COVID-19 detection based on combined domain features

Omar Munthir Al Okashi¹, Ismail Taha Ahmed¹, Leith Hamid Abed²
¹College of Computer Sciences and Information Technology, University of Anbar, Anbar, Iraq
²Department of Computer Systems Techniques, Anbar Technical Institute, Middle Technical University, Anbar, Iraq

Article Info

ABSTRACT

The computed tomography (CT) scan delivers more detailed information and higher judgment accuracy than a chest X-ray, which has a wide range of uses in diagnosing and decision-making to aid medical professionals. This paper proposed a method to detect COVID-19 from CT scan images using the combination of spatial domain and transform domain features. Using the lung segmentation step, the CT image is first processed and segmented, and then various domain features are extracted. From these domain features, the highest combined domain features (CDF) are obtained. Finally, the detection task is completed using random forest (RF) and Naive Bayesian (NB) classifiers. The proposed method is tested using a dataset of CT scan images, and the results are compared to several current techniques. The results showed that our method based on CDF outperforms previous methods, with an overall accuracy of nearly 98%. As can be shown, CDF is the best domain feature to apply for detecting COVID-19.

Keywords:
Combined domain features
Computed tomography
COVID-19
Naive bayesian
Random forest

Corresponding Author:
Ismail Taha Ahmed
College of Computer Sciences and Information Technology, University of Anbar
Anbar, Iraq
Email: ismail.taha@uoanbar.edu.iq

1. INTRODUCTION

COVID-19 that resulted from corona virus is one of the most recent epidemic that threat the world and it mainly affects lungs of the human. To diagnose this disease, doctors usually tend to image the chest. Computed tomography (CT) imaging technique provides more clear information and greater judgment accuracy than the chest X-ray as shown in Figure 1. Figure 1(a) shows the samples of X-ray COVID-19 image and Figure 1(b) shows CT scan COVID-19 image. Therefore, CT is one of the best choices that used to detect COVID-19. Usually, doctors need to inspect the CT images carefully before make a decision whether the lung is infected or not. However, the huge development in machine learning techniques and computer hardware could aid in make fast and accurate decision for COVID-19 depending on learning process.

Several researchers discussed automatic detection of COVID-19 from different medical images such as X-ray and CT scan images [1], [2]. The work presented in [3] was based on extract handcrafted features local binary patterns (LBP) and gray level co-occurrence matrix (GLCM) from X-ray images and using them to train a neural network model to classify COVID-19 cases from other non-COVID cases. In term of feature fusion, [4] suggested to fuse the histogram of oriented (HOG) and CNN features that were extracted from X-ray images to train a convolutional neural network (CNN) model to detect COVID-19. Combination of LBP and different textural features that extracted from X-ray images to train a COVID classifier was also proposed in [5]. However, the authors didn’t specify whether the combination enhance the result. The detection of COVID-19 from chest X-ray image was also presented in [6]. The proposal is mainly focused on merging extracted CNN and wavelet transform features before using them to train a random forest (RF)
model to detect COVID-19 cases in a two-level fashion. Zhang et al. [7] was aiming to combine between handcrafted and deep learning features to train an SVM classifier to classify an X-ray images of chest to healthy, regular pneumonia, and COVID-19. Classification of chest X-ray images into COVID and non-COVID was also presented in [8]. The idea was to extract GLCM features and then using them to train latent-dynamic conditional random field (LDCRFs) model for fine classification. Despite the fact that X-ray is an available and cheap method to diagnose COVID-19, it still not the best way to perform this task due to lack of clarity in the chest image. Some researchers focused on automatic detection of COVID-19 from CT images. Ameer and Mohammed [9] focused on handcrafted features (GLCM) extracted from CT images of infected and not infected lung to classify them using Euclidian distance. Wu et al. [10] Proposed to train a RF classifier using a texture feature that extracted from CT images by modified wavelet transform and matrix computation analysis. The resulted model aimed to recognize COVID-19 from other infectious pneumonias.

After reviewing the literature about COVID-19 detection, we found that it is important to put more effort in this subject. This is mainly to increase the accuracy and reduce the high dimensionality of extracted features which will lead to enhance the time complexity. Therefore, in our paper, we proposed a COVID-19 detection approach based on several domains such as spatial and transform domains to improve classification results. Random forest (RF) and Naive Bayesian (NB) are two common classifiers which will be used in the classification process. The structure of the article is broken down as shown in: Section 2 outlined the proposed method; Section 3 discusses the experimental findings; The conclusions are presented in Section 4.

2. PROPOSED METHOD

The proposed approach for effective COVID-19 detection, which consists of four steps, is described in the following subsections (pre-processing, features extraction, combined feature vector, and classification). Figure 2 depicts the steps of the proposed method as a flowchart. See the following subsection for more details on each step.

2.1. Lung segmentation

The CT scan slide that shows lungs contains other unimportant details such as ribs, bones along with the information of the patient. To make a useful features extraction and images classification at the later steps, we need to make an accurate segmentation to the lungs and exclude other objects. The main steps for the lungs (in Figure 3) segmentation are given as:

a) Image conversion from RGB to gray-scale are shown in Figure 3(a). This is mainly to reduce complexity in dealing with the image without losing any important information.
b) Threshold the gray scale image Figure 3(a) with a suitable value. The thresholding value should be high enough to highlight bright unwanted objects as shown in Figure 3(b).

c) Complement the resulting image Figure 3(b) and remove the borders to keep the lungs objects Figure 3(c).

d) Label the objects in the binary image and using the area property to keep the two biggest objects (lungs) and exclude others Figure 3(d).

e) Mask the gray scale image Figure 3(a) by the binary image Figure 3(d) to get the final segmented lung image Figure 3(e).

f) Automatic cropping for the segmented lung image Figure 3(e) to produce the final cropped segmented image Figure 3(f). This cropping is based on labeling the mask image Figure 3(d) to find the required landmarks (upper, lower, left and right) for accurate cropping.

![Figure 3](https://example.com/figure3.png)

Figure 3. Steps of lung segmentation: (a) the initial grayscale image, (b) the binary image, (c) complemented image, (d) lung’s mask, (e) segmented lung image, and (f) cropped segmented lung image

2.2. Features extraction

The features extraction stage is crucial in the image classification process. Here, the features are extracted from spatial and transform domains. From spatial domain, segmentation-based feature texture analysis (SFTA), local binary patterns (LBP) and gray level co-occurrence matrix (GLCM) features are extracted. The requirement is to extract useful features from various domains as explained in the next subsections.

2.2.1. Spatial domain

Due to its role in revealing information concealed inside the image that is difficult to discern with the human eye, texture analysis plays a particular role in most image processing domains. Texture features is one of the most effective feature extraction approaches for CT images. A total of 94 texture features were collected for each image. They are classified into three categories: SFTA, LBP, and GLCM features. Next, we will go over each feature in detail.

a). Segmentation-based feature texture analysis (SFTA)

Two steps make up the core procedure of the segmentation-based feature texture analysis (SFTA) [12] technique. In the first stage, decomposing the supplied grayscale image into a collection of binary
images is achieved. The data was decomposed using the two-threshold binary decomposition (TTBD) approach. For each binary image formed utilizing the fractal dimension from its regions’ boundaries, SFTA feature vectors are computed in the second phase. We also compute the mean gray level and size of the areas. SFTA is the texture feature of choice for analyzing a texture image because of its stability as well as its minimal calculation cost. As a result, applying SFTA features extraction to COVID-19 detection is intriguing. For more details, see [12]. The SFTA features have a dimension of 1x21.

b). Local binary patterns (LBP)

According to Ojala et al. developed the local binary patterns (LBP) approach, which is a visual descriptor used to describe textural properties inside an image [13], [14]. LBP is a robust texture descriptor that has been developed and tested in a variety of texture classification applications. A window of a given size (e.g., 3x3) moves over the entire image to acquire LBP features. A comparison process is performed between the center of the window and all other surrounding pixels through its movement. If any of the neighboring pixels are less than the center pixel, their value is set to 0, otherwise it is set to 1, as seen in Figure 4. Figure 4(a)-(g) show the main process for feature vector creation using LBP. The LBP thresholding procedure is illustrated in (1).

\[
LBP_{p,r} = \sum_{i=0}^{P-1} S(g_i - g_c) 2^i S(x) = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases}
\]

Where \( P \) denotes the window's neighboring pixels, \( R \) denotes the radius of a neighborhood, \( g_i \) denotes the intensity of the window's neighboring pixels, and \( g_c \) is the value of the central pixel. This method generates a feature vector with 59 values.

\[ (1) \]

Figure 4. Feature vector creation using LBP approach [3], (a) image division, (b) thresholding, (c) first binary creation, (d) second binary creation, (e) binary converted into decimal no, (f) binary no gneration, and (g) final LBP

c). Gray level co-occurrence matrix (GLCM)

This method involves using a gray level co-occurrence matrix (GLCM) established by Haralick et al. [15] to extract features from a CT image. This matrix records the number of times a pixel of a particular gray level intensity interacts with another pixel in the image via a rule or configuration known as offset. It can be compared horizontally at 0 degrees, vertically at 90 degrees, diagonally at 45 degrees, or 135 degrees.

The GLCM characterizes an image by generating a histogram of co-occurring greyscale values at a given offset and direction across it [15], [16]. As seen in Figure 5, features are found by directly using the greycomatrix() function from the skimage library to each image with an offset of 1 and in four distinct directions (0, 45, 90, 135). This method creates a 14-valued feature vector.

COVID-19 detection based on combined domain features (Omar Munthir Al Okashi)

2.2.2. Curvelet domain

Curvelet transforms are a type of image representation that uses many fewer large-magnitude coefficients than spatial domain representations. Geometry bases in various locations, scales, and orientations are used to accomplish this. As indicated in (2) [17], the image is divided into sets of curvelet coefficients $C(j, l, k_1, k_2)$ using the fast discrete curvelet transform (FDCT). Curvelet 2.1.2 toolbox, accessible at http://www.curvelet.org, was used to do the curvelet transformation.

$$C^D(j, l, k_1, k_2) = \sum_{0 \leq m < M} \sum_{0 \leq n < N} f[m, n] \phi^{D}_{j,l,k_1,k_2}[m, n]$$  \hspace{1cm} (2)

Where $\phi_{j,l,k_1,k_2}$ is a curvelet coefficients at various scales, $j$, orientations, $l$ and location, $(k_1, k_2)$.

The curvelet decomposition can separate an image into three levels: coarse, detail, and fine. The coarse category was assigned to the low-frequency coefficients. Fine was ascribed to the high-frequency coefficients. Detail was attributed to the middle-frequency coefficients. The scale $j$ is from finest to coarsest, and angle $l$ starts at the top-left corner and advances clockwise, according to fast discrete curvelet transform (FDCT) WARPING [16].

In this paper, each image is split into five levels of scales based on its size. Different scales have varying amounts of sub bands in varied orientations. The scales 1, 2, 3, 4, and 5 have 1, 16, 32, 32, and 1 sub bands, respectively, for a 5 level decomposition. There is no additional process in some of the five levels decomposition because level 1 and 5 include only one sub-band without orientation. Fine scale coefficients indicate the presence of local information in an image [18], [19]. Large-magnitude coefficients only appear in portions of the image that include fine details. In addition, the Curvelet transform provides an effective image representation with far fewer large-magnitude coefficients. As a result, the energy is calculated using the mean() of the top 0.1% curvelet coefficients at finest scale 4 ($\theta^4$) as defined by (3):

$$\text{Energy} = (\mu(\theta^4))$$  \hspace{1cm} (3)

because the finest scale 4 is configured to encompass 32 separate sub bands, we compute the mean values of each sub-band as shown in (4).

$$\mu(\theta^4) = m_1, m_2, m_3, ..., m_{32}$$  \hspace{1cm} (4)

Finally, for each image, a total of 32 features were obtained.

2.2.3. Combined domain features (CDF)

Multiple features can be extracted from multiple domains such as spatial domain and transform domain. In this phase, we merge the feature vectors obtained in steps 2.2.1 and 2.2.2 which have feature dimensions of 94 and 32, respectively. The total dimension of the CDF vector is 126.
2.3. Classification

Because a CT image can be classified as infected or non-infected, we need to identify ways to classify those images. Machine learning is the finest solution to employ. Machine learning algorithms are presented as a means of learning and making smart decisions automatically. The classification step employs a RF and a NB classifier.

2.3.1. Random forest (RF)

RF is a common machine learning method that builds a classifier using a decision tree (DT) based ensemble technique [20]. RF is faster and more accurate than other methods, and it can handle huge features with little samples using decision trees [21]. In terms of efficiency, RF beats classic machine learning algorithms such as artificial neural networks and support vector machines [22].

2.3.2. Naive Bayesian (NB)

NB classifier is built on Bayes’ theorem. It’s a simple probabilistic classifier that calculates a set of probabilities by counting the frequency and combinations of values in a dataset. It is assumed that the chance of one characteristic has no influence on the likelihood of the others [23], [24].

3. RESULTS AND DISCUSSION

The experimental results of applying our proposed approach to the CT scan images dataset are presented and analyzed in this part. The experiment was based on an HP laptop with an Intel Core i7 processor operating at 2.60 GHz and 8 GB of RAM running Microsoft Windows 10 64-bit (OS). The proposed technique was developed in MATLAB version R2020a.

3.1. Data set

Because the CT provides more clear information and greater judgment accuracy than the chest X-ray, this study focused solely on the CT scan images dataset for evaluate the proposed method. The data set used in this study consists of 60 CT scan images that were collected from Fallujah General Hospital. The data set are categorized as shown in: The first group provides non-infected CT scan images and it involves 30 images. The CT scan images with infection are included in the second group with total number of 30 images. Some of CT scan image examples (after lung segmentation step) from the dataset are shown in Figure 6(a)-(b).

![Figure 6. Samples of segmented lungs CT scan images in the dataset (a) non-infected images and (b) infected images](image)

The dataset is divided into two groups due to the lack of an independent dataset. The first group contains 70% of the images utilized in the training. The second group contains 30% images that will be utilized to test the suggested method. Throughout the evaluation experiments, a k-fold cross-validation approach (with k=10) was utilized to generate accurate and durable results independent of the training and test datasets.

3.2. Performance evaluation

The proposed method’s performance is assessed using the following assessment metric. The rate of correctly categorized images defines total accuracy as shown in (5).

\[
\text{Total Accuracy} = \frac{TP + TN}{(TP + TN + FP + FN)} \times 100 \% \tag{5}
\]
COVID-19 detection based on combined domain features (Omar Munthir Al Okashi)

3.2.1. Random forest

Table 1 shows that the RF classifier's accuracy for spatial, transform, and CDF is 83%, 70%, and 98%, respectively. Figure 7 plots the detection accuracy for the RF classifier (blue line). As a result, it's reasonable to suppose that the RF classifier outperforms the other.

3.2.1. Naive Bayesian

According to the experimental results, the accuracy of the NB classifier is 75%, 66%, and 85% for spatial, transform, and CDF, respectively. Figure 7 plots the detection accuracy for the NB classifier (red line). As a consequence, after the RF classifier, the NB classifier came in second.

As can be shown, CDF is the best domain feature to apply for detecting COVID-19. The explanation for this is that in all classifiers, the CDF feature surpassed all other domain features. In all of the classifiers, the spatial domain feature has the second-best domain feature performance. The Transform domain feature has a lower performance than the other domain features. When CDF features are included, the approach yields a greater detection accuracy (98%) as can be observed. Table 1 shows that the proposed method based on CDF can be successful in detecting COVID-19 while also enhancing detection accuracy.

Table 1. Performance evaluation result

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Spatial domain feature</th>
<th>Transform domain feature</th>
<th>Combined domains feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random forest</td>
<td>83</td>
<td>70</td>
<td>98</td>
</tr>
<tr>
<td>Naive Bayes</td>
<td>75</td>
<td>66</td>
<td>85</td>
</tr>
</tbody>
</table>

3.3. Comparison with existing brain CT images detection works

The performance comparison with other similar works [3], [9], [10], [25], [26] in the literature was undertaken to highlight the improved performance of the proposed method over current state-of-the-art in terms of accuracy. Table 2 and Figure 8 show the findings of the comparative evaluation. Our proposed combined domain feature-based RF classifier attained an accuracy of roughly 98%, which appears to surpass the majority of previous approaches, according to the comparative results.
4. CONCLUSION

This paper proposed a method to detect COVID-19 from the CT scan images using the combination of spatial domain and transform domain features. Using the lung segmentation step, the CT image is first processed and segmented, and then various domain features are extracted. From these domain features, the highest CDF are obtained. Finally, the detection task is completed using RF and NB classifiers. The proposed method is tested using a dataset of CT scan images, and the results are compared to several current techniques. The results show that our method outperforms previous methods, with an overall accuracy of nearly 98%. Pretrained CNN models will be employed in the future due to the scarcity of medical images.

REFERENCES


Table 2. Comparison results of existing techniques

<table>
<thead>
<tr>
<th>Techniques</th>
<th>Feature extraction</th>
<th>Image type</th>
<th>Classifier</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pereira et al. [25]</td>
<td>LBP, EQP, LDN,</td>
<td>CT-Scan</td>
<td>Predictive cluster trees (PCT)</td>
<td>89.0</td>
</tr>
<tr>
<td>Ameer and Mohammed [9]</td>
<td>LETRIEST, BSIF, LPQ</td>
<td></td>
<td>and generates a single decision tree (DT)</td>
<td></td>
</tr>
<tr>
<td>Imani [26]</td>
<td>GLCM</td>
<td>X-Ray</td>
<td>SVM and random forest</td>
<td>76/94</td>
</tr>
<tr>
<td>Wu et al. [10]</td>
<td>MP, Gabor filters</td>
<td>CT-Scan</td>
<td>Random forest</td>
<td>82.26</td>
</tr>
<tr>
<td>Santos and Melin [3]</td>
<td>Non subsampled</td>
<td>CT-Scan</td>
<td>Random forest</td>
<td>88.54</td>
</tr>
<tr>
<td>Proposed</td>
<td>Combined domain</td>
<td>X-Ray</td>
<td>Feedforward neural network.</td>
<td>98</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Techniques</th>
<th>Feature extraction</th>
<th>Image type</th>
<th>Classifier</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predictive cluster trees</td>
<td>and generates a</td>
<td></td>
<td>Random forest</td>
<td>82.26</td>
</tr>
<tr>
<td>Random decision tree</td>
<td>DT</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Random forest</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Feedforward neural network</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proposed</td>
<td></td>
<td></td>
<td></td>
<td>98</td>
</tr>
</tbody>
</table>

Figure 8. Comparison of detection accuracy of various existing methods
COVID-19 detection based on combined domain features (Omar Munthir Al Okashi)