Classification of Tumor Epithelium and Stroma in Colorectal Cancer based on Discrete Tchebichef Moments

Rodrigo Nava[1(✉)], Germán González[2], Jan Kybic[1], and Boris Escalante-Ramírez[2]

[1] Czech Technical University in Prague, Czech Republic
uriel.nava@gmail.com
[2] Facultad de Ingeniería, Universidad Nacional Autónoma de México, Mexico

Abstract. Colorectal cancer is a major cause of mortality. As the disease progresses, adenomas and their surrounding tissue are modified. Therefore, a large number of samples from the epithelial cell layer and stroma must be collected and analyzed manually to estimate the potential evolution and stage of the disease. In this study, we propose a novel method for automatic classification of tumor epithelium and stroma in digitized tissue microarrays. To this end, we use discrete Tchebichef moments (DTMs) to characterize tumors based on their textural information. DTMs are able to capture image features in a non-redundant way providing a unique description. A support vector machine was trained to classify a dataset composed of 1376 tissue microarrays from 643 patients with colorectal cancer. The proposal achieved 97.62% of sensitivity and 95% of specificity showing the effectiveness of the methodology.

Keywords: Colorectal cancer, Tchebichef moments, Tissue microarray, Tumor classification, Support vector machine

1 Introduction

Colorectal cancer (CRC) is the third most common type of cancer worldwide with more than 1.4 million cases registered in 2012 [4]. As population aging continues growing more people are susceptible to CRC: around 70% of cancer mortality occurs among adults over 65 years [7]. Furthermore, almost half of the population will develop at least one benign intestinal tumor during its lifetime [10]. In most cases, CRC begins as a benign polyp or adenoma, which is characterized by accumulation of cells at the epithelial layer of the gastrointestinal track. A small fraction of polyps evolves through accumulation of genetic alterations yielding carcinomas. Such a sequence is called adenoma-carcinoma sequence (ACS) [17].

Cancer progression through lymphatic or blood vessels (metastasis) to the liver and lungs is the principal cause of death and occurs in up to 25% of patients [2]. In contrast to ACS, colorectal metastasis is not strongly associated with alterations in any genes but with the healthy cells that surround the tumors. Such cells, called stroma, are usually composed of connective tissue. They are essential for the maintenance of both normal epithelial tissue and their malignant counterpart. Oncogenic changes in

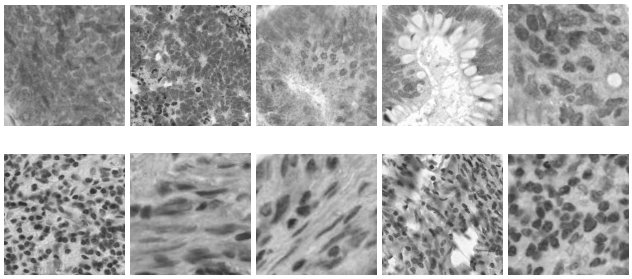


Fig. 1. Samples of colorectal cancer in digitized tissue microarrays (only red channel) from the database used in [12]. First row shows pure tumor epithelium and second row shows tumor stroma extracted from a paraffin block.

the epithelial tissue modify the stromal host compartment, which is responsible for establishing and enabling a supportive environment and eventually promotes growth and metastasis. Hence, stroma plays a fundamental role in allowing development and progression of the disease [1], [2], [8].

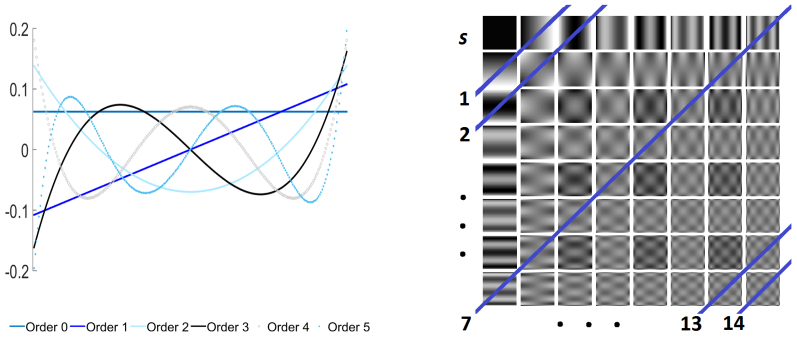
Tissue microarrays (TMAs) are the gold standard for determining and monitoring the prevalence of alterations associated with colorectal carcinogenesis [19]. This procedure collects small histological sections from unique tissues or tumors and places them in an array to form a single paraffin block, (see Fig. 1). Typical TMAs may contain up to 1000 spots that are used for simultaneous interpretation. Hence, the large amount of information is the main drawback of the manual assessment and the motivation of this study. In addition, the identification of regions of interest depends on visual evaluation of histology slide images by pathologists, which introduces a bias.

Texture analysis has been used in segmentation of epithelial tissue in digital histology previously. For instance, Wang [20] proposed a Bayesian estimation method for classification of tumoral cells in tissue microarrays of lung carcinoma. Tumor and stroma from prostate tissue microarrays were classified in [9,11,3]. Foran et al. [6] developed a software platform to compare expression patterns in tissue microarrays using texton-based descriptors and intensity histograms. To the best of our knowledge, automated analysis of CRC in tissue microarrays is relatively new. Linder et al. [12] used a methodology based on local binary patterns (LBPs) [18] and contrast information called (LBP/C) to classify tumor epithelium and stroma. Here, we use the same dataset and propose a novel descriptor based on discrete Tchebichef polynomials.

Next, we present a detailed description of the methodology. A comparison between our proposal and LBPs was also performed using k-NN and support vector machine (SVM) as classifiers.

2 Material and methods

We propose a methodology composed of three stages. First, for each image, feature extraction is performed on overlapped sliding windows using discrete Tchebichef polynomials. Then, all the local Tchebichef vectors from a single image are grouped and



(a) (b)

Fig. 2. Set of scaled Tchebichef kernels. **(a)** 1-D discrete Tchebichef polynomials of order s from 0 to 5. **(b)** Ensemble of 2-D discrete Tchebichef polynomials. The magnitude of the moment of order s is calculated by summing of the correlation indexes, p,q so that $s = p+q$. Graphically, the sum is carried out diagonally.

characterized by statistic moments in order to build a single vector of 234-bins length that can be viewed as the texture signature. Finally, a SVM is trained using a subset of 656 samples, whereas the performance of the proposal is assessed on an independent set of 720 tissue microarray samples.

2.1 Dataset

We used the dataset provided and described in detail in [12], which consists of 1376 samples of tissue microarray of tumor epithelium and stroma from 643 patients with CRC annotated by expert pathologists, (see Fig. 1). The samples were divided into two parts. The training subset is composed of 656 images: 400 samples representing tumor epithelium and 256 representing tumor stroma. A separate subset, consists of 425 images of tumor epithelium and 295 images that represent tumor stroma, was used as validation set. The dataset does not contain private information of patients.

Prior to extract Tchebichef feature vectors, the tissue samples were scaled by a 0.5 factor, the mean was subtracted, and only the red channel was used. Blue and green channels were discarded because they do not have relevant information.

2.2 Discrete Tchebichef Moments

Generally speaking, moments are scalar quantities that characterize a function of interest. They are computed as projections between the function $f(x,y)$ and a polynomial basis $r_{pq}(x,y)$ within the region $\Omega : T_{pq} = \iint_{\Omega} r_{pq}(x,y) f(x,y)\, dxdy$ where p and q are non-negative integers and $s = p+q$ represents the order of the moment. Therefore, T_{pq} measures the correlation between the function $f(x,y)$ and the corresponding polynomial $r_{pq}(x,y)$ [5].

Discrete Tchebichef moments (DTMs) were originally proposed by Mukundan et al. [15] to overcome limitations of conventional orthogonal moments such as Zernike and Legrendre. DTMs are based on a normalized version of discrete Tchebichef polynomials scaled by a factor that depends on the size of the image N, (see Fig. 2(a)).

The scaled discrete Tchebichef polynomials, $\widehat{t}_p$, can be generated using the following recurrent relation:

$$\begin{aligned}
\widehat{t}_0(x) &= \frac{1}{\sqrt{N}}, \\
\widehat{t}_1(x) &= (2x+1-N)\sqrt{\frac{3}{N(N^2-1)}}, \text{ and} \\
\widehat{t}_p(x) &= K_1 x\widehat{t}_{p-1}(x) + K_2\widehat{t}_{p-1}(x) + K_3\widehat{t}_{p-2}(x)
\end{aligned} \tag{1}$$

with $x = 0, 1, \ldots, N-1$.

$K_1 = \frac{2}{p}\sqrt{\frac{4p^2-1}{N^2-p^2}}$, $K_2 = \frac{1-N}{p}\sqrt{\frac{4p^2-1}{N^2-p^2}}$, and $K_3 = \frac{p-1}{p}\sqrt{\frac{2p+1}{2p-3}}\sqrt{\frac{N^2-(p-1)^2}{N^2-p^2}}$ are the coefficients that ensure stability in case of large order polynomials [14].

DTMs are computed by projecting a given image, $I(x,y)$, onto the basis of $\widehat{t}_p$. The moment T_{pq} is calculated according the following formula:

$$T_{pq} = \sum_{x=0}^{N-1}\sum_{y=0}^{N-1}\widehat{t}_p(x)\widehat{t}_q(y)I(x,y) \tag{2}$$

T_{pq} quantifies the correlation between the image, $I(x,y)$, and the kernel $\widehat{t}_p(x)\widehat{t}_q(y)$, see Fig. 2(b).

One way to understand this relationship is that the greatest the magnitude of T_{pq}, the greatest the similarity between the given image and the polynomials $\widehat{t}_p$ that oscillate at similar rates to the image. Hence, it is possible to build a feature vector, $M(s)$, that captures similarities along X- and Y-axes as follows:

$$M(s) = \sum_{p+q=s}\left|T_{pq}\right| \tag{3}$$

with $s = 0, 1, \ldots, 2N-2$.

$M(s)$ provides a unique description in the expanded Tchebichef space by capturing oscillating behavior of all textures that constitute the image.

2.3 Feature Extraction

Feature extraction with DTMs was introduced by Marcos et al. [13] on synthetic textures and used by Nava et al. [16] on emphysematous tissues. However, they compute a single vector using the whole image, which implies calculating high-order moments. According to [15], large Tchebichef vectors may introduce an error due to stability in the oscillations. Here, we present a modification based on sliding windows by implementing the following steps:

The scaled images are processed using a window of 40×40 pixels; the accuracy was used as the performance measure to evaluate the optimal window's size. The window

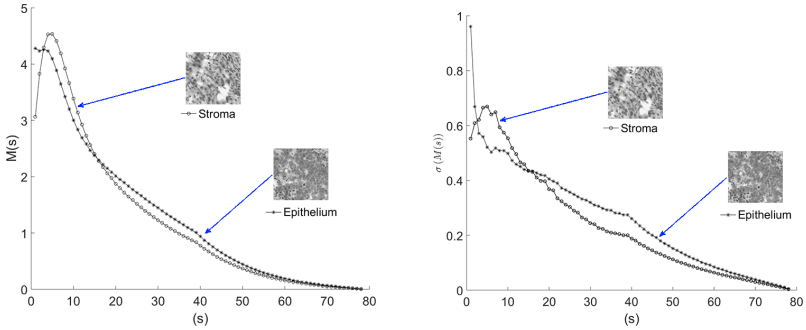


Fig. 3. DTM signatures. **(a)** Average and **(b)** standard deviation vectors obtained from tumor epithelium and stroma tissues.

is moved from the upper-left corner to the lower-right corner by 20 pixels per iteration, this means an overlap of 50%.

The corresponding $M(s)$ vectors are calculated on each window position. After this process is conducted over all possible windows, we obtained a set of vectors $M_i(s)$ where i indicates the window position. Since the images in the dataset are not the same size, then i varies among images. The feature vector is build as follows:

$\forall i \in$ the given image:

$$\begin{aligned} \bar{t} = [&\mu(M_i(1)), \sigma(M_i(1)), \beta(M_i(1)), \ldots, \\ &\mu(M_i(2N-2)), \sigma(M_i(2N-2)), \beta(M_i(2N-2))] \end{aligned} \tag{4}$$

where μ and σ are the mean and the standard deviation respectively. The operator β is the defined as: $\beta(x) = \frac{\sigma(x)}{\kappa(x)^{1/2}}$ and κ is the kurtosis.

Eq. (4) represents a novel way to describe texture characteristics. Note that the moment of order $s = 0$ is not used because it represents the mean value of the image. Furthermore, correlated coefficients between tumor epithelium and stroma are discarded by applying the p-test. The test reflects statistically significant differences ($p < 0.001$) between both groups, the features with a p-value greater than the threshold p are not included. The average Tchebichef signatures for both classes are shown in Fig. 3.

2.4 Classifier

A SVM and a k-NN classifier were implemented to validate our proposal. The classifiers were trained using a subset of 656 images and a different set with 720 images was used in the validation stage. Both image datasets were processed in the same manner described previously and the accuracy was the measure to assess the performance of the proposal.

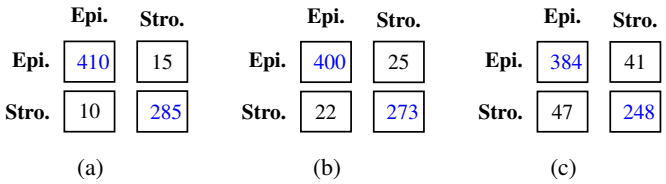


Fig. 4. Final classification results. Epi. and Strom. stand for tumor epithelium and tumor stroma, respectively. **(a)** DTMs with SVM. **(b)** DTMs with k-NN; and **(c)** LBPs with SVM.

3 Experimental results

Using a standard linear SVM classifier, our proposal labeled incorrectly 25 images out of 725, that means an accuracy of 96.53%. 15 images were wrongly classified as epithelium, whereas 10 samples were labeled as tumor stroma. We also computed the performance using k-NN with k = 11; the number of neighbors was not relevant in the classification performance. The best results are shown as confusion matrices in Fig. 4.

For comparison purposes, the LBP descriptor described in [18] was implemented. For each image, on every window position a feature vector was built by concatenating $\text{LBP}_{8,1}$ and $\text{LBP}_{16,2}$ histograms. Then, all the LBP feature vectors from a single image were grouped and characterized by the first two statistic moments: mean and standard deviation. Furthermore, we include results reported in [12] where the same database was used. Linder et al. propose a combined LBP/C descriptor to characterized the tumor texture.

We computed the ROC curve for our proposal, (see Fig. 5). The area under the ROC curve (AUC) is 0.9847, such a value is pretty similar to the AUC reported by Linder et al. We also calculated the $\text{F}_1\text{-Score} = 2 * \frac{\text{Precision}_{*}\text{Sensitivity}}{\text{Precision}_{+}\text{Sensitivity}}$ and all the results are summarized in Table 1.

4 Conclusions

We propose a novel method based on discrete Tchebichef Moments to classify tumor epithelium and stroma in a large database of colorectal cancer collected from TMAs.

Table 1. Comparison and classification results. All the data are expressed in (%). Bold values represent the best results.

Method	**Precision**	**Sensitivity**	**Specificity**	**F_1-Score**
DTMs/SVM	**96.47**	97.62	**95**	96.94
DTMs/KNN	94.12	94.79	91.61	94.45
LBPs/SVM	90.35	89.1	85.81	89.72
LBPs/KNN	91.53	83.48	85.83	87.32
LBP/C [12]	95.53	**99.02**	93.87	**97.19**

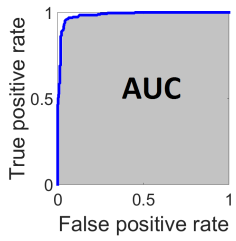


Fig. 5. ROC curve for DTMs/SVM proposal. The achieved AUC is 0.9847.

We have shown that Tchebichef moments possess the ability to describe textures by projecting the image of interest onto a polynomial basis where its sinusoidal-like behavior provides a suitable representation of all the textures that constitute the image. The sliding window approach improves the descriptor stability by discarding high-order moments and avoids the curse of dimensionality.

As in [12], our proposal achieved an accuracy rate above 96% (only 2 images below the LBP/C descriptor). Our method classifies better the epithelium tissue than LBP/C. Nevertheless, it is not possible to claim that there is a better performance because the difference between accuracies is only 0.28%. DTMs performance is about 6% better than LBPs, which indicates that our proposal captures texture variations in a better way. Furthermore, our proposal does not use contrast information, therefore, it is not necessary to quantize the images to get the local variance.

Acknowledgments

The authors extend their gratitude to Prof. Dr. Johan Lundin for providing the images. This publication was supported by the European social fund within the project CZ.1.07/2.3.00/30.0034 and UNAM PAPIIT grant IG100814. R. Nava thanks Consejo Nacional de Ciencia y Tecnología (CONACYT). G. González thanks CONACYT–263921 scholarship. J. Kybic was supported by the Czech Science Foundation project 14-21421S.

References

1. Calon, A., Lonardo, E., Berenguer-Llergo, A., Espinet, E., Hernando-Momblona, X., Iglesias, M., Sevillano, M., Palomo-Ponce, S., Tauriello, D.V., Byrom, D., Cortina, C., Morral, C., Barcelo, C., Tosi, S., Riera, A., Attolini, C., Rossell, D., Sancho, E., Batlle, E.: Stromal gene expression defines poor-prognosis subtypes in colorectal cancer. Nat Genet 47(4), 320–329 (2015)
2. Conti, J., Thomas, G.: The role of tumour stroma in colorectal cancer invasion and metastasis. Cancers 3(2), 2160 (2011)
3. Doyle, S., Feldman, M., Tomaszewski, J., Madabhushi, A.: A boosted bayesian multiresolution classifier for prostate cancer detection from digitized needle biopsies. Biomedical Engineering, IEEE Transactions on 59(5), 1205–1218 (2012)

4. Ferlay, J., Soerjomataram, I., Ervik, M., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., Parkin, D., Forman, D., Bray, F.: GLOBOCAN2012 v1.0, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11 (2014), http://globocan.iarc.fr/
5. Flusser, J., Suk, T., Zitová, B.: Introduction to Moments, pp. 1–11. John Wiley & Sons, Ltd (2009)
6. Foran, D.J., Yang, L., Chen, W., Hu, J., Goodell, L.A., Reiss, M., Wang, F., Kurc, T., Pan, T., Sharma, A., et al.: Imageminer: a software system for comparative analysis of tissue microarrays using content-based image retrieval, high-performance computing, and grid technology. Journal of the American Medical Informatics Association 18(4), 403–415 (2011)
7. Hayat, M.: Introduction: Colorectal cancer. In: Hayat, M. (ed.) Colorectal Cancer, Methods of Cancer Diagnosis, Therapy, and Prognosis, vol. 4, pp. 3–9. Springer Netherlands (2009)
8. Isella, C., Terrasi, A., Bellomo, S.E., Petti, C., Galatola, G., Muratore, A., Mellano, A., Senetta, R., Cassenti, A., Sonetto, C., Inghirami, G., Trusolino, L., Fekete, Z., De Ridder, M., Cassoni, P., Storme, G., Bertotti, A., Medico, E.: Stromal contribution to the colorectal cancer transcriptome. Nat Genet 47(4), 312–319 (2015)
9. Janowczyk, A., Chandran, S., Madabhushi, A.: Quantifying local heterogeneity via morphologic scale: Distinguishing tumoral from stromal regions. Journal of pathology informatics 4(Suppl) (2013)
10. Jemal, A., Bray, F., Center, M., Ferlay, J., Ward, E., Forman, D.: Global cancer statistics. CA: A Cancer Journal for Clinicians 61(2), 69–90 (2011)
11. Kwak, J.T., Hewitt, S.M., Sinha, S., Bhargava, R.: Multimodal microscopy for automated histologic analysis of prostate cancer. BMC cancer 11(1), 62 (2011)
12. Linder, N., Konsti, J., Turkki, R., Rahtu, E., Lundin, M., Nordling, S., Haglund, C., Ahonen, T., Pietikäinen, M., Lundin, J.: Identification of tumor epithelium and stroma in tissue microarrays using texture analysis. Diagnostic Pathology 7(1), 22 (2012)
13. Marcos, J.V., Cristóbal, G.: Texture classification using discrete Tchebichef moments. J. Opt. Soc. Am. A 30(8), 1580–1591 (2013)
14. Mukundan, R.: Some computational aspects of discrete orthonormal moments. IEEE Transactions on Image Processing 13(8), 1055–1059 (2004)
15. Mukundan, R., Ong, S., Lee, P.: Image analysis by Tchebichef moments. IEEE Transactions on Image Processing 10(9), 1357–1364 (2001)
16. Nava, R., Marcos, V., Escalante-Ramírez, B., Cristóbal, G., Perrinet, L., Estépar, R.: Advances in texture analysis for emphysema classification. In: Ruiz-Shulcloper, J., Sanniti di Baja, G. (eds.) Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications (CIARP). Lecture Notes in Computer Science, vol. 8259, pp. 214–221. Springer Berlin Heidelberg (2013)
17. Nicholson, A.D., Guo, X., Sullivan, C.A., Cha, C.H.: Automated quantitative analysis of tissue microarray of 443 patients with colorectal adenocarcinoma: Low expression of bcl-2 predicts poor survival. Journal of the American College of Surgeons 219(5), 977–987 (2014)
18. Ojala, T., Pietikäinen, M., Maenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence 24(7), 971–987 (2002)
19. Simon, R., Mirlacher, M., Sauter, G.: Tissue microarrays in cancer diagnosis. Expert Review of Molecular Diagnostics 3(4), 421–430 (2003)
20. Wang, C.W., Fennell, D., Paul, I., Savage, K., Hamilton, P.: Robust automated tumour segmentation on histological and immunohistochemical tissue images. PloS one 6(2), e15818 (2011)