American Journal of Artificial Intelligence
2024, Vol. 8, No. 2, pp. 41-47
https://doi.org/10.11648/j.ajai.20240802.12

Research Article

Investigation of Histological Image Classification Methods Using Different Feature Extraction Techniques

Nomaz Mirzaev, Farkhod Meliev*

Laboratory of Biometric Systems, Digital Technologies and Artificial Intelligence Research Institute, Tashkent, Uzbekistan

Abstract

This paper examines the performance of different machine and deep learning algorithms in classifying colon histological images using different feature extraction methods. The relationship between the feature extraction methods and the selected machine learning methods to improve the classification accuracy is analyzed. Widely used methods like local binary patterns, histograms of oriented gradients, Gabor filter and Dobeshi wavelets are investigated for feature extraction from colon histological images. The features extracted by histogram of oriented gradients and Gabor filter methods were used as a single joint feature vector. And popular machine learning methods such as Support vector machine, Decision trees, Random forest, k-nearest neighbors and Naive Bayesian method were used to classify the selected images. The paper also investigates ensemble methods using gradient bousting and voting classifier as examples. The authors also focus on the study of convolutional neural networks as they are one of the main deep learning methods at the moment. The classification methods selected for analysis are compared in terms of classification accuracy and time taken for training and recognition. All pre-defined and adjustable parameters of both feature extraction methods and classification methods were personally selected by the authors as a result of experimental studies, which were conducted using a software tool created in the Python programming language on a set of LC25000 histological images. The software created is easily customizable and can be used in the future to investigate classification methods on other datasets.

Keywords

LBP, HOG, Dobeshi, Feature Extraction, Classification

1. Introduction

Histologic images are an integral part of medical diagnosis, providing visual data about the tissues and cells of the body [1]. Automating the analysis of these images using machine learning and feature extraction methods is crucial for improving the accuracy and speed of diagnosis. In recent years, the development of deep learning and feature extraction methods has led to significant progress in the development of histology image classification algorithms [2].

Classification of histological images is a challenging task due to the high heterogeneity and complexity of the data [3]. Traditional analysis methods are often inefficient, which requires the application of modern machine learning approaches. Feature extraction methods play an important role in this context by transforming raw image data into a more informative representation that can be effectively utilized by classifiers.

*Corresponding author: nomazmirza@rambler.ru (Farkhod Meliev)

Received: 6 July 2024; **Accepted:** 6 August 2024; **Published:** 20 August 2024

There are many approaches to extract features from images [4]. One of the most common methods is the use of Histograms of Oriented Gradients (HOG), which is often used in object recognition tasks [5]. Another important method is the use of Gabor filters, which can extract texture features from an image [6]. Wavelet transforms can also extract important structural features from images at different scales, which is particularly useful for analyzing complex textures and patterns in histological data [7]. Local Binary Patterns (LBP) - another powerful method for analyzing image textures that is widely used in computer vision and pattern recognition. LBP is a texture operator that transforms an image into a new view that emphasizes textural features. The basic idea of the method is to create a binary code for each pixel of an image based on the comparison of intensity values of the pixel and neighboring pixels [8]. Recently, however, the focus has been on deep learning methods such as Convolutional Neural Networks (CNN), which can automatically extract features at different levels of abstraction [9].

After feature extraction, the next step is to use them for image classification. Common classification methods include naive Bayesian classification methods, k-nearest neighbor (kNN) methods, decision trees, and ensemble methods such as Random forests and Gradient boosting. Each of these methods has its own advantages and limitations that need to be considered when selecting an approach for a particular problem.

The objective of this study is to comparatively analyze different methods of histological image classification using different feature extraction techniques, also, to expand the knowledge in the field of automated histological image analysis, to contribute to the development of more accurate and efficient diagnostic methods in medical practice and to achieve higher accuracy of histological image classification.

The scientific novelty of the study is as follows:

Analysis of feature extraction methods:

Comprehensive analysis of different methods of LBP, HOG, Dobeshi wavelet, and Gabor filter as methods for feature extraction from histological images. Evaluation of the performance of these methods in histologic image classification tasks.

Combining feature extraction approaches:

An approach combining features extracted by multiple methods to create a more informative and diverse feature set is studied. The effectiveness of the combined approach compared to using each method separately is investigated, which can improve classification accuracy.

Comprehensive testing and validation:

Some combinations of feature extraction methods and classification algorithms have been tested on a large histological image dataset.

Applications of machine learning and deep learning:

The effectiveness of various Machine learning algorithms (e.g. Support Vector Machine, Random Forest) and Deep learning algorithms (e.g. convolutional neural networks) in classifying histological images was investigated.

The relationship between the type of extracted features and the chosen classification algorithm is analyzed.

2. Problem Statement

Let us denote the histological image as I, where $I \in \mathbb{R}^{m \times n}$ is a matrix of pixel intensities of size $m \times n$.

The task of classification is to assign each I image class labels $y \in \{1, 2, \ldots, K\}$, where K is the number of classes.

To achieve this goal, the following steps are followed:

Feature extraction: using different methods to convert an image I into a feature vector f.

Classification: application of machine learning algorithms to classify images based on feature vectors f.

3. Materials and Methods

3.1. Feature Extraction Methods

3.1.1. Local Binary Templates

It is a method for extracting texture features of an image that analyses local pixel intensity patterns. The basic idea of the method is to compare pixel values in the vicinity of the central pixel and encode the result as a binary number.

The basic steps of the LBP method:

1) Definition of environment:

For each pixel p with coordinates (i, j) in image I, determined the surroundings with radius R and P neighboring pixels. The surroundings can be square, circular.

2) Calculation of intensity differences:

Compares intensity central pixel $I(i, j)$ with the intensities of its neighbors I_p, Where $p \in \{0,1, \ldots, P-1\}$.

3) Binarization:

The Heaviside function [10] is used to obtain the binary value of $s(x)$:

$$s(x) = \begin{cases} 1 \text{ if } x \geq 0 \\ 0 \text{ if } x < 0. \end{cases} \quad (1)$$

Thus, the binary value for each neighbor is calculated as:

$$s(I_p - I(i, j))$$

4) Formation of binary code:

The resulting binary values are combined into a single binary number:

$$LBP(i, j) = \sum_{p=0}^{P-1} s(I_p - I(i, j)) \cdot 2^p \quad (2)$$

This value is the LBP code for the central pixel.

5) Construction of a histogram of LBP codes:

A histogram of occurrence frequencies of each LBP code is constructed for the whole image. The histogram is used as a

feature vector describing the image texture.

3.1.2. Histograms of Oriented Gradients

Histogram of Oriented Gradients is a feature extraction method used in computer vision to describe local objects and textures. The HOG method is based on the distribution of intensity gradient directions in local areas of the image.

The basic steps of the HOG method:

1) Computation of image gradients:

For each pixel of the image $I(i,j)$ we calculate gradients in the direction of axis x and axis y:

$$G_x(i,j) = I(i,j+1) - I(i,j-1),$$

$$G_y(i,j) = I(i+1,j) - I(i-1,j)$$

Let's use the obtained values to calculate the gradient value $G(i,j)$ and the direction of the gradient $\theta(i,j)$:

$$G(i,j) = \sqrt{G_x(i,j)^2 + G_y(i,j)^2} \quad (3)$$

$$\theta(i,j) = arctan\left(\frac{G_x(i,j)}{G_y(i,j)}\right) \quad (4)$$

2) Dividing an image into cells:

The image is divided into small cells and blocks of equal size.

3) Construct histograms of oriented gradients for each cell:

For each cell, a histogram of gradient directions is plotted and divided into a fixed number of bins. For each pixel (i,j) in the cell, the gradient value is weighted $G(i,j)$ and this value is added to the corresponding histogram bin determined by the gradient direction $\theta(i,j)$.

4) Block normalization:

Several neighboring cells are combined into a block. The block feature vector is formed by combining the histograms of all cells in the block. The resulting vector is normalized to reduce the effect of light variation and contrast:

$$v' = \frac{v}{\sqrt{\|v\|_2^2 + \epsilon^2}} \quad (5)$$

where v is the histogram vector for the block, $\|v\|_2$ - its norm, and ϵ is a small value to prevent division by zero.

5) Formation of a finite vector of features:

Normalized vectors of all blocks are combined into one final feature vector describing the image.

3.1.3. Gabor Filter

Gabor filters are powerful tools for extracting texture features from images. They are used in various computer vision tasks such as object recognition, texture classification and edge detection. The filter is a band-pass filter optimized for extracting spatial frequency features in images.

The main steps of feature extraction using Gabor filter are:

1) Selection of Gabor filter parameters:

Defining a set of filter parameters such as frequency, direction, bandwidth and scale. Usually several filters with different parameters are used to cover different spatial and frequency characteristics of the image.

2) Creating Gabor Filters:

Each Gabor filter $g(x,y)$ is a complex exponent with a Gaussian envelope. The general formula of the Gabor filter:

$$g(x,y;\lambda,\theta,\psi,\sigma,\gamma) = exp\left(-\frac{x'^2+\gamma^2y'^2}{2\sigma^2}\right)\cdot cos\left(2\pi\frac{x'}{\lambda}+\psi\right)$$

where $x' = xcos\theta + ysin\theta$, $y' = -xsin\theta + ycos\theta$, λ - is the wavelength (determines the frequency), θ - filter orientation, ψ - phase shift, σ - width of the Gaussian envelope, γ - aspect ratio (determines ellipticity).

3) Applying a filter to an image:

Applying each Gabor filter to the image using convolution:

$$I_{filtered}(x,y) = I(x,y) * g(x,y;\lambda,\theta,\psi,\sigma,\gamma) \quad (6)$$

where $*$ denotes the convolution operation.

4) Extraction of amplitude and phase features:

The amplitude and phase components are extracted from the resulting complex image:

$$A(x,y) = |\, I_{filtered}(x,y)\,|$$

$$\phi(x,y) = arg(I_{filtered}(x,y)) \quad (7)$$

5) Construction of a feature vector:

For each pixel or for each local region of the image, a feature vector including amplitude and/or phase components for all applied Gabor filters is generated.

6) Post-processing and normalization:

Feature vectors can be normalized or transformed to improve their performance and meet the requirements of machine learning algorithms. Additional processing methods such as block averaging, histograms and dimensionality reduction techniques can be applied.

3.1.4. Dobeshi Wavelets

Dobeshi wavelets (DW) are among the most popular and widely used wavelets in image and signal processing problems. They have good time-frequency characteristics and can efficiently decompose images into different frequency components. Let us consider the basic steps in extracting features from images using Dobeshi wavelets.

1) Selection of wavelet and decomposition level:

The type of wavelet is defined (e.g. db1, db2, etc.). Dobeshi wavelets are denoted as dbN, where N is the wavelet order. In our case, db1 was used. Then the number of decomposition levels is determined. Usually several levels are chosen to capture features at different scales.

2) Application of wavelet transform:

A discrete wavelet transforms (DWT) is applied to the image to decompose it into sub-bands of frequencies. At each level of decomposition, the image is divided into four sub-bands: approximation, horizontal detail, vertical detail and diagonal detail.

3) Feature extraction:

From the resulting sub-bands, features such as mean, standard deviation, energy, etc. are extracted. The features can be extracted for each decomposition level and each sub-band.

4) Construction of a feature vector:

A feature vector is formed, combining all extracted features from different levels and sub-ranges. The feature vector can be normalized or further processed to improve its characteristics and can be used in classification, clustering, object recognition and other machine learning tasks.

3.2. Classification Methods

Machine learning algorithms such as SVM (Support Vector Machine), Random Forest, Design Tree, kNN, Naive Bayes, Gradient Boosting, Voting Classifier and convolutional neural networks are investigated for classification of histological images.

Feature extraction from images was carried out by the following parameters:

1) When using LBP: $R = 1$, $P = 8 * R$, the method for counting is universal;
2) When using HOG: cell size 8×8 block size 2×2;
3) When using the Gabor filter: $\lambda = 4$, $\theta = 1$, every next step $\theta = \theta/4 * \pi$, $\psi = 10$, $\sigma = 0{,}5$, $\gamma = 0$;
4) When using Dobeshi wavelet: wavelet type is selected as db1, decomposition level is 3.

The features extracted by the HOG and Gabor methods were used merged into one array.

The following parameters were used in training the models:

1) *When using Random Forest method:* the number of trees in the forest is 100. The random selection is equal to 42;
2) *When using the SVM method:* the kernel type is selected as radial-basis. The number of iterations is 10000;
3) *When using the Decision Tree method:* the function for evaluating the quality of partitions is set as "gini". The number of random number generation for mixing data during tree construction is equal to 42;
4) *Using the kNN method:* the number of nearest neighbours used for predictions is 5;
5) *When using the Naive Bayes method:* the smoothing parameter is equal to 10^{-9};
6) *When using the Gradient Boosting method*: the number of trees in the ensemble is 100. Randomness for reproducibility is equal to 42;
7) *When using Voting Classifier method:* list of methods included in the ensemble: logistic regression, support vector method, decision tree. Voting type "soft" predictive class based on averaging the predicted probabilities of all evaluators.
8) *When using convolutional neural networks:* number of hidden layers 2, GlobalAveragePooling2D global averaging was used for each image channel, 128 neurons were used in the first hidden layer and 64 in the second, 'Relu' activation function was used in both layers, output layer with 2 neurons and softmax activation function which converts the outputs into probabilities summing to 1.

4. Results

Histological images of colon (Figure 1) from the LC25000 set of 5000 images for each of the two classes (malignant, benign) were used for the experimental studies [11].

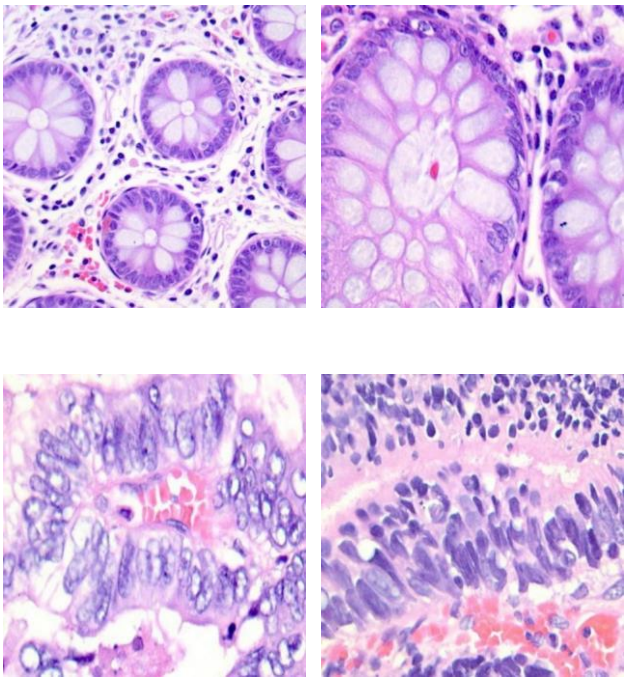

Figure 1. *Histological images of the colon.*

The above classification methods were compared in terms of recognition quality, training and recognition time consumption. The experimental results of classification by the used methods and using CNN are shown in Tables 1 and 2, respectively.

Table 1. Experimental results of the selected methods.

Classification method	Feature extraction method								
	LBP			HOG+ Gabor filter			Dobeshi wavelets		
	Time (sec.)		Accuracy (%)	Time (sec.)		Accuracy (%)	Time (sec.)		Accuracy (%)
	Training	Recognition		Training	Recognition		Training	Recognition	
SVM	0.36	0.34	87.24	2598	1118.93	83.41	254.35	202.91	76.33
Random forest	1.71	0,04	90.73	71.04	0.26	82.01	48.56	0.15	79.17
Design Tree	0.05	0.02	84.75	207.97	0.24	61.39	85.64	0.06	65.42
kNN	0.014	0.05	88.01	0.65	4.89	50.32	0.15	2.01	51.17
Naive Bayes	0.02	0.04	64.58	4.42	1.58	78.62	1.38	0.44	78.72
Gradient Boosting	2.16	0.02	84.42	5560.5	0.3	79.63	1799.39	0.11	81.51
Voting Classifier	2.43	0.32	87.22	53669.9	1627.83	80.27	2069.11	224.53	69.46

Table 2. Experimental result when using CNN.

Classification method	Time (sec.)		Accuracy (%)
	Training	Recognition	
CNN	12936.11	194.7	98,2

5. Discussion

A study of histological image classification methods using different feature extraction techniques has identified the most effective approaches for automatic diagnosis based on histological images. The main methods considered in this study include traditional classification and feature extraction methods such as local binary patterns, histograms of oriented gradients, Gabor filters, as well as modern approaches based on convolutional neural networks.

Experimental results showed that Random forest method showed the best classification accuracy (90.73%) when using LBP method for feature extraction. Ensemble Gradient Boosting method showed good classification result (81.51%) when using Dobeshi wavelet as a feature extraction method. The SVM method showed high classification accuracy (83.41%) when using the combined features extracted by HOG and Gabor filter methods. The convolutional neural networks showed the best performance compared to traditional feature extraction methods. In the experiments, CNNs achieved the highest classification accuracy (98.2%), outperforming traditional methods, although it takes quite longer time to train the data. This confirms that deep learning and CNNs are the most promising approaches for automatic diagnosis tasks based on histological images.

6. Conclusions

A study of histology image classification methods using different feature extraction methods and classifiers showed that different approaches have their own advantages and disadvantages.

In feature extraction using LBP, it was shown that Random Forest method has the best classification performance compared to other methods. This proves that this classification method is suitable for tasks that require fast learning on large datasets. And with the combined use of features extracted by HOG method and Gabor filter, the best classification result was shown by SVM, although it took more time for its training and recognition compared to Random Forest method, which also showed not bad classification result. Ensemble Gradient Boosting method was the best among other methods in feature extraction using Dobeshi wavelet.

Although the training time was longer than other methods, the recognition time and accuracy was comparatively high. Studies have also shown that, with a large amount of data for training the use of convolutional neural networks give the best classification result, although it takes quite a long time to train them.

Thus, this work allowed us to explore the possibilities of popular classification methods using different methods for feature extraction. It can be concluded that, when the need for fast training and classification is recommended to use Random Forest, kNN or SVM methods, and when the requirement of a sufficiently accurate classification using convolutional neural networks show comparatively better results. The results of the study can be useful in improving the accuracy and reliability of automated diagnostics and used in the creation of systems aimed at classifying histological images.

Abbreviations

HOG	Histograms of Oriented Gradients
LBP	Local Binary Patterns
CNN	Convolutional Neural Networks
DW	Dobeshi Wavelets
DWT	Discrete Wavelet Transforms
SVM	Support Vector Machine
kNN	k-Nearest Neighbor

Author Contributions

Nomaz Mirzaev: Conceptualization, Formal Analysis, Methodology, Data curation

Farkhod Meliev: Investigation, Resources, Writing – original draft, Writing – review & editing

Data Availability Statement

Not applicable.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Gurcan, M. N., Boucheron, L. E., Can, A., Madabhushi, A., Rajpoot, N. M. and Yener, B. Histopathological image analysis: A review. In IEEE Reviews in Biomedical Engineering, 2009, vol. 2, pp. 147-171. https://doi.org/10.1109/RBME.2009.2034865

[2] De Matos, J., Ataky, S. T. M., de Souza Britto Jr, A., Soares de Oliveira, L. E. and Lameiras Koerich, A. Machine learning methods for histopathological image analysis: A review, Electronics. 2021, 10(5), 562. https://doi.org/10.3390/electronics10050562

[3] Aeffner, F., Zarella, M. D., Buchbinder, N., Bui, M. M., Goodman, M. R., Hartman, D. J., Lujan, G. M., Molani, M. A., Parwani, A. V., Lillard, K. and Turner, O. C. Introduction to digital image analysis in whole-slide imaging: a white paper from the digital pathology association, Journal of pathology informatics. 2019, 10(1), p. 9. https://doi.org/10.4103/jpi.jpi_82_18

[4] Mutlag W. K. et al. Feature extraction methods: a review, Journal of Physics: Conference Series. – IOP Publishing, 2020. – T. 1591. – №. 1. – C. 012028. https://doi.org/ 10.1088/1742-6596/1591/1/012028

[5] Dalal, N., Triggs, B. Histograms of oriented gradients for human detection. In IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), 2005, San Diego, CA, USA, pp. 886-893 vol. 1. https://doi.org/10.1109/CVPR.2005.177

[6] Manjunath, B. S. and Ma, W. Y. Texture features for browsing and retrieval of image data. In IEEE Transactions on pattern analysis and machine intelligence 1996. 18(8), pp. 837-842. https://doi.org/10.1109/34.531803

[7] Mallat SG. A theory for multiresolution signal decomposition: the wavelet representation. In IEEE transactions on pattern analysis and machine intelligence 1989. 11(7), pp 674-693. https://doi.org/10.1109/34.192463

[8] Ojala, T., Pietikäinen, M., Harwood, D. A comparative study of texture measures with classification based on featured distributions. Pattern recognition. 1996, 29(1), pp. 51-59. https://doi.org/10.1016/0031-3203(95)00067-4

[9] LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature, 521(7553), 436-444. https://doi.org/10.1038/nature14539

[10] Wyld, Henry William, and Gary Powell. Mathematical methods for physics. CRC Press, 2020. https://doi.org/10.1201/9781003037460

[11] Borkowski, A. A., Bui, M. M., Thomas, L. B., Wilson, C. P., DeLand, L. A. and Mastorides, S. M., 2019. Lung and colon cancer histopathological image dataset (lc25000). arXiv preprint arXiv:1912.12142.

Biography

Nomaz Mirzaev: Since 2020 – Doctor of Technical Sciences. He defended his doctoral dissertation in 2020 at the Tashkent University of Information Technologies on the topic "Methods for the synthesis of recognition operators in the space of large characters" 01/05/03 - Theoretical foundations of computer science. From 2021 to the present, he is the project manager at the Research Institute for the Development of Digital Technologies and Artificial Intelligence. He has published 150 scientific papers, 50 of them after defending dissertations.

Farkhod Meliev is a Doctoral student at the Research Institute for the Development of Digital Technologies and Artificial Intelligence. Conducts research in the field of Digital technologies and artificial intelligence.
His research activities focus on the application of advanced methods and technologies for visual data analysis and processing. His work includes the study of various algorithms and approaches aimed at improving the quality and accuracy of automatic image processing. Also explores the application of popular machine learning methods in the field of medical diagnostics.

Research Field

Nomaz Mirzaev: Pattern recognition, mathematical statistics, probability theory, machine learning.
Farkhod Meliev: Image processing, machine learning