INTERNATIONAL JOURNAL
FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

www.ijraset.com

Call: ◎08813907089 | E-mail ID: ijraset@gmail.com

Script Independent Document Analysis for Forensics

Rakesh Patel[1], Mili Patel[2], Kamlesh Tiwari[3]
[1,2]Kirodimal Institute of Technology Raigarh
[3]BITS Pilani

Abstract: The main objective of this research is to utilize ink analysis for the identification of forgeries in handwritten documents, particularly checks. The completion of this study involves the application of motif feature extraction techniques, including Peano scan motif, Directional local motif, and third-order recombination between RGB planes. Following Method 1, the Peano scan motif is employed.

To detect ink forgeries, 112 check images were examined, each written with fourteen different pens. A MATLAB function is applied to extract features for each of the mentioned approaches. The Weka application is utilized for both testing and training to validate the accuracy of the techniques.

Typically, motif approaches like Peano scan motif, Directional local motif, and third-order recombination are used for identifying comparable items in databases, such as content-based picture retrieval systems. What sets this project apart is its innovative use of ink level forgery detection. The recombination approach, involving modified color planes and the previously mentioned Peano scan motif method, introduces a novel aspect to the study.

I. INTRODUCTION

A. Problem Statement

We need to create a model that can identify different types of handwritten documents, such as checks, based on the degree of ink fraud. For every picture in the dataset, a collection of feature vector files utilizing motif features must be created in order for Weka software to be utilized for training and testing.

B. Methodology

The dataset incorporates the handwritten text of each check, denoted as the input picture. Following the conversion of the input picture into a binary format, the quest begins for square submatrices containing entirely black pixels in dimensions of 2x2 and 3x3. This process is fundamental as it addresses the challenge of identifying ink level forgeries in check photos, where there is a limited background.

To efficiently tackle the assignment involving overlapping subproblems and an optimal substructure, dynamic programming is employed. This strategic approach not only accomplishes the task but also significantly reduces the processing time.

In the context of a matrix consisting solely of 1s and 0s, the goal is to ascertain the largest square sub-matrix or the count of square sub-matrices. Dynamic programming is employed for this purpose, taking into consideration the values diagonally, to the left, and above the current index. The same principle is applied to identify square submatrices of sizes 2x2 and 3x3 in this scenario.

The feature extraction modules receive the located square sub-matrix as a parameter, and upon calculation of the appropriate features, they output a vector. This process is repeated for each identified square sub-matrix, and the resulting output is stored in the corresponding Excel file, where each row represents a single feature vector.

Every check is associated with two files containing screenshots of text written in different pens, generating two feature vector files named 1motif and 2motif. Subsequently, two additional vector feature files, named same and diff, are created. The absolute difference between feature vectors of images within the same folder is calculated. Using the diff function, the absolute difference is determined between images from separate directories, specifically one from folder 1 and the other from folder 2.

A pen association file contains information about the pens used in each check image. This information is utilized to create the primary feature vector file for training and testing.

Given that there are 14 pens associated with the check photos, 14 files are created, each corresponding to a specific pen. For each pen, checks unrelated to it are identified, relevant feature vectors are gathered from the same and different files, and then inserted into the pen-specific file. This process is repeated for each pen, and the search for a check continues

II. METHODS FOR FEATURE EXTRACTION

This study identifies ink level forgeries in handwritten documents, particularly checks, through the application of motif feature extraction algorithms. When presented with a query image, these characteristics are commonly employed to retrieve similar images from a database. Scan patterns, with variations such as horizontal, vertical, diagonal, and primary diagonal, represent the simplest form. The third method incorporates recombination and utilizes a scan pattern as a subroutine. Here is a brief overview of each approach:

1) Peano Scan Motifs Are Extracted From The Picture And Used For Ink Level Forgery Detection

This process focuses on non-overlapping 2x2 submatrices. The starting position is the top-left pixel of each 2x2 matrix. The subsequent point is identified by selecting the point with values closest to the current pixel values that has not been visited. Each location visit results in six distinct motif pattern types. In a specific case, all values in the matrix are identical, resulting in a total of seven instances.
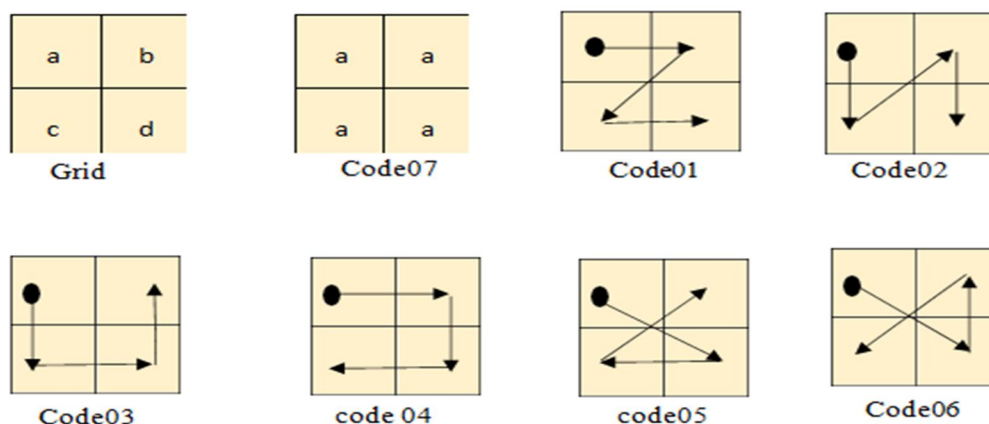

Figure 1: Peano scan pattern

The seven distinct pattern types that may be produced with the above-mentioned traversal approach are depicted in the figure above.
This approach will yield a total of 7x3=21 features as there are three RGB planes.

2) Detecting Ink Level Forgeries Using Characteristics Derived From Directional Local Motif

This approach takes into account 3x3 submatrices. This approach handles the four directions in each submatrix: the two diagonals, the vertical, and the horizontal, all of which pass through the matrix's center.
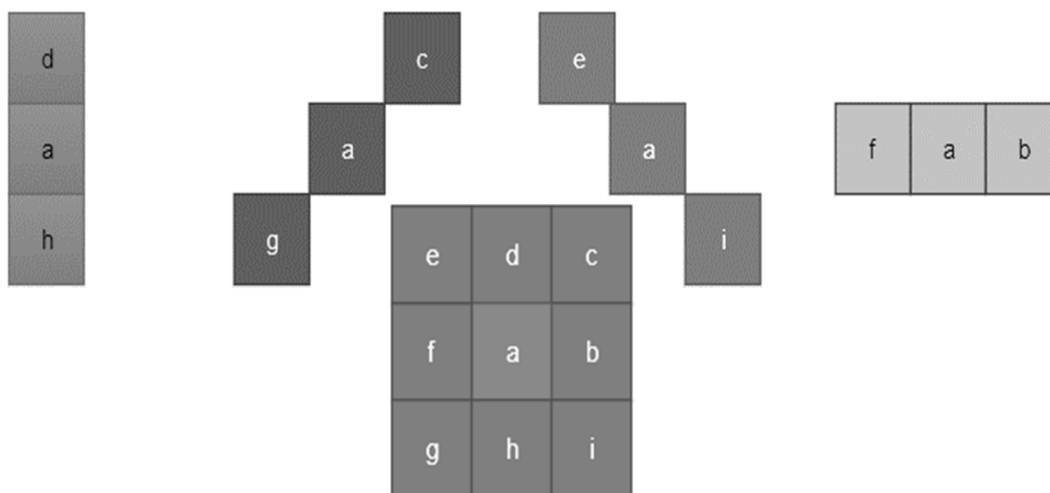

Figure 2: Directed patterns using Method 02

Similar to the preceding illustration with the center point, four directions are recognized: two diagonals, one vertical, and four horizontal.

As will be covered on the next page, there are seven possible feature kinds based on the values.
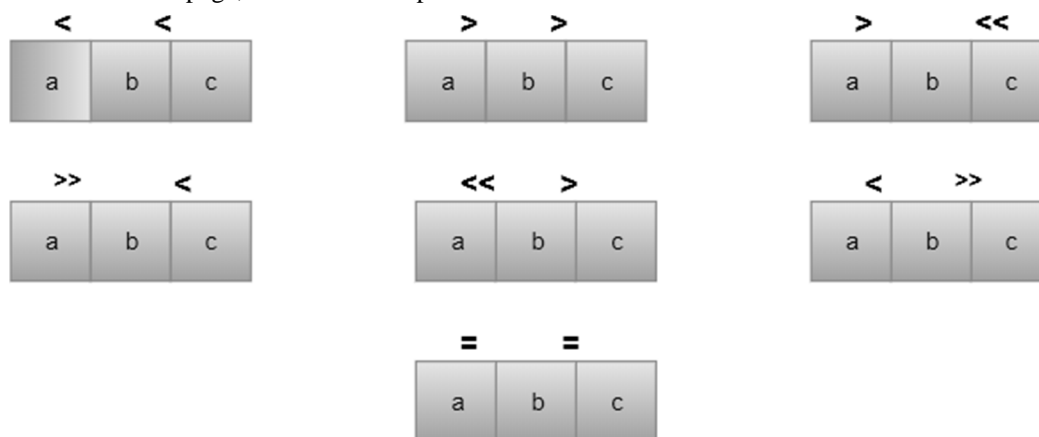

Figure 3: Patterns with zero degrees based on vector values

As seen in the above figure, there are seven possible outcomes based on the values of a, b, and c.

There are 7x4 = 28 features when looking at a single plane, and since there are three planes, the total is 28x3 = 84 features.

An index is read and the count is incremented whenever a specific kind of feature pattern is identified.

The variable count, which keeps track of the number of times this kind of 3x3 matrix has been found, is used to do normalization.

3) Recombination Between Rgb Planes And A Second Application Of Procedure 1 Are Used To Identify Ink Level Forgeries

This motif feature takes into account the connection between the red, green, and blue planes, in contrast to previous motif calculations. 2x2 submatrices are taken into account in RGB planes in this approach. After the planes have undergone recombination, procedure 1 is used.
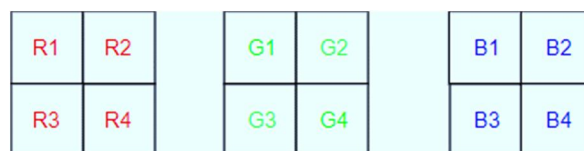

Figure 4: RGB planes of a selected area of the picture

It uses the top-left pixel value of each red, green, and blue matrix once to take into account the connection between the red, green, and blue planes.

For example, the top-left value of the red matrix will be chosen, and the top-left values of the green and blue matrices will be substituted.

Likewise, the top-left value of the green matrix will be chosen, and it will be substituted for the top-left value of the red and green matrix.

The top-left value of the red and green matrices will be substituted with the top-left value of the blue matrix after it has been selected.

As a result, nine distinct matrices in total will be produced using the top-left values of the red, green, and blue matrices. Since it employs technique 1, there are seven possibilities for each of them; therefore, a total of 9x7=63 characteristics are recognized.
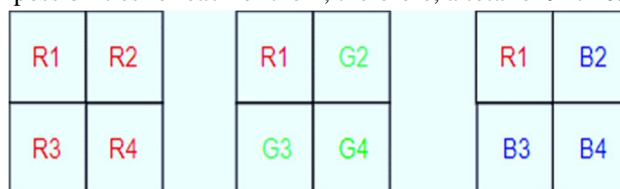

Figure 5: Red pixel replacement.

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue XI Nov 2023- Available at www.ijraset.com

As seen in the figure above, the top-left value of the green and blue matrix is substituted for the top-left value of the red pixel. Peano scan motif is estimated as before after substitution.
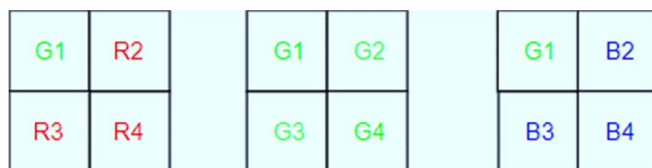

Figure 6: Green pixel replacement.

In a similar manner, the top-left value of green is utilized, changed, and the Peano motif is then calculated.
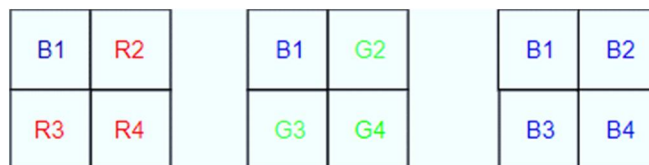

Figure 7: Blue pixel replacement.

The blue matrix is treated similarly, as was previously mentioned.

Up until now, as was previously indicated, this was utilized to get comparable photos. This study aims to identify forgeries at the ink level. Up to recently, methods for identifying ink level forgeries included RMSE and Euclidean distance. This technique is novel in that it uses motif traits to identify ink level forgeries in photos of checks.

a) Details Regarding the DataSet

There are 112 cheque folders in the dataset. Two sub-folders with processed pictures of checks are included in every cheque folder. Because of its link with two pens, it comprises two folders.

Example: Folder 1 (Cheque 83654)

Cheque 83654 folder 02:
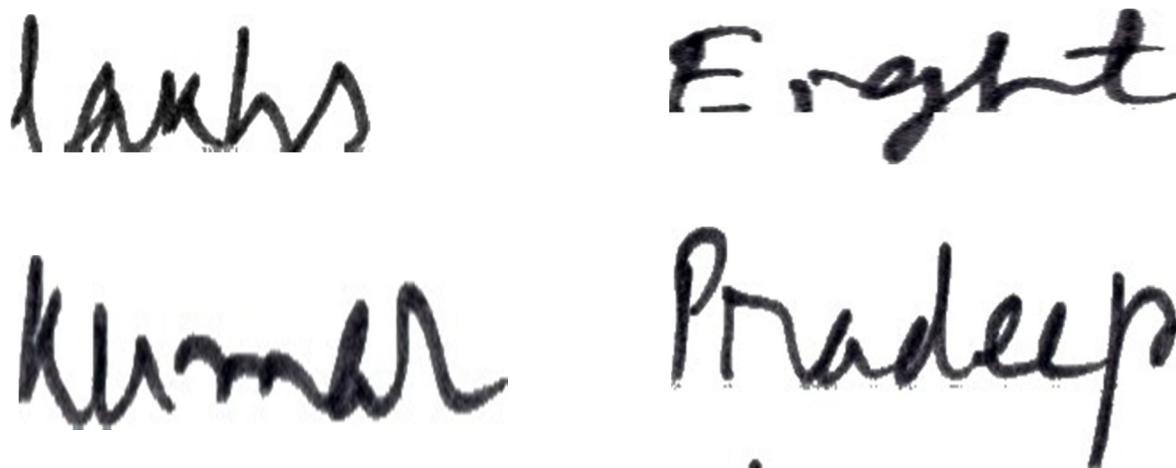
Figure 8: Example dataset for checks

Cheque 100830 folder 01:

Cheque 100830 Folder 02:

Figure 9: Dataset of sample checks

S.No.	1st Pen	2nd Pen	Documents No.
1	11	12	083654
2	12	13	083655
3	13	14	083656
4	14	8	083657
5	1	3	083658
6	2	4	083659
7	3	5	083660
8	3	4	0100828
9	4	5	0100829
10	5	6	0100830
11	6	7	0100831
12	7	1	0100832
13	8	9	0100833

14	9	10	0100834
15	10	11	0100835
16	4	6	0120611
17	5	7	0120612
18	6	1	0120613
19	7	2	0120614
20	8	10	0120615
21	9	11	0120616
22	10	12	0120617
23	11	13	0120618
24	12	14	0120619
25	13	8	0120620
26	13	9	0309061

Figure 10: Pen association matrix portion

The serial number appears in the first column. Details regarding the pen number attached to the check are provided in the second and third columns. The check number is in the fourth column.

Only pen associations for a total of 26 checks are displayed here. 112 checks and their corresponding pen associations total.

Matlab Code:-

The tasks of ink level forgery detection have been completed by writing the following routines:

- The Peano motif function code is:

```
function patternCode = findPattern(imageMatrix)
    % Input is a 2 x 2 matrix containing intensity values.
    % This function finds a pattern based on the following rule:
    % Starting from the top-left pixel, move in the direction where the
    % absolute difference between the intensities is the least.
    % This results in the formation of six patterns: 'Z', 'N', 'U', 'C',
    % 'gamma', and 'alpha', which are coded from 1 to 6 sequentially.
    % Code 7 is returned when all values are the same.
    % Check if all pixel intensities values are the same
    if all(imageMatrix(:) == imageMatrix(1))
     patternCode = 7;
     else
      % Initialize starting pixel
      currentPixel = imageMatrix(1, 1);
      visited = zeros(2, 2);
      % Initialize pattern code
      patternCode = 0;
      % Loop until all pixels are visited
      while ~all(visited(:))
       % Find unvisited neighbors
       neighbors = find(~visited);
       % Calculate absolute differences
       differences = abs(imageMatrix(neighbors) - currentPixel);
       % Find the index with the minimum difference
       [~, minIndex] = min(differences);
```

```
      % Update current pixel and mark as visited
   currentPixel = imageMatrix(neighbors(minIndex));
    visited(neighbors(minIndex)) = 1;
     % Increment pattern code
      patternCode = patternCode + 1;
    end
  end
end
```

- Compute Directional Motif:
  determines the kind of pattern by calling the getF function.

```
% Input is a 3 x 3 matrix used for feature computation. The matrix is a
% portion of an image that, when binarized, contains all black pixels.
% Get horizontal vector from input
horizontalVector = [input(2,1), input(2,2), input(2,3)];
% Get vertical vector
verticalVector = [input(1,2), input(2,2), input(3,2)];
% Get principal diagonal vector
principalDiagonalVector = [input(1,1), input(2,2), input(3,3)];
% Get other diagonal vector
otherDiagonalVector = [input(1,3), input(2,2), input(3,1)];
% Feature array stores the feature for respective directions
featureArray = zeros(1, 2);
% Checking the values and finding the patterns
featureArray(1) = getFeature(horizontalVector);
featureArray(2) = getFeature(verticalVector);
% Output the feature array
output = featureArray;
% Subroutine
function patternCode = getFeature(inputVector)
  % This function returns a code according to the values of the vector
  f = inputVector(1);
  a = inputVector(2);
  b = inputVector(3);

  if a >= f && a <= b
    patternCode = 1;
  elseif a <= f && a >= b
    patternCode = 2;
  elseif a <= f && a <= b && f <= b
    patternCode = 3;
  elseif a <= f && a <= b && b <= f
    patternCode = 4;
  elseif a >= f && a >= b && f <= b
    patternCode = 5;
  elseif a >= f && a >= b && b <= f
    patternCode = 6;
  elseif f == a && a == b
    patternCode = 7;
```

```
        end
end
```

Using the peano scan motif (method 1) and recombination between RGB planes:

Uses Peano motif as a subroutine to get motif patterns after replacement.

➢ *Input:* The three planes of the RGB matrix.

➢ *Output:* Generate a code for each of the 9 possible matrices using the Peano motif.

➢ *Procedure:*

❖ This approach involves recombination among the red (r), green (g), and blue (b) planes, all of which are 2x2 matrices.

❖ Begin by taking the top-left pixel of the red plane and use it to replace the top-left pixels of the green and blue planes. This operation results in three matrices where red is considered.

❖ Next, take the top-left pixel of the green plane and use it to replace the top-left pixels of the red and blue matrices. This process creates another set of three matrices where green is considered.

❖ Finally, take the top-left pixel of the blue plane and use it to replace every top-left pixel of the red and green matrices. Following this step, a total of 9 matrices are obtained.

➢ *Code Generation:*

❖ For each of the 9 matrices, apply the Peano function.

❖ Store the results in a 3x3 matrix where each value ranges from 1 to 7.

➢ *Output:*

❖ Return the 3x3 matrix as the output, representing the generated codes for each of the 9 matrices using the Peano motif.

Principal method for extracting every feature:

➢ *Input:* Images of checks from each folder.

➢ *Output:* Extraction of all features.

➢ *Algorithm:*

❖ Begin by taking the input image and storing its red, green, and blue planes.

❖ Binarize the input image.

❖ Traverse the image to identify 2x2 and 3x3 matrices where all pixels are black. Use the dynamic programming concept for calculating these matrices.

❖ Send the identified matrices to their respective functions for feature extraction.

❖ If a 2x2 matrix with all black pixels is found, call the Peano function and the recombination method. Store the obtained results.

❖ If a 3x3 matrix with all black pixels is found, call the directional motif function. Store the result.

❖ Normalize the stored results.

❖ End.

This algorithm processes check images from various folders, extracts features using different methods based on identified black pixel matrices, and normalizes the obtained results. The key steps involve binarization, dynamic programming for matrix calculation, and subsequent feature extraction based on matrix types.

The feature vector files 1motif.xlsx and 2motif.xlsx will be present in each of the check folders after the above mentioned function has been completed. To compute same.xlsx and diff.xlsx files, using the subtract.m function.

% Input: Two Excel files, 1motif.xlsx & 2motif.xlsx

% Output: Two Excel files, same.xlsx & diff.xlsx

% Initialize the feature vectors for same.xlsx and diff.xlsx

sameFeatureVectors = [];

diffFeatureVectors = [];

% Load feature vectors from 1motif.xlsx and 2motif.xlsx

featureVectors1 = readmatrix('1motif.xlsx');

featureVectors2 = readmatrix('2motif.xlsx');

% Compare feature vectors of images written with the same pen

```
for i = 1:size(featureVectors1, 1)
  for j = i+1:size(featureVectors1, 1)
     % Calculate absolute difference and store in same.xlsx
     diff = abs(featureVectors1(i, :) - featureVectors1(j, :));
     sameFeatureVectors = [sameFeatureVectors; diff];
  end
end
% Compare feature vectors of images written with different pens
for i = 1:size(featureVectors1, 1)
  for j = 1:size(featureVectors2, 1)
     % Calculate absolute difference and store in diff.xlsx
     diff = abs(featureVectors1(i, :) - featureVectors2(j, :));
     diffFeatureVectors = [diffFeatureVectors; diff];
  end
end


% Write the results to Excel files
writematrix(sameFeatureVectors, 'same.xlsx');
writematrix(diffFeatureVectors, 'diff.xlsx');


% Output: Return the paths to same.xlsx and diff.xlsx
output = {'same.xlsx', 'diff.xlsx'};
```

To obtain the final feature vector file taking into account each pen at a time, utilize another MATLAB function called getcomp.m. To determine whether or not the current check is connected to the current pen, using a pen association file or matrix. Initially, the final feature vector file corresponding to this pen is enhanced using the feature vector relating to same and diff. Finally, the feature vectors of every check connected to this pen are concatenated. And so it is with others. Since there are 14 pens, there will be a total of 14 feature vectors.

Depending on the technique being utilized, this function can be changed. There will be different columns utilized depending on the technique. When the three approaches presented are combined, there are 168 characteristics in all. Since there are seven times three equals twenty-one features in the first technique, the first twenty-one columns will be taken from the same file and added to the related pen's file. The characteristics for the second technique are 7x4x3= 84. Therefore, in this instance, columns 22 through 105 will be taken from the same and diff file. We have 9 x 7 = 63 characteristics for the last technique, which is the recombination method. Therefore, the same identical diff file will be used to extract columns 106 through 168.

```
    % Input: Pen Association matrix, same.xlsx, and diff.xlsx of each cheque
    % Output: 14 different CSV files

    % Load Pen Association matrix
    penAssociation = readmatrix('PenAssociationMatrix.csv');

    % Load feature vectors from same.xlsx and diff.xlsx
    sameFeatureVectors = readmatrix('same.xlsx');
    diffFeatureVectors = readmatrix('diff.xlsx');

    % Initialize cell array to store CSV file names
    csvFiles = cell(14, 1);

    % Iterate through each pen
    for pen = 1:14
```

```
    % Initialize feature vectors for this pen
    penFeatureVectors = [];

    % For all cheque images
    for cheque = 1:size(penAssociation, 1)
        % Check if this pen is associated with this cheque
        if penAssociation(cheque, pen) == 1
            % Grab feature vector associated with this method from same.xlsx and diff.xlsx
            featureVector = [sameFeatureVectors(cheque, :), diffFeatureVectors(cheque, :)];

            % Store vector in the file corresponding to this pen
            penFeatureVectors = [penFeatureVectors; featureVector];
        end
    end

    % Write the feature vectors to a CSV file for this pen
    csvFileName = ['Pen', num2str(pen), '_FeatureVectors.csv'];
    writematrix(penFeatureVectors, csvFileName);

    % Store the CSV file name
    csvFiles{pen} = csvFileName;
end

% Output: Return the names of the 14 CSV files
output = csvFiles;
```

Following the following function's execution, we will have 14 files for every method that may be utilized to the Weka program for testing and training.

III. CONCLUSION

Feature extraction is a crucial task in image processing, especially in recognizing elements like digits. Motif features, commonly employed in content-based image retrieval systems, have not been previously utilized for ink level forgery detection in cheque images. The integration of feature extraction methods, namely Peano scan motif, directional motif, and recombination atop method 1, has proven to be highly successful in detecting ink level forgery, achieving a maximum accuracy of approximately 88% using method 3.

Peano motif extraction yields 7 types of features for each of the 3 planes (R, G, B), resulting in a total of 21 features. Directional motif extraction provides 28 features for each plane (four directions), summing up to 84 features. The recombination method, layered over method 1, produces 7 features for each of the 9 possible matrices, totaling 63 features.

Comparative analysis reveals that method 1, relying solely on Peano motif, exhibits lower performance with accuracies around 70%. Method 2 performs better, achieving accuracies around 80%. Method 3 surpasses both, leveraging the relationship between RGB planes to generate 9 matrices. Within each of these matrices, method 1 (Peano scan) extracts 7 features, resulting in a total of 63 features. This comprehensive approach yields accuracies ranging between 85-88%, establishing method 3 as the most effective among the three.

REFERENCES

[1] Jhanwar, N., Chaudhuri, S., Seetharaman, G., & Zavidovique, B. (2004). Content based image retrieval using motif cooccurrence matrix. Image and Vision Computing, 22(14), 1211-1220.

[2] Vipparthi, Santosh Kumar, and S. K. Nagar. "Expert image retrieval system using directional local motif XoR patterns." Expert Systems with Applications 41.17 (2014):8016-8026.

[3] Subrahmanyam, M., Wu, Q. J., Maheshwari, R. P., & Balasubramanian, R. (2013).Modified color motif co-occurrence matrix for image indexing and retrieval. Computers& Electrical Engineering, 39(3), 762-774.

[4] Pass, G., Zabih, R., & Miller, J. (1996, November). Comparing Images Using ColorCoherence Vectors. In ACM multimedia(Vol. 96, pp. 65-73).

[5] Chang, P., & Krumm, J. (1999). Object recognition with color cooccurrence histograms. In Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149) (Vol. 2, pp. 498-504). IEEE.

[6] Y. Rui, T.S. Huang, S. Chang, Image retrieval: current techniques, promising directions and open issues, Journal of Visual Communication and Image Representation 10 (1999) 39–6

INTERNATIONAL JOURNAL
FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089   (24*7 Support on Whatsapp)