enhanced with more complicated pooling approaches and investigating other characteristics to analyze the kernels. In future, we plan to extend this framework to other deeper layers of the CNNs and attempt to model visual contents in compact feature histograms.

Compliance with Ethical Standards

Funding This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea Government (MSIP) (No.2016R1A2B4011712).

Conflict of Interest The authors declare that there is no conflict of interest.

Ethical Approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

- Sainju, S., Bui, F.M., and Wahid, K.A., Automated bleeding detection in capsule endoscopy videos using statistical features and region growing. *J. Med. Syst.* 38:25, 2014.
- Ahmad, J., Sajjad, M., Mehmood, I., Rho, S., and Baik, S.W., Saliency-weighted graphs for efficient visual content description and their applications in real-time image retrieval systems. *J. Real-Time Image Proc.* 1–17, 2016.
- Murala, S., Maheshwari, R., and Balasubramanian, R., Directional binary wavelet patterns for biomedical image indexing and retrieval. *J. Med. Syst.* 36:2865–2879, 2012.
- Smeulders, A.W., Worring, M., Santini, S., Gupta, A., and Jain, R., Content-based image retrieval at the end of the early years. *Pattern Analysis and Machine Intelligence, IEEE Transactions on.* 22: 1349–1380, 2000.
- Nowaková, J., Prilepok, M., and Snášel, V., Medical image retrieval using vector quantization and fuzzy S-tree. *J. Med. Syst.* 41:18, 2017.
- Messing, D. S., Van Beek, P., and Errico, J. H., The mpeg-7 colour structure descriptor: Image description using colour and local spatial information. In: *IEEE International Conference on Image Processing (ICIP), Thessaloniki, Greece*, pp. 670–673, 2001. <http://dx.doi.org/10.1109/ICIP.2001.959134>.
- Liu, G.-H., and Yang, J.-Y., Content-based image retrieval using color difference histogram. *Pattern Recogn.* 46:188–198, 2013.
- Liu, G.-H., Zhang, L., Hou, Y.-K., Li, Z.-Y., and Yang, J.-Y., Image retrieval based on multi-texton histogram. *Pattern Recogn.* 43: 2380–2389, 2010.
- Liu, G.-H., Li, Z.-Y., Zhang, L., and Xu, Y., Image retrieval based on micro-structure descriptor. *Pattern Recogn.* 44:2123–2133, 2011.

10. Wang, X., and Wang, Z., A novel method for image retrieval based on structure elements' descriptor. *J. Vis. Commun. Image Represent.* 24:63–74, 2013.
11. Ahmad, J., Sajjad, M., Rho, S., and Baik, S.W., Multi-scale local structure patterns histogram for describing visual contents in social image retrieval systems. *Multimed. Tools Appl.* 75:12669–12692, 2016.
12. Lowe, D.G., Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* 60:91–110, 2004.
13. Bay, H., Ess, A., Tuytelaars, T., and Van Gool, L., Speeded-up robust features (SURF). *Comput. Vis. Image Underst.* 110:346–359, 2008. <http://dx.doi.org/10.1016/j.cviu.2007.09.014>.
14. Li, T., Mei, T., Kweon, I.-S., and Hua, X.-S., Contextual bag-of-words for visual categorization. *IEEE Trans. Circ. Syst. Video Technol.* 21:381–392, 2011.
15. Haas, S., Donner, R., Burner, A., Holzer, M., and Langs, G., Superpixel-based interest points for effective bags of visual words medical image retrieval. In: *MICCAI International Workshop on Medical Content-Based Retrieval for Clinical Decision Support*, pp. 58–68. Berlin, Heidelberg: Springer, 2011.
16. Yang, J., Jiang, Y.-G., Hauptmann, A. G., and Ngo, C.-W., Evaluating bag-of-visual-words representations in scene classification. In: *Proceedings of the international workshop on multimedia information retrieval, Augsburg, Bavaria, Germany*, pp. 197–206, 2007.
17. Wang, S., Lu, S., Dong, Z., Yang, J., Yang, M., and Zhang, Y., Dual-tree complex wavelet transform and twin support vector machine for pathological brain detection. *Appl. Sci.* 6:169, 2016.
18. Zhang, Y.-D., Zhao, G., Sun, J., Wu, X., Wang, Z.-H., Liu, H.-M., et al., Smart pathological brain detection by synthetic minority oversampling technique, extreme learning machine, and Jaya algorithm. *Multimed. Tools Appl.* 1–20, 2017. <http://dx.doi.org/10.1007/s11042-017-5023-0>.
19. Wang, P., Krishnan, S. M., Kugean, C., and Tjoa, M., Classification of endoscopic images based on texture and neural network. In: *Engineering in Medicine and Biology Society, 2001. Proceedings of the 23rd Annual International Conference of the IEEE, Istanbul, Turkey*, pp. 3691–3695, 2001.
20. Wang, S.-H., Du, S., Zhang, Y., Phillips, P., Wu, L.-N., Chen, X.-Q., et al., Alzheimer's disease detection by Pseudo Zernike moment and linear regression classification. *CNS Neurol. Disord. Drug Targets (Formerly Current Drug Targets-CNS & Neurological Disorders)*. 16:11–15, 2017.
21. Krizhevsky, A., Sutskever, I., and Hinton, G. E., Imagenet classification with deep convolutional neural networks. In: *Advances in neural information processing systems, Lake Tahoe, Nevada*, pp. 1097–1105. Curran Associates, Inc., USA, 2012.
22. Ahmad, J., Sajjad, M., Mehmood, I., and Baik, S.W., SiNC: Saliency-injected neural codes for representation and efficient retrieval of medical radiographs. *PLoS One.* 12:e0181707, 2017.
23. Krizhevsky, A., and Hinton, G. E., Using very deep autoencoders for content-based image retrieval. In: *Proceedings of the 19th European Symposium on Artificial Neural Networks, Bruges, Belgium*, pp. 489–494, 2011.
24. Zhang, Y.-D., Zhang, Y., Hou, X.-X., Chen, H., and Wang, S.-H., Seven-layer deep neural network based on sparse autoencoder for voxelwise detection of cerebral microbleed. *Multimed. Tools Appl.* 1–18, 2017. <http://dx.doi.org/10.1007/s11042-017-4554-8>.
25. Qi, Y., Song, Y.-Z., Zhang, H., and Liu, J., Sketch-based image retrieval via Siamese convolutional neural network. In: *IEEE International Conference on Image Processing (ICIP), Phoenix, AZ, USA*, pp. 2460–2464, 2016.
26. Vishnuvarthanan, A., Rajasekaran, M.P., Govindaraj, V., Zhang, Y., and Thiagarajan, A., An automated hybrid approach using clustering and nature inspired optimization technique for improved tumor and tissue segmentation in magnetic resonance brain images. *Appl. Soft Comput.* 57:399–426, 2017.
27. Lu, S., Wang, S., and Zhang, Y., A note on the marker-based watershed method for X-ray image segmentation. *Comput. Methods Prog. Biomed.* 141:1–2, 2017.
28. Pons, J., and Serra, X., Designing efficient architectures for modeling temporal features with convolutional neural networks. In: *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2017), New Orleans, USA*, pp. 2472–2476, 2017.
29. Zeiler, M. D., and Fergus, R., Visualizing and understanding convolutional networks. In: Fleet, D., Pajdla, T., Schiele, B., and Tuytelaars, T. (Eds), *Computer Vision – ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part I*, pp. 818–833. Cham: Springer International Publishing, 2014. http://dx.doi.org/10.1007/978-3-319-10590-1_53.
30. Babenko, A., Slesarev, A., Chigorin, A., and Lempitsky, V., Neural codes for image retrieval. In: Fleet, D., Pajdla, T., Schiele, B., and Tuytelaars, T. (Eds.), *Computer Vision – ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part I*, pp. 584–599. Cham: Springer International Publishing, 2014. http://dx.doi.org/10.1007/978-3-319-10590-1_38.
31. Razavian, A. S., Azizpour, H., Sullivan, J., and Carlsson, S., CNN features off-the-shelf: an astounding baseline for recognition. In: *IEEE Conference on Computer Vision and Pattern Recognition Workshops, 23–28 June, Columbus, OH, USA*, pp. 512–519, 2014. <http://dx.doi.org/10.1109/CVPRW.2014.131>.
32. Ahmad, J., Mehmood, I., Rho, S., Chilamkurti, N., and Baik, S.W., Embedded deep vision in smart cameras for multi-view objects representation and retrieval. *Comput. Electr. Eng.* 61C:297–311, 2017.
33. Ahmad, J., Mehmood, I., and Baik, S.W., Efficient object-based surveillance image search using spatial pooling of convolutional features. *J. Vis. Commun. Image Represent.* 45:62–76, 2017.
34. Li, C., Huang, Y., and Zhu, L., Color texture image retrieval based on Gaussian copula models of Gabor wavelets. *Pattern Recogn.* 64: 118–129, 2017.
35. Pogorelov, K., Randel, K. R., Griwodz, C., Eskeland, S. L., de Lange, T., Johansen, D., et al., Kvasir: a multi-class image dataset for computer aided gastrointestinal disease detection. In: *Proceedings of the 8th ACM on Multimedia Systems Conference, Taipei, Taiwan*, pp. 164–169, 2017.
36. Wang, S., Chen, M., Li, Y., Shao, Y., Zhang, Y., Du, S., et al., Morphological analysis of dendrites and spines by hybridization of ridge detection with twin support vector machine. *PeerJ.* 4: e2207, 2016.
37. Yu, L., Feng, L., Chen, C., Qiu, T., Li, L., and Wu, J., A Novel Multi-Feature Representation of Images for Heterogeneous IoTs. *IEEE Access.* 4:6204–6215, 2016.