Detecting Duplicate Regions in Digital Images Using Improved Localization Method

A Dissertation submitted in partial fulfilment of the requirement for the

Award of degree of

MASTER OF TECHNOLOGY

IN

INFORMATION SYSTEM

Submitted by

AKSHAT KUMAR

(2K12/ISY/02)

Under the esteemed guidance of

RITU AGARWAL

(Asst. Professor, IT Department)

Department of Information Technology

Delhi Technological University

Bawana Road, Delhi – 110042

(2012-2014)

CERTIFICATE

This is to certify that **Akshat Kumar (2K12/ISY/02)** has carried out the major project titled **"Detecting Duplicate Regions in Digital Images Using Improved Localization Method"** for the award of Master of Technology degree in Information System by Delhi Technological University.

The major project is a bonafide piece of work carried out and completed under my supervision and guidance during the academic session **2012-2014**. The matter contained in this report has not been submitted elsewhere for the award of any other degree.

(Project Guide)

Ritu Agarwal

Assistant Professor

Department of Information Technology

Delhi Technological University

Bawana Road, Delhi-110042

ACKNOWLEDGEMENT

I express my sincere thanks and deep sense of gratitude to my project guide**, Ritu Agarwal**, Assistant Professor, Department of Information Technology, Delhi Technological University, for her valuable motivation and guidance, without which this study would not have been possible. I consider myself fortunate for having the opportunity to learn and work under her supervision and guidance over the entire period of association.

I humbly extend my words of gratitude to other faculty members of this department for providing their valuable help and time whenever it was required.

Akshat Kumar
Roll No. 2K12/ISY/02
M.Tech. (Information System)
E-mail: akshatrathore@live.com

ABSTRACT

Now days, the uses of digital images has increased and so has the forgery mechanisms. Copy-Move forgery is one of the types of image forgery. In the case of copy-move forgery one small region of image is replaced with another region within the same image. Many algorithms already developed for detecting copy-move regions but the major problem for most of the algorithms are localization of duplicate regions within same image after detecting copy-move part. Here we present an efficient approach for detecting duplicate regions in a digital image. In our method, we have used PCA (Principal Component Analysis) to the input image for obtaining compressed image. For obtaining feature vector of each block, we divide the image into number of overlapping blocks. These feature vectors are calculated with the help of principal component analysis. Here we represent eigenvector as feature vector. Lexicographic sorting is henceforth applied to locate the similar feature vector of each block. If two or more feature vectors have same value then this indicates that image forgery has been done. But we cannot guarantee of image forgery because in some cases, feature vectors may be same in an image. Hence, we calculate total number of connected blocks with in the same distance. If the numbers of blocks are greater than a threshold value then image has been forged by duplicate regions, but if numbers of blocks are less than a threshold value then image has not been forged. After that, localization method is used for getting number of pixels that have been copied and pasted. With the help of our proposed method, duplicate regions in an image can be detected more accurately and with less computational complexity as compared to other methods.

LIST OF FIGURES

Figure No.	Title	Page No.
1.1	An example of Copy-Paste Forgery	3
3.1	Copy Move Forgery Detection Classification	16
4.1	Block diagram of proposed algorithm	25
5.1	Normal image without noise	33
5.2	Tampered image of cricket and forest	34
5.3	Detection Result of Tampered image of cricket and forest	34
5.4	Tampered image with Gaussian noise	35
5.5	Detection result of tampered image with Gaussian noise	35
5.6	Tampered image with salt and pepper noise	36
5.7	Detection result of tampered image with salt and pepper noise	36

Table of Contents

CERTIFICATE…… (i)
ACKNOWLEDGEMENT…….. (ii)
ABSTRACT…….. (iii)
LIST OF FIGURES……..… (iv)

CHAPTER 1. INTRODUCTION……..……..1
1.1 Introduction.....1
1.2 Image forgery.....2
1.3 Classification of Forgery Detection Techniques 3
1.3.1 Active Approach 4
1.3.2 Passive Approach.....4
1.4 Feature Extraction Techniques 5
1.4.1 Principal Component Analysis (PCA).....6
1.4.2 Discrete Cosine Transform (DCT).....6
1.4.3 Discrete Wavelet Transform (DWT) 6
1.4.4 Fourier Mellin Transform (FMT) 6
CHAPTER 2. PRINCIPAL COMPONENT ANALYSIS……8
2.1 Introduction.....8
2.2 Mathematical tools used in PCA 8
2.2.1 Standard Deviation 9
2.2.2 Variance 10
2.2.3 Covariance 10
2.2.4 Eigenvectors.....11
2.2.5 Eigenvalues.....12
2.3 PCA Algorithm.....12
2.4 Application of PCA 15

CHAPTER 3. LITERATURE REVIEW……..……..16

3.1 Block Based Method.....16

3.1.1 Moment-based (BLUR, HU, and ZERNIKE) 16

3.1.2 Dimensionality reduction-based (PCA, SVD, KPCA, and PCA-EVD) 18

3.1.3 Intensity-based (LUO, BRAVO, LIN, CIRCLE, and PCMIFD).....19

3.1.4 Frequency-based (DCT, DWT, FMT, PHT, DyWT, QCD, LBP, and Curvelet) .21

3.2 Keypoint-based methods.....23

CHAPTER 4. PROPOSED METHOD……25

4.1 Divide Image into Overlapping Blocks 26

4.2 Feature Extraction through PCA.....26

4.3 Lexicographical Sorting on Feature Vector.....27

4.4 Detecting duplicate connected blocks.....27

4.5 Localization of copy-move region.....28

4.6 Algorithm Steps 29

CHAPTER 5. EXPERIMENTAL RESULTS…….31

5.1 Selection of Parameter values.....31

5.2 Detection of Copy-Move Forgery.....33

CHAPTER 6. CONCLUSION AND FUTURE WORK…….38

6.1 Conclusion……38

6.2 Future Work.....38

REFERENCES…….40

Chapter 1
Introduction

1.1 Introduction

Images and videos have become the main information carriers in digital media, now days. There are numbers of powerful tools available today which can be used to modify the images, even by non-professionals. To maintain the authenticity and integrity of digital media, many research studies are going on to improve image forgery detection. In digital imaging both the acquisition and the tampering process techniques are likely to leave some marks. Hence the task of forensic scientists or experts lies in detecting these key points by exploiting existing knowledge on digital image mechanism. Today detecting image forgery is an emerging field of research and no proper mechanism exists for this which tells accurately and in less time that an image is tampered or not and if an image is tampered then how to locate the duplicate regions in the image.

Copy-Move forgery is one of the types of image forgery. In the case of copy-move forgery one small region of image is replaced with another region within the same image. Such type of image forgery can be used to hide an undesired object in the image. The main characteristics of this type of image forgery is that, since the region is copied and pasted in the same image itself, so texture, colour pattern and the noise are compatible with the rest of the image. This makes it hard to detect the forgery.

The main aim of digital image forensic is to provide tools or soft wares that are used to detect image forgery. This new stream of image forgery is related form existing multimedia security related research domain such as steganography and watermarking. In digital watermarking [1] we hide a keypoint in the image for protecting its copyright. Steganography [2] includes in communication secretly via some media. Steganography is an art of concealing an image, massage, or file within another image, massage or file.

In the recent time large amount of image manipulation is seen in the field of magazine publishing, scientific journals, fashion industry and court rooms etc. There are many types of image forgery [3] which can be performed to remove the authenticity and integrity of digital images. So it is important to develop a method or algorithm for detecting these forgeries in such a way that it gives efficient result with less computational complexity.

1.2 Image forgery

Forgery is a method in which one can modify or delete the image data with the intent others or to take advantage. So we can say that forgery deals with altered or modified objects. The result of forgery is to create an artificial document, data or image that loses its confidentiality and authenticity. Following are the few types of forgery techniques that exist nowadays.

- Fluent Forgery- In fluent forgery, we do some minor changes in data or image. For example in case of letter or word when we change the direction of movement in letter, spacing between letter, size of letter and the relative location of letter parts. This type of forgery is called fluent forgery.
- Copied Forgery- In copied forgery, one can copy some parts or features from original content or object and uses that features in his or her work.
- Photomontage forgery- Photomontage forgery is a type of copied forgery. This type of forgery can be done in many ways and it is perfectly design so that original signature and copied part is transplanted in another document. Only microscopic examination can reveal that forgery is done or not.
- Self-Forgery- In self-forgery, one can forge one's own signature or feature in order to deny it at a later stage. In this case of forgery, a person is unable to free himself from his own handwriting which exposes him.

Now days, Image forgery is very common with the evolution of powerful tools that are used for forgery. In image forgery, a person can edit or modify the image for taking some advantage etc.. So if the image is modified then it loses its confidentiality and authenticity.

Many types of image forgeries [4] are possible but one of the most common types of image forgery is copy-move forgery. It is one of the simplest approaches for manipulating a digital image because there are various types of software available for doing so. In this type of image forgery, a region from one part of an image is copied and pasted into another part of the same image. Such type of image forgery can be used to hide an undesired object in the image as in Figure 1.1. Image tampering, cloning or splicing are done to create a forged image. Digital forged image sometimes looks so real that one cannot distinguish it from original image. So by doing this integrity, authenticity of an image is lost. Several authors have proposed different approaches for detecting copy-move forgery. One direct method for detecting this type of forgery is exhaustive search techniques but it is infeasible because of the process complexity.

(a) (b)

Figure 1.1: An example of copy-paste forgery: (a) the normal image and (b) the tampered image

For detecting copy-paste forgery, we first divide an image into overlapping blocks. Then we find feature vectors of every blocks. There are many algorithm exists from which we obtain feature vector of blocks such as PCA, DCT, and DWT etc. By comparing feature vector of every block, we decide that forgery has been done or not. If image is forged then next step is localization. Localization is the major problem of detecting copy-move forgery. In which we have to decide which pixels are copied and pasted. In Localization we take maximum possible pixels for obtaining efficient answer from tampered image. Localization algorithm should be very effective so that we can obtain better result with minimum computational complexity.

In digital forged image, tampering process leaves some specific key points. The aim of forensic expert to detect those key or traces for detecting corrupted image.

1.3 Classification of Forgery Detection Techniques

Now days, image forgery detection is very hot area of research and a large number of algorithms exist for copy move detection. Exhaustive search technique [39] is one of the simple techniques for detecting copy-paste forgery wherein the original image and its shifted version are overlaid looking for pixels that is copied and pasted. This method is simple and very effective for small sized image but the computational complexity and time is more. Even this method is impractical for medium and large size of image.

Based on various types of algorithm for copy move forgery, researcher has classified these algorithms in two approaches that are active and passive.

1.3.1 Active Approach

Active approach was very popular method at an early age for detecting copy move forgery. Active approach has some limitation due to this most of researcher does not use this method. In this approach, at the time of creating the image, digital image requires some pre-processing steps such as to add watermarking [1] and digital signature for the security of image. So if forgery has done in the image then main task of forensic expert to check watermarking and signature of an image for authentication and integrity of an image.

But now days there are millions of image exist in the internet that have no digital signature and watermarking. So in such case active approach will not work for detecting copy move forgery or ensuring authentication and integrity of an Image. So because of this limitation some researchers do not use this approach.

1.3.2 Passive Approach

There is no need of watermarking and digital signature in case of passive approach for detecting copy move forgery unlike active approach. Passive approach [5] does not need any previous information about the image. Here the main task of forensic experts to detect specific change that forgery bring into an image.

There are three methods that mostly used for manipulation of digital image- tampering, splicing and cloning.

- Tampering is a method through which one can manipulate the image to achieve the specific result. In this method we can add or delete some specific objects for getting advantage.
- Splicing, it is also known as composition in which the digital splicing of two or more images result a single composite image. We can say that we create a single image by combining two or more images
- Cloning is also known as copy-paste forgery through which one region of an image has been copied and pasted with in the same image. It is called copy move forgery.

Some researcher has classified the detection techniques in another two categories that are feature based techniques and block based technique.

Block based methods [6] require dimension reduction or compression. There are many algorithm exist for this like principal component analysis (PCA) etc. It assumes that copied

part of an image has not change in any post processing. When we compress the image then there must be loss of data. In this technique we divide the image into number of block. These blocks are equal in size. This technique fails when there is a case of rotation, scaling and resizing. So it does not deal with geometric transformation of copied region.

Feature based techniques [6] deals with geometric transformation such as rotation, scaling and resizing. Here we consider on features of an image rather than blocks so we match features instead of blocks. In block based approach we match every pixel of each block to every pixel of another block. But here we take feature of each block and match those features. These features do not change with respect to rotation and translation. Scale invariant feature transform (SIFT) is popular technique that is based on this approach.

1.4 Feature Extraction Techniques

Feature Extraction is an important step in image forgery detection. When we divide the image into overlapping or non-overlapping blocks then we find features of every block. So feature extraction technique must be more effective and less computational complex.

We divide the features into two specific category general features and domain specific features.

- General features are those features in an image that do not depend on application such as colour, texture and shape. That is again divided into several categories.
  Pixel-level Features- These features are examined at every pixel such as colour and location etc.
  Local Features- Features that are obtained as a local level in an image or the results of sub-division of an image.
  Global Features- these are those features that are examined for the whole area of image.
- Domain specific features are those features that are application dependent such as fingerprints.

There are number of feature extraction techniques available now days that can be used for copy move forgery detection. When we perform the analysis of large and complex data then major problem include number of variable that is used. So analyses of data that have large number of variables require large amount of memory and more computation.

There are following important algorithms that generally, we used in feature extraction.

1.4.1 Principal Component Analysis (PCA)

PCA [7] is a well-known feature extraction algorithm. It is also used for dimension reduction. Principal Component Analysis is a general method for obtaining similar pattern in image and highlighting differences. In this approach, we deals with mean, standard deviation, variance, co-variance, eigenvectors and eigenvalues.

We represent the image in the form of matrix. So after calculation covariance of the matrix, we have to find out eigenvectors and eigenvalues of matrix. After that we sort the eigenvalue and take minimum eigenvalue. We remove those Eigenvectors that correspond to minimum eigenvalues as these contain less information so we don't lose much information about data.

1.4.2 Discrete Cosine Transform (DCT)

DCT [8] calculates continues data set in form of cosine function that is changing at different frequencies. DCT used in number of areas of science. When we use cosine instead of sine function then it is hard for reduction of dimensionality, because some cosine functions are used for approximation of unusual signal, while for differential equations, it needed as specific value of boundary condition.

Discrete Cosine Transform is a Fourier transform that is similar to the DFT. DCT are similar to DFT according to operation on real data with even symmetry.

1.4.3 Discrete Wavelet Transform (DWT)

DWT [9] [10] [15] is used in those wavelet transform for which the wavelets are sampled discretely in functional analysis and numerical analysis. The key advantage of discrete wavelet transform is that it has Fourier transforms in the form of temporal resolution so it gives both location information (location in time) and frequency.

We can use DWT in data compression and in signal coding for representing a discrete signal in reduced form.

1.4.4 Fourier Mellin Transform (FMT)

Fourier Mellin Transform [11] is used for reconstruction, pattern recognition and image data retrieval. It mainly deals with to find out similarity in large data. Because of its scale invariance property, it is mainly used in computer science for the analysis of algorithms. Of a

scaled function, the magnitude of the Mellin Transform is same as the magnitude of the original one. This property of FMT is analogous to Fourier Transform's shift invariance property.

This property of Mellin transform is very useful in image recognition. An image of an object is easily scaled by this property when the object is moved away or towards from the camera.

Chapter 2
Principal Component Analysis

2.1 Introduction

PCA [7] is a statistical technique that used in face recognition, image compression and feature extraction [14] etc. it is a powerful tool for analysing large set of data. It is used to find out the similar patterns and differences in high dimension data. Another important advantage of principal component analysis is that we can compress an image or the data set by removing dimensions of data set. So we apply PCA in large set of database for reducing numerical computation. Principal Component Analysis [12] [13] uses an orthogonal transformation that is used to convert a set of correlated variables into a set of linearly uncorrelated variables that is known as principal components. The numbers of original variables are greater than or same as the number of principal components.

Principal Component Analysis is a well-known pattern recognition technique. It is also known as Hotelling transform in multivariate quality control, spectral decomposition in noise and vibration, Karhunen–Loeve transform (KLT) in signal processing and proper orthogonal decomposition (POD) in mechanical engineering.

This method is mostly used as a tool for making predictive models and in exploratory data analysis. The result of a Principal Component Analysis is mainly described as a component scores or factor scores. Factor analysis usually deals with eigenvector. Eigenvector is also used for dimension reduction. There is particular eigenvector for a specific eigenvalue. We can remove that eigenvector that is corresponds to minimum eigenvalue. Because that eigenvector contain less information as compared to others. By removing eigenvector we can reduce the dimensions of dataset.

2.2 Mathematical tools used in PCA

Principal Component Analysis uses lots of mathematical terms, before going to further discussion we have to understand all mathematical terms and results of that tools. Usually PCA is used for to find out the pattern recognition, feature extraction and dimension reducibility. In case of image that is represented in the form of matrix, for this we have to understand about matrix algebra that includes standard deviation, variance, covariance,

eigenvector and eigenvalue. When we apply principal component analysis in any algorithm or application then we have to use these mathematical terms and we must what these terms specifies and what would be the output when we apply these functions in our data set.

2.2.1 Standard Deviation

Standard deviation is a measure of how data spread out in two or more sets. In statistics, the standard deviation is represented by the Greek letter sigma, σ that measures the amount of variation or dispersion in a large set of data. A low standard deviation specifies that the values in the data sets are similar to the mean of data set and high standard deviation specifies that the values in the data sets are dispersing in a large range. The standard deviation of a random variables is the square root of its variance. The main advantage of SD, unlike the variance, standard deviation is represented as the same unit as we represent the data. If the measurement of data is percentage than the standard deviation will also represent as percentage.

Definition of standard deviation indicates that it is the average distance from the mean of the data set to a point.

$$s = \sqrt{\frac{\sum_{i=1}^{n}(Xi-\overline{X})2}{n-1}} \quad (1)$$

Here, **s** is the symbol of standard deviation. $\overline{X}$ is the mean of data set and n is indicating number of values in the data set. At first we calculate the squares of the distance form every data point to the mean of the data set, then we apply the summation on squares of the distances and after that we divide by n-1.

If the data set contains equal value then its standard deviation will be equal to zero. Because all the values in the data set are same. So there is no dispersion in the data set. If a data set in which difference in the value is high then its standard deviation value will also high as compared to data set that contain values with small differences. The value of standard deviation may be positive or negative because here we calculate the square root of a value.

$\overline{X}$ indicates the mean of the data set that is also known as the average of data set. Mean is calculated by taking sum of all value in the data set and after that divide the result by number of values in the data set.

$$\tilde{X} = \frac{\sum_{i=1}^{n} Xi}{n} \quad (2)$$

Standard deviation of a set of values indicates how much variation there is from the average or mean. There are number of area where standard deviation is used such as weather forecasting, in financial calculation, geometric interpretation and in number of areas and applications.

2.2.2 Variance

Variance is also similar to standard deviation. It is also the measurement of the dispersion of data in a data set. Variance is always non negative number because square of a number is always positive but standard deviation may be negative because in this we calculate square root of a number. Variance is calculated by below given formula.

$$s2 = \frac{\sum_{i=1}^{n}(Xi-\overline{X})2}{(n-1)} \quad (3)$$

S^2 is the symbol representation for variance. $\overline{X}$ is mean of data set and n is total number of values in the data set. If a constant is added to all values of the data set, then variance will not change.

2.2.3 Covariance

Standard deviation and variance are purely 1-dimensional measurement of data set. If there is any relationship exist between two or more dimension such as if in a class room, data set of student contain height as one dimension and mark as another dimension and we have to identify that there is any effect of height of the student on their percentage. Then we cannot analyse this on the basis standard deviation because it operates only on 1 dimension. So standard deviation and variance of a dimension is independent to another dimension. But sometimes we have to analyse that how much vary one dimension of a data set to another dimension of the same data set.

Covariance is such type of measure that finds out that relationship between two dimensions of data set. The formula of covariance is written as-

$$cov(X,Y) = \frac{\sum_{i=1}^{n}(Xi-\overline{X})(Yi-\overline{Y})}{n-1} \quad (4)$$

If we examine the covariance between one dimension and itself then it will give variance of that dimension as shown in figure

$$var(X) = \frac{\sum_{i=1}^{n}(Xi-\overline{X})(Xi-\overline{X})}{n-1} \quad (5)$$

If we have to calculate the covariance between 3-dimension such as (a, b, c) then we measure the covariance between a and b dimension, a and c dimension and b and c dimension. For an example if we have to measure the covariance of a data set that contain two dimension, in which one is h that is hours studied by student and second is m that is marks obtain by student. Now we have to measure that there is any dependency exist between hour dimension and marks dimension. For this value of covariance is not as important as sign of covariance. If the value is positive then it indicates that both values will increase together hence marks of student will increase as the number of hours of studied increased. If the sign of covariance is negative then that sign indicate, if value of one dimension will increase then another dimension will decrease. When covariance value is zero then it indicates that both dimensions are independent to each other.

With the help of covariance value of data set, we obtain the covariance matrix. If data set contain n dimension then covariance matrix contain n rows and n columns and each entry in covariance matrix indicate covariance value between two dimensions of data set.

For an example if we have to write the covariance matrix of 3 dimensional data (x, y, z) set then it contain 3 rows and columns and the entries in the covariance matrix are like:

$$c = \begin{pmatrix} cov(x,x) & cov(x,y) & cov(x,z) \\ cov(y,x) & cov(y,y) & cov(y,z) \\ cov(z,x) & cov(z,y) & cov(z,z) \end{pmatrix} \quad (6)$$

From the above written matrix we can say that the matrix is symmetrical about the diagonal because the value of cov(a, b)=cov(b, a).

2.2.4 Eigenvectors

Eigenvector is a special set of vector of matrix that can be in only square matrix and not every square matrix has eigenvector. Suppose if the size of the matrix is n * n then it has n eigenvectors. For an example, when we multiply a matrix M with a vector V and the resultant

vector R is exactly λ times of vector v then vector v is one of the eigenvector of matrix m. It is property of transformation matrix due to which eigenvector arises.

$$M * V = \lambda R \quad (7)$$

The eigenvectors of the matrix are perpendicular to each other i.e. they make right angle to each other or we can say they are orthogonal to each other. With the help of length of the vector, we cannot say, it is eigenvector or not so the length does affect a vector but direction of a vector does.

In some cases eigenvector is also called feature vector that specify the feature of the block. In PCA eigenvector is used as features of data or image and it also use in image compression. Eigenvector that corresponds to minimum eigenvalue contain less information as compared to other eigenvector so by removing this vector, we loss less information that will not give much effect on image.

There are many applications of eigenvector such as it is used in physics for stability analysis and in the physics of rotating bodies, in market analysis, in principal component analysis for image recognition, and in PageRank from google.

2.2.5 Eigenvalues

From above given equation 7 we know vector V is equal to vector R that is λ times of vector V. That λ is called as eigenvalue of that matrix M. Eigenvalue and eigenvector always comes in pairs. For each eigenvalue there is associated eigenvector.

Eigenvalues are also used in factor analyses that are used for choosing to obtain number of factor in overall factor analysis. Factor analysis is also used for verifying and exploring patterns in a set of correlation coefficients.

2.3 PCA Algorithm

Principal Component Analysis (PCA) is a method from which we find the similarity and differences in a large set of data and for pattern recognition. Once we find out the pattern in the data set then we can compress the data. It is a very important tool for analysing data. PCA algorithm includes number of steps that are following.

Step 1: Get some data

At first we have some data in which we apply PCA algorithm. Data may be one dimensional or contain more than one dimensional data set. Image is represented in the form of matrix. In image analysis we have matrix as data in which we apply Principal Component Analysis. The entry of the matrix represents the intensity value of each pixel. So in image we have to find out the similarity and differences of feature vector with the help of eigenvector or we can also apply PCA for image compression in an image.

Step 2: Subtract the mean

After getting data in which we will apply PCA algorithm, first we calculate the mean of every data dimension. The mean is average value of every dimension. After that we subtract data of each dimension from its mean value. So if a data set that have two dimension x and y then they have mean value of each dimension $\overline{X}$ and $\overline{Y}$ respectively. Data of x dimension is subtracted from its mean that is $\overline{X}$ and data of y dimension is subtracted from$\overline{Y}$.

Step 3: Calculate the Covariance matrix

As we have discussed earlier, if data set that have n dimension then the size of covariance matrix will be n * n. The entries of covariance matrix contain covariance value of each dimension. Covariance matrix is symmetrical along its diagonal.

Step 4: Calculate the eigenvector and eigenvalue of covariance matrix

As we know that covariance matrix is a square matrix, we calculate eigenvector and eigenvalue from this matrix. These eigenvector and eigenvalue give important information about the matrix or data set. They provide information about the pattern of the data. Eigenvector corresponds to minimum eigenvalue contain less information. These eigenvector also represents as feature vector.

Step 5: Choosing component and forming a feature vector

After getting eigenvector and eigenvalue of data set we choose which eigenvector we have to ignore and which eigenvector is principal component of data set. With the help of this, we compress the data and remove the dimension of data set for reducing numerical complexity. Eigenvector corresponds to highest eigenvalue is called as principal component of the data set.

Once we find the eigenvector of data set next step is to sort the eigenvector. This will give us the component that can be used in various applications. With the help of this we remove the component of data set of minimum importance. By doing this we may lose some information but it does not give much effect on result. But we lose corresponding dimension of data set. So finale data set have less dimensioned then original data set. If we have n dimension in data set then we calculate n eigenvector and eigenvalue then we choose only q eigenvector for further processing, then final data set has only q dimension.

Step 6: Deriving the new data set

It is the last step in Principal component analysis. Once we decide which component will remain in the data set and form feature vector and which component we have to ignore, then we take the transpose of the feature vector and then multiply it.

$$Final\ Data = Row\ Feature\ Vector * Row\ Data\ Adjust \quad (8)$$

Row Feature Vector is the matrix of eigenvectors in the columns so eigenvector are in the rows of the matrix with the highest significant eigenvector at the top. Row Adjust Data is the mean data that is in the columns of the matrix, with every row have a separate dimension.

Now when we want our original data back from the new set of data then we have to do exactly reverse process of the method from which we are getting final data set.

$$Row\ Adjust\ Data = Row\ Feature\ Vector^{-1} * Final\ Data \quad (9)$$

Row Feature Vector^{-1} is the inverse of Row Feature Vector. When we consider all eigenvector in our feature vector then the inverse of feature vector is same as the transpose of the feature vector. So above written equation is same as following:

$$Row\ Adjust\ Data = Row\ Feature\ Vector^{T} * Final\ Data \quad (10)$$

Now for getting original data we have to add the mean of the original data into this equation. So the equation will be:

$$Row\ Adjust\ Data = Row\ Feature\ Vector^{-1} * Final\ Data + Original\ Mean \quad (11)$$

We can also apply this formula when we do not consider all eigenvector in our feature vector. It will give exact transform of final data into original data set.

2.4 Application of PCA

Principal Component Analysis is used in number of application such as image recognition, pattern analysis and image compression etc. But there are also other application rather than image analysis in which we apply PCA algorithm such as neuroscience, finance, biomedical, to analyse change detection in SAR images, pharmacy, health, architecture, agriculture and taxonomy etc.

Principal Component Analysis is used in neuroscience for identifying the particular behaviour of stimulus (a thing that evokes a specific functional reaction in an organ or tissues) that starts the neuron's activity.

Principal Components Analysis is also used in data mining. In data mining we have to mine the data in a large set database or recognize relevant information. In large amount of data, we have to identify pattern or similarity and differences in data. In data mining, PCA classifies similar data in one cluster.

Principal Component Analysis is also used in weather forecasting. Principal component analysis (PCA) is a standard tool for descriptive analysis that is used for forecasting. Usually, summarizing a large set of data with number of factors may result in missing information. Therefore, factor-based forecasting with principal component (PC) estimation can give better result to extract a small number of latent factors which could contain the most information of data set and improve the prediction accuracy for macroeconomic variables. According to current literature, the dynamic factor model (DFM) has received much importance, and become a feasible solution for forecasting problem of many potential useful predictors.

There is also other number of areas in which Principal Component Analysis is used for getting feasible solution with minimum computational complexity.

Chapter 3
Literature Review

Copy paste forgery is one of the types of image forgery in which one or more regions of an image is copied and pasted with in the same image. Lots of researches have been done on this topic and there are number of algorithms available for this. But the main problems are accuracy, time complexity, scaling, rotational factor of image and localization etc. Detection techniques of copy paste forgery are classified into two category block based and key point based as shown in figure 3.1.

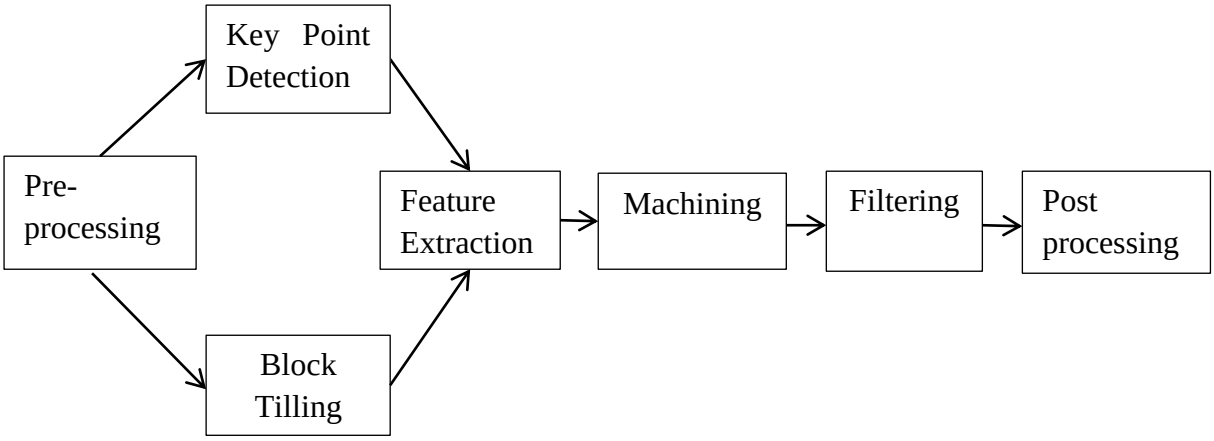


Figure 3.1: Copy Move Forgery Detection Classification

3.1 Block Based Method

There are number of techniques exist now a days that are based on block based method. In block based method, we first divide the image into equal size of overlapping or non-overlapping blocks instead of taking whole image at once. After that we find out the features of every block. These features are compared to each other for detecting copy-move forgery. If the blocks that have same feature then it mean forgery has been done in the image but if there are no blocks that have similar feature then forgery in an image has not been done in the image. There are following block based algorithm exist.

3.1.1 Moment-based (BLUR, HU, and ZERNIKE)

1). Saic and Mahdian [16] has given an algorithm that is based on block based method. In this method we take blur moment for representing the copied part of the image. BLUR moment

are invariant with respect to additive nose and blur degradation. In this method, first we divide the image into equal size of blocks. After that, image is rotated by a specific size of block. In this method we represented every blocks by blur invariant. In this method for removing the dimension of feature vector, we use principal component transformation. We take a threshold value for finding the blocks that have same features. Once we find out the blocks that have same feature after that we must have to verify it. We verify it by counting total number of connected blocks. If two blocks that have same features but have different neighbourhood then they considered as a false positive. With the help of this technique, we detect copy-paste forgery that have blurred duplicated region. But the major disadvantage of this algorithm is computational complexity.

2). Zhang, Wang, Dai and Liu [17] have detected copy-paste forgery with the help of Hu moment. In this algorithm, we have used Gaussian pyramid for compression of the image. This method is invariant to different post-processing like blurring and lossy compression. In this method we first divide the image into equal size of overlapping blocks. After that we calculate eigenvalue and eigenvector by applying the Hu moment to every blocks of the image. These eigenvectors are sorted lexicographically. After that a threshold value is chosen for reducing false detections of image blocks. By using morphological techniques, they performed matching process of the blocks.

3). Mohamadian and Pouyan [18] have given a new technique for detecting copy-paste forgery in an image. In this algorithm, we have used SHIFT algorithm and Zernike moments. SHIFT algorithm cannot detect flat copied part in the image so in this case we are using Zernike moments. In this method first, we extract SIFT feature points. After getting feature points, we match these feature points with each other. In this method, we have used hierarchical clustering for minimizing the possibility of false detection. We can reduce the false detection by using hierarchical clustering. Because in this algorithm, when two cluster have minimum of three similar features then we assume that the image is corrupt. By taking a threshold values, they calculated matching blocks. In this algorithm, we make cluster and find similar cluster that have similar features. There is one disadvantage of SIFT algorithm, it cannot detect flat copy-paste forgeries. It can overcome with the help of Zernike moments. This algorithm takes less time as compared to other algorithm.

3.1.2 Dimensionality reduction-based (PCA, SVD, KPCA, and PCA-EVD)

1). Farid and Popescu [19] have given an algorithm for detecting copy-move forgery in an image that is based on block based method. The main approach of this method is lies upon dimensionality reduction. For dimensionality reduction, we have used PCA algorithm. PCA is mainly used for find out the similarity measure, dimension reduction and feature extraction. In this method, we convert colour image into grayscale image. The image is then divided into equal size of overlapping or non-overlapping blocks. After that for every block, we have to find out the feature vector. For this we calculate eigenvector and eigenvalue. These eigenvector is also called as feature vector of blocks. After obtaining feature vector, we sort these feature vectors lexicographically. We match these sorted feature vector against each other for detecting copy-move forgery. PCA is able to detect even minor variations which are arises due to noise or a lossy compression. With the help of PCA, we can compress the image by removing the dimension of data set. For this we identify eigenvector that corresponds to minimum eigenvalue. Because, this eigenvector contains less information about image. So we can remove this. PCA is very simple and gives accurate result as compared to others algorithm. It has less computational complexity as compare to other algorithm for detecting copy-paste forgery in an image. However this algorithm does not give accurate result in case when image quality is low and block size decreases.

2). Rang ding and Ting [20] have given another copy-paste forgery detection that is based on block based method. Here, we have used Singular Value Decomposition for detecting copy-paste forgery in an image. This method searches similar regions in a tampered image by calculating correlation between copied region and pasted region. The first step of this algorithm is to divide an image into equal size of overlapping blocks. After that for obtaining feature vector of every block, we apply SVD to each block of a tampered image. We match these feature vectors for getting similar feature vector. Matching blocks are finding by transforming every features of block into k-d tree. When the matching blocks are found then it means forgery in an image has been done. This algorithm is not robust for JPEG compression. We use a threshold value for increasing the robust ness of the algorithm and also for reducing the false matching. This algorithm is less complex and robust to post-processing.

3). Noda, Mori Bashar and ohnishi [21] have developed an algorithm for detecting copy-move forgery in an image. This algorithm can detect an image forgery which has an additive noise and lossy JPEG compression. Here in this algorithm, we have used Kernel Principal Component (KPCA) and Discrete Wavelet Transform. First we divide the image into equal size of overlapping and non-overlapping blocks. After that we calculate feature vector of every block in the form of DWT vector and KPCA vector. Then we make a matrix of these feature vectors and apply lexicographical sorting on this feature vector. We use sorted blocks for finding the similar points. To reduce the possibility of false detection, we take a threshold value. This approach is better than PCA approach for finding feature vector of every block of an image. We used this method because of its robustness property of block matching. This algorithm can detecting copy-paste region more accurately as compared to others.

4). Xingming and Zimba [22] propose a method for detecting copy-paste image forgery. This method can detect duplicity that even involves rotation of varying degrees. In this approach we use Discrete Wavelet Transform. We obtain low frequency sub band by applying DWT over in whole image. For detecting duplicacy in an image, we divide the image into equal size of blocks. After that, we perform Principal Component Analysis from which eigenvalues and eigenvectors are obtained. We form matrix from this feature vectors and sort these vectors lexicographically. This sorting method is less complex. After that, we calculate the normalized shift vector and offset frequency. We apply morphological processing in this offset frequency for getting final results. This method is better than common Principal Component Analysis method by compressing an image in the beginning of the process. This method fails to detect forgeries that involve rotation, scaling and heavy compression. Another disadvantage of this method is that it cannot give accurate result when the size of copied and pasted region is very small.

3.1.3 Intensity-based (LUO, BRAVO, LIN and PCMIFD)

1). Qiu, Luo and Huang [23] have given another algorithm of block based techniques. This algorithm use intensities of block of an image for detecting copy-paste forgery. In this method we first divide an image into number of overlapping blocks. After that with help of additive white guassian noise, we divide the blocks of a tampered image into two equal parts and four directions. A block characteristics vector is calculated for every block that is lexicographically sorted. Here there is no need to represent a duplicated region of each pair of similar block feature. In this method we use SHIFT vector algorithm. We consider only those

shift vectors whose occurrence is highest as compared to others. The pairs whose shift vectors are much different from this value are discarded. After that we apply some technique for ensuring, whether forgery in an image has been done or not. This method is robust to post-processing. This method is also less complex. It also does not properly work when forgery in an image is done in a very small region. This algorithm fails when the images have large smooth regions and are highly distorted.

2). Nandi and Bravo [24] have given a block based method for detecting copy-paste forgery in the image. Most of the algorithm fails when image forgery includes reflection, scaling and rotation. This algorithm deals with this kind of post-processing. In this algorithm first we take a window of specific size for sliding an image pixel by pixel. After that we calculate feature vector of blocks that are colour dependent. We sort these feature vectors lexicographically and after matching process is performed. By sliding an image pixel by pixel, we minimize the number of searches. In this method we calculate four features. Three out of four features calculated independently that are blue, green and red components. The fourth feature is entropy of luminance channel. This fourth feature is used for discarding the blocks that contain less information. This algorithm is better than other algorithm as compared to computation and time complexity.

3). Lin et al. [25] propose a new method for detecting copy-paste forgery that is based on block based method. In this algorithm we have used intensity of pixels of an image. We calculate average intensity of single block. In this algorithm we first divide the image into equal size of blocks that are again divided into four blocks. By finding the difference between average intensity and individual intensity, we calculate relative intensity. We do this for each block and obtain feature vector. We use radix sort method for sorting these feature vectors instead of lexicographical sorting technique. This algorithm will not work when corrupted image is rotated by some angle. This algorithm will give proper result when any post processing is done in the image like adding guassian noise and JPEG compression.

4). Sandeep, Sridevi and Mala [27] propose an algorithm for detecting copy move forgery in real time. Detection of copy-move forgery is done in parallel environment. This process is start by first converting an image form colour image into grayscale image. After that we divide the image into equal size of overlapping or non-overlapping blocks. We find out the intensity feature of each block. After that these intensity feature vector is sorted. For sorting they develop another algorithm in which lexicographic sort uses radix sort. With the help of this sorting we can easily detect duplicate block of a tampered image. After that matching

process is performed for finding similar features. If similar feature is found then it means forgery in an image has been done else there is no forgery in the image. We cannot use common techniques like PCA, DWT and SVD for detecting duplicate regions in an image in real time application because of its high computation complexity. This technique has better performance over many other conventional algorithms. We cannot apply this algorithm in colours image.

3.1.4 Frequency-based (DCT, FMT, QCD and LBP)

1). Soukal, Fridrich and Lukas [28] propose an algorithm for detecting copy-move forgery. This algorithm is based on block based method. This method uses DCT coefficient for detecting image forgery. In this algorithm, first we take a window of a specific size. We slide the window pixel by pixel in the image and obtain number of similar block of equal size. We take an array and enter the pixel value of every block into this array. After that we sort the array lexicographically for getting similar feature of blocks in the rows of the matrix. We use this sorted matrix for finding the tampered region in an image. If there in any entry in the matrix that has similar values in the row then it means forgery in the image has been done. But if the matrix contains unique entriy in the row then there is no forgery has been done in the image. For calculating DCT coefficient, we take a Qfactor. For Qfactor we take a particular value. Before matching process, the array is again sorted lexicographically. However, this algorithm cannot recognize differences between large identical textures of a natural image.

2). Memon, Bayram and Sencar [30] develop a new method for detecting copy-paste forgery in an image. This algorithm is based on frequency measurements of blocks. This algorithm uses Fourier Mellin Transform. This method is robust to JPEG compression. This algorithm starts with dividing the image into number of equal size of blocks. After that we apply Fourier Transform to every block. FMT is robust to blurring, noise, translation effect and scaling. FMT is rotational invariant with respect to small rotation angle. We obtain feature vector by quantization and re-sampling of blocks. After that, for getting similar vector, we sort these feature vectors by lexicographical sorting techniques. For minimizing the false positive, we have to calculate the number of connected blocks. Because a natural image has number of similar feature vector due to its similar texture. This algorithm is rotation invariant when rotation has done up to 10 degree. This algorithm cannot detect copy-move forgery in

an image when very small portion of an image is copied. But the major advantage of this algorithm is that it is invariant to noise.

3). Ghorbani, Firouzmand and Faraahi [33] describe an algorithm that is based upon block based technique for detecting duplicate regions in an image. This algorithm use QCD (Quantization Coefficients Decomposition). QCD is performed upon DWT and DCT coefficients. We apply DWT for getting four sub-bands. In this algorithm, we first divide the image into equal size of blocks. Only low frequency sub-band is used for detecting copy-move forgery in an image. After dividing the image into blocks, we apply Discrete Cosine Transform to get DCT feature vectors. After that Quantization Coefficients Decomposition is applied on the Discrete Cosine Transform vectors. We make a matrix with the help of these feature vectors. To minimize computational time, we sort the matrix lexicographically. After that we define a threshold value. We count the number of times a shift vector comes. If the count value is greater than the threshold value then it means forgery in an image has been done. But if the count value is less than the threshold value, it means forgery in an image has not been done. This algorithm is efficient for detecting copy-move forgeries as compared to other algorithms. This technique will not work when image has been through via different post processing like rotation, heavy compression and scaling. This technique is not rotational invariant.

4). Li et al. [34] develop an algorithm for detecting copy-paste forgery in an image. It is based on block based method. In this algorithm, they have used a grayscale operator call Local Binary Pattern (LBP). In this algorithm first, we transform the coloured image into grayscale image. In this algorithm, we use guassian low pass filter because when an image has gone through via several types of post-processing technique like lossy JPEG compression and noise contamination etc. then the high frequency components are not stable in this case. For increasing the detection performance, we use a Gaussian low pass filter more than twice time. Firstly we divide the image into equal size of blocks. After that with help of Local Binary Pattern, we find out the feature vector of every block. Local Binary Pattern is rotational invariant. We make a matrix with the help these feature vectors for finding similar blocks. To reduce the computational time, we sort the matrix lexicographically. After that, we use Euclidean distances to obtain matching blocks. Euclidean distance is obtained for each feature vector and then compared with a threshold value. The blocks that are matched indicate the forged regions in an image. They also detect some false regions. So for this, we

use filtering to minimize the false positives. After that, for reducing the false positive, we perform morphological erosion and morphological processing. This algorithm is invariant to flipping and rotation. However, this algorithm cannot detect image forgery that involves rotation at different angles.

3.2 Keypoint-based methods

1). Zhang and Guo [35] propose a keypoint based method for detecting copy-move forgery. In this algorithm, we use Scale Invariant Feature Transform (SIFT) for making the method more robust against any type of post-processing techniques. In this algorithm, we take correlation between original image region and pasted region. At first, we calculate SHIFT key point, after that we match these keypoints to each other for detecting copy-move forgery. If any SHIFT keypoints are matched then it means image forgery has been done in the image but if there are no similar keypoints then image forgery has not been dome in the image. By identifying the nearest neighbour, matching process is done for every keypoints. After that we take a threshold value. This threshold value is the ratio of two neighbour one is closest and another one is second closest. In this method for selecting threshold value, we take one corrupted and apply detection technique for different value of threshold. This algorithm uses SIFT algorithm successfully for detecting the copy-move forgery. However, this algorithm is not efficient when the tampered region is small.

2). Bo, Junwen, Guangjie and Yuewei [36], develop a new method for detecting copy-move forgery by using SURF (Speeded up Robust Features) algorithm, this algorithm is developed by Herbert Bay et al. this method involves key point detection and its description. We use Hessian matrix for key point detection and Haar wavelets for assigning the orientation. We calculate dominant orientation and describe the orientation of the interest point descriptor. We chose Haar wavelets because it is invariant to the illumination bias. The SURF descriptors are used for matching. For increasing the robustness and avoid false detections, we use a threshold value. We chose a specific value of threshold and test this technique on different images and they are successful. This technique is successful in locating the tampered regions even if post processing is done on the images. We perform post processing like scaling, blurring and rotation on the corrupted images. We use this algorithm for testing and it is successful in showing its robustness for post processing. However, we cannot find the exact boundaries of the tampered region.

3). A study by Zheng, Haoa and Zhub [37] propose a new technique for detecting copy-move forgery. This technique is based on keypoint metching of original region of an image and corrupted region of an image. Keypoints in corrupted regions and original regions should be consistent and they should be distributed evenly over the entire image. This ensures that large similar textures also produce considerable number of keypoints. This method is developed to scan and remove the keypoints at first time. This ensures that there is no impact of noise on them. We scan the keypoints again and find the features for all keypoints. They develop a new method for finding the features and represent these features into a matrix. This algorithm differs from SIFT method in the way of determining features. By recognizing the consistent keypoints in the matrix, this algorithm detects copy-move forgery in the image.

Chapter 4
Proposed Method

We have proposed an efficient algorithm for detecting copy-move regions in the same image. Fig.4.1 shows the block diagram of our proposed algorithm in which we first divide the image into overlapping blocks for detecting the connected blocks that are copied and pasted. The main idea for detecting duplicate regions are that the distance between duplicate block pairs would be same because each block is shifted with same value. After that we have to find out the features of every block. There are many approaches available for extracting features of blocks such as DCT, PCA, SVD, PHT and FMT etc. Here we are using Principle Component Analysis (PCA). After getting the feature vector of every block, we apply lexicographically sorting on that vector for taking the decisions on duplicated regions and then locate the duplicate regions.

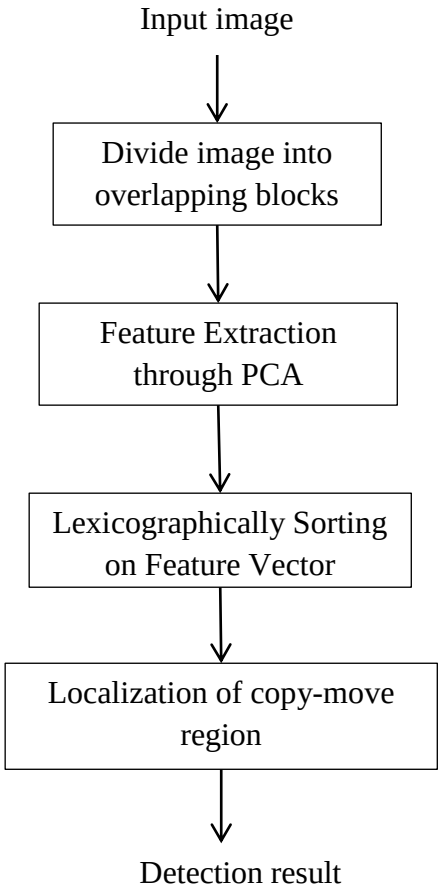


Figure 4.1: Block diagram of proposed algorithm

4.1 Divide Image into Overlapping Blocks

The first step for detecting duplicated regions in an image is to divide an image in a number of overlapping blocks. For an image of size M x M, it is divided into (M-n+1) x (M-n+1) overlapping blocks of size n × n. These blocks are represented as vectors of n^2 dimensions.

The main idea to dividing image into overlapping block is to detect connected blocks that are copied and pasted. The copied region would consist of many overlapping blocks. Block based technique essentially compares blocks in an efficient way and provides invariance to some transformations through an appropriate choice of the method of representation. The size of each block is taken in such a way that minimum blocks are made without leaving any copied region.

4.2 Feature Extraction through PCA

We have used principal component analysis (PCA) for feature extraction. PCA is a very useful statistical technique and it is a common technique for finding pattern in high dimension data. It is a way of identifying patterns in data and expressing the data in such a way that represent their similarity and differences.

The following steps are used to find similar pattern in an image through PCA.

1).At first represent the image in a matrix form of M×N size where M represents number of columns and N represents number of rows in an image. Rows of the pixels of an image are placed one after the other to form one dimension image.

2).The mean is then subtracted from each of the data dimensions to obtain the average across each dimension.

3).The covariance matrix is calculated. Covariance is always measured between two dimensions. If we calculate the covariance between one dimension then it is called variance.

4). Calculate the eigenvectors and eigenvalues of the covariance matrix. All the eigenvector of a matrix are perpendicular no matter how many dimensions are considered. So they are orthogonal.

Eigenvalues are closely related to eigenvectors. Eigenvector and eigenvalues always come in pairs.

5). Eigenvector with the highest eigenvalue is the principle component of the data set. Once eigenvectors are found from the covariance matrix next step is to arrange them from highest to lowest according to eigenvalues. Now eigenvector corresponds to lowest eigenvalue can be ignored. By doing this some information may get lost.

6). Obtain feature vector from eigenvector. Feature vector is constructed by taking the eigenvectors that you want to keep from the list of eigenvectors and forming a matrix with these eigenvectors in the columns.

PCA method for feature extraction of blocks in an image is an efficient and computationally easy as compared to others. PCA provides a foundation for approaching the fields of machine learning and dimension reductions. Since patterns in data can be hard to find in data of high dimensions, for this PCA is a powerful tool to analyse similarity and differences of data. This technique is used in also for image compression.

4.3 Lexicographical Sorting on Feature Vector

After getting the feature vectors of each block a matrix is created with the help of these feature vectors. This matrix is constructed in such a way that all rows of the matrix would represent the blocks of the image and columns would indicate the feature vectors. Now by looking at the matrix, if the two blocks in the images are similar then their feature vector and corresponding rows in the matrix would also be similar. The detection can be done by lexicographical sorting [12]. In lexicographical sorting, comparison assumes both sequences to be of the same length. To ensure they are the same length, the shorter sequence is usually padded at the end with enough "blanks" (a special symbol that is treated as coming before any other symbol). This also allows ordering of phrases.

A lexicographical ordering may not coincide with conventional alphabetical ordering. For example, the numerical order of Unicode code point does not always correspond to traditional alphabetic orderings of the characters, which vary from language to language.

4.4 Detecting duplicate connected blocks

In case of most of the images, they have many similar blocks so finding duplicate blocks in an image is not sufficient for detecting duplicate regions in the same image. For this there should be number of connected blocks with in the same distance. So for this we have to

calculate the distance between two blocks. To find out the distance between two duplicate blocks we used Euclidian distance method.

$$dx(i,j) = |xi - xj|$$
$$dy(i,j) = |yi - yj| \quad (12)$$

For measuring how many blocks are detected as similar within the same distance, a distance measure vector D is constructed. Initial value of D is set to 0.whenever the distance between two blocks are calculated, the corresponding value of D is incremented by 1.

$$D(dx, dy) = D(dx, dy) + 1 \quad (13)$$

Now we have to decide a threshold value for$D(dx, dy)$. If value of $D(dx, dy)$ is greater than the threshold value than it indicates that blocks are copied and pasted with in the same connected distance. So detecting duplicates is based on the value of$D(dx, dy)$. It is not sufficient to find out the duplicate regions in the image, also we have to find out number of duplicate region of connected blocks within the same distance.

4.5 Localization of copy-move region

This is the last step of our proposed algorithm in which we have to detect as many pixels as possible for getting accurate answer from corrupted digital image. For localization we proposed a better algorithm here so that we can get better answer very effectively with less computational complexity.

Let us suppose that we have the location of one pixel in original image as $f(x, y)$ and if that pixel is copied from one location and pasted into another location within the same image then at first we have to find out the distance between copied pixel and pasted pixel suppose that is (Δx, Δy). After shifting the pixel we get the same pixel at location $f'(x + \Delta x, y + \Delta y)$. Now from this we will get the shifted pixel at $f'(x.y)$.

Only those pixels which are copied and pasted have to be shifted and the pixel that is not copied and pasted remain unchanged. Suppose D is the set of those pixels that are copied and pasted.

$$f'(x,y) = \begin{cases} f(x + \Delta x, y + \Delta y) & (x,y) \in D \\ 0 & (x,y) \notin D \end{cases} \quad (14)$$

After finding out spatial position of pasted pixels, we know that the pixels in the two regions have same coordinates in the two images. Now we have to calculate the difference between these two images.

$$f\Delta(x,y) = \begin{cases} |f'(x,y) - f(x,y)| & (x,y) \in D \\ f(x,y) & (x,y) \notin D \end{cases} \quad (15)$$

Because some images go through several types of post processing so the value of corresponding pixels are not entirely same as such so for this we to find out the threshold value for this that leads to better localization in copy-move region. For this we have to take out the minimum threshold value. If we take the greater threshold value in this case then it will give more ambiguous result.

$$f\Delta(x,y) = \begin{cases} f\Delta(x,y) & f\Delta(x,y) > T \\ 0 & f\Delta(x,y) \geq T \end{cases} \quad (16)$$

From the above given approach, we can easily detect copy-move forgery. This method is very efficient, feasible and less computationally complex.

4.6 Algorithm Steps

The detailed steps of above given approach is as follows:

1). First we divide the corrupted image into overlapping blocks.

2).After dividing the image into overlapping blocks, Principle Component Analysis (PCA) is used to find feature vector of each block. PCA helps in finding similarity and differences in various set of data and it is also used for the purpose of image compression with the help of eigenvector and eigenvalue.

3).After getting the feature vector of each block, we apply lexicographical sorting on that feature vectors. The blocks that are copied and pasted have their corresponding feature vector similar.

4). Obtaining duplicate region is not sufficient in the digital image because in most of the image there are many blocks that have same feature. So we find the distance between duplicate blocks. If the distance between duplicate blocks is greater than threshold value then forgery has been done in that image.

5).After taking the decision on blocks whether it is corrupted or not, next step is localization. In localization we first shift the image according to corresponding pixel distance that is calculated.

6).After shifting the image, we will find out the difference between original image and shifted image. Because images go through various post processing method so corresponding pixels value are not entirely same so for this we have to define a threshold value for this. If the pixel value is less than threshold value then that pixel is taken for localization.

Chapter 5
Experimental Results

The proposed method with improved localization method has been implemented. We have conducted a number of experiments on tampered images for analysing the performance and evaluating the results. The experiments are cautiously designed so that the behaviour of the method for different test images can be easily observed. The performance of the method is depend upon: image size, block size, mean, eigenvalue, eigenvector and threshold value. In this algorithm, the main idea is to calculate total number of connected blocks in a duplicate region that is depend upon the threshold value. So we must specify a particular threshold value so that we can get efficient result. In simple image minimum threshold value is taken but for noisy images maximum threshold value is taken.

The copied regions can be indicated with the help of correlation. But, the copied regions are obtained easily from the features of the image. When the results of the copy-move detection are presented to the user then it must not be difficult for the user for identifying the copied regions.

5.1 Selection of Parameter values

The values of the parameters that we have used in our algorithm are depending on the image that we are using whether it is low resolution or high resolution. It is also depend on the system configuration. If our system has low memory and we are using a large image for processing then it will give an exception '*out of memory*'. Here, we are using a maximum number of feature points, as compared to 80 in the original standard [38], because in order to be able to obtain feature points even from non-salient parts as textures like grass and sand, which often are used to hide objects in copy–move forgeries.

First, we convert the coloured image into grey-scale image in which we have to detect copy-move forgery. We have to divide the image into overlapping blocks. For this, we have to define the block size. Here we are taking the size of side of one block is 4.

If our system is working on low memory then we have to resize the image into small size because we work on large size image then it will an exception like out of memory. So here we resize the image into 64 * 64.

After adjusting the size of image we define some derived parameter such as size of image, size of block, total number of blocks in an image.

We define the size of image according to its width and image height.

N (size of image) = imagewidth * imageheight

Similarly we define the size of each block. In this experiment we define the size of side of each block is 4:

B (size of block) = bside * bside

After defining size of image and size of block, we calculate numbers of block formed in an image that is calculated by the following formula:

$$NB = (sqrt(N) - sqrt(B) + 1)^2 \quad (17)$$

In this experiment, we are using principal component analysis for getting feature vector and compression of image. With the help of PCA, we get eigenvalue and eigenvector. These eigenvector can be represent as feature vector. There is eigenvector corresponds to every eigenvalue. The eigenvector corresponds to minimum eigenvalue contain less information about image so we can remove that eigenvector. Because by removing this we will not loss much information. With this process, we can compress the image.

For calculating the eigenvector and eigenvalue, we first calculate covariance matrix. Covariance is such type of measure that finds out that relationship between two dimensions of data set. The formula of covariance is written as-

$$cov(X,Y) = \frac{\sum_{i=1}^{n}(Xi-\tilde{X})(Yi-\tilde{Y})}{n-1} \quad (18)$$

In this algorithm, we have defined two threshold values. One is for getting number of connected block. If total number of connected block is greater than a particular threshold value then it means forgery has done in the image but if total numbers of connected blocks are less than it means forgery has not done in the image. In this experiment, this threshold value is set to 20. That means if total number of connected block is greater than 20 then forgery on the image has done.

Because some images go through several types of post processing so the value of corresponding pixels are not entirely same as such. So for this we define a threshold value for this that leads to better localization in copy-move region. For this we have to take out the minimum threshold value. If we take the greater threshold value in this case then it will give more ambiguous result.

5.2 Detection of Copy-Move Forgery

First, we conduct an experiment in a normal image that does not contain any noise. We have applied this experiment on a set of data. Here, we explain this in two images. These are the cricket and the forest image which are shown in figure 5.1:

(a)

(b)

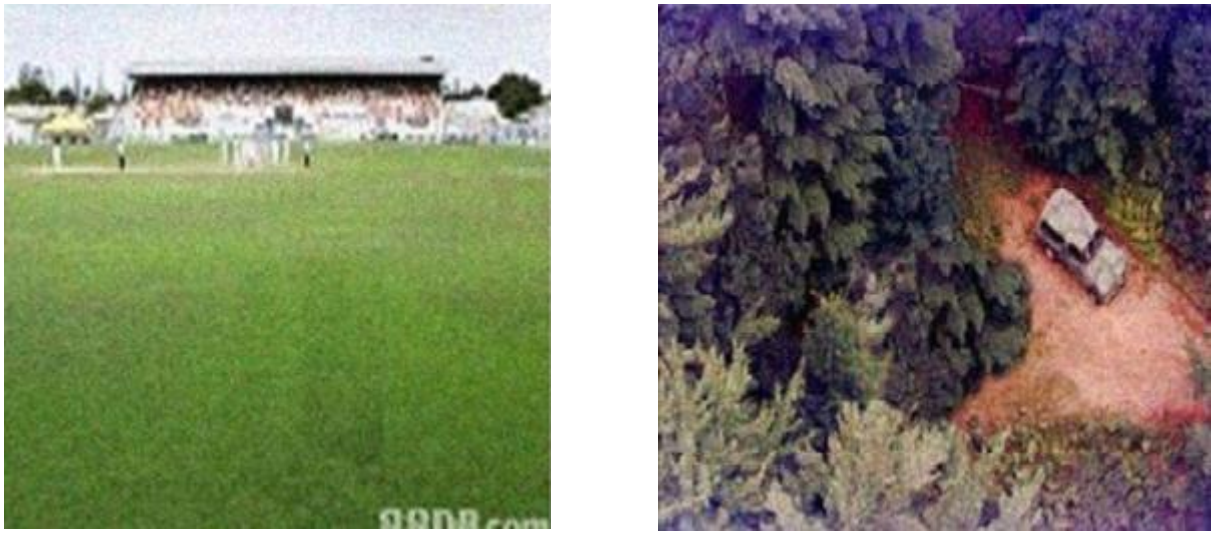
Figure 5.1: Normal image without noise (a) cricket, (b) forest

(a) (b)

Figure 5.2: Tampered image of cricket and forest

(a) (b)

Figure 5.3: Detection Result of Tampered image of cricket and forest (a) cricket image (b) forest image



(a) **(b)**

Figure 5.4: Tampered image with Gaussian noise **(**a) cricket image with Gaussian noise (b) forest image with Gaussian noise

(a) (b)

Figure 5.5: Detection result of tampered image with Gaussian noise **(**a) cricket image with Gaussian noise (b) forest image with Gaussian noise

(a) (b)

Figure 5.6: Tampered image with salt and pepper noise **(**a) cricket image with salt and pepper noise (b) forest image with salt and pepper noise

(a) (b)

Figure 5.7: Detection result of tampered image with salt and pepper noise **(**a) cricket image with salt and pepper noise (b) forest image with salt and pepper noise

From the figure 5.5 and 5.7, we can see that our result is independent of noise. First, we add Gaussian noise into tampered image then we apply our algorithm to find out copy-move forgery after that we check it for salt and pepper noise. So the method that we are using is invariant to noise.

Chapter 6
Conclusion and Future Work

6.1 Conclusion

We present an effective method to detect duplicate regions in digital images. With the use of advance software even a non-professional can do copy-move forgery. In order to detect these kind of forgery effectively, an algorithm must be accurate and less computationally complex. Our main aim in this approach is to detect duplicate regions with high accuracy with less complexity.

In our approach we have used Principal Component Analysis (PCA) for acquiring the feature vector of each block in the image but the performance relies upon detection of duplicate regions and localization of pixels. We are using eigenvector for reducing the dimensions of the image. We are using a novel and efficient localization method.

Despite some achievement, some challenges remain in digital image forgery like robustness of existing tools. When the image is noisy, then the results are different from the expected results, so this algorithm is not invariant for noise. However, when the image does not contain noise then our results show that our approach can detect duplicate regions very accurately and efficiently.

6.2 Future Work

There are some challenges that we have faced during experiments for detecting copy-move forgery like- tampered image with rotation, tampered image with noise and tampered image with compression. In some cases when a very small region of the image is corrupted then it creates some problem in this method. In this case, proposed method gives improper result.

There are following enhancement that we can do in our algorithm:

- Detecting change in satellite images, video forgery and spliced images etc.
- Getting accurate results in case of very small region of the image is corrupted in the same image.
- Using alternate algorithm for getting feature vector instead of principal component analysis (PCA).

- Detecting duplicate regions in an image with rotation and scaling.

However as the number of forgery techniques keeps on increasing and the anti-forensic tools that is freely available on the internet through which even a normal person can create tampered image, detection will become more and more difficult. So in future there will always be scope of further improvement.

References

[1] M. Chandra, S. Pandey and R. Chaudhary, "Digital Watermarking Technique for Protecting Digital Images" published in ICCSIT, IEEE conference vol. 7, pp. 226-233, july 2010.

[2] Niels Provos and Peter Honeyman, University of Michigan, "Hide and Seek: An Introduction to Stagnography" published by IEEE computer society, 1540-7993/03.

[3] M. Sridevi, C. Mala and Siddhant Sanyam, "Comparative study of Image Forgery and Copy-Move Techniques" published in springer, advances in computer science, Eng. & Appl. AISC 166, pp. 715-723, 2012.

[4] Judith A. Redi, Weim Taktak, Jean-Luc Dugelay, "Digital Image Forensics: a booklet for Beginners", Multiledia tools and applications in Springer, An International Journal, 10.1007/s11042-010-0620-1.

[5] Jiu Zhulong, Li Xianghua, Zhao Yuqian, "Passive Detection of Copy Paste Tampering for Digital Image Forensics", IEEE conference on Intelligent Computation Technology and Automation in 2011.

[6] Pravin Kakar and N. Sudha, "Exposing Postprocessed Copy-paste Forgeries Through Transform-Invarient Features", IEEE transaction on Information Forensics and Security, vol. 7, No. 3,, june 2012.

[7] B. Moore, University of Toronto, "Principal Component Analysis in Linear Systems Controllability, observability, and model reduction" IEEE transaction on Automatic Control, vol. 26, ISSN: 0018-9286, pp. 17-33, feb 1981.

[8] Lu Jinpeng and Jiang Dalin, "Survey on the technology of image processing based on DCT compressed domain" published in Multimedia Technology (ICMT), IEEE Conference ISBN: 978-1-61284-771-9, pp.-786-789, july 2011.

[9] Yang Chun Ling, Gao Wen-Rui and Po Lai-Man, "Discrete wavelet transform-based structural similarity for image quality assessment" image conferencing, ICIP 2008, 15th IEEE conference ISSN: 1522-4880, pp. 370-380, oct. 2008.

[10] Li Guangzhen, N. Nitta and Yu Xiaoyi, "A discrete wavelet transform based recoverable image processing for privacy protection" image conferencing, ICIP 2008, 15th IEEE conference pp. 1372-1375, oct. 2008.

[11] Guo Xiaoxin, Xu Zhiwen, Lu Yinan and Pang Yunjie, "An Application of Fourier-Mellin Transform in Image Registration" published in Computer and Information Technology, CIT 5th IEEE conference, pp.-619-623, sep. 2005.

[12] V. Gaidhane, V. Singh and M. Kumar, "Image Compression Using PCA and Improved Technique with MLP Neural Network" published in ARTcom, IEEE conference, pp. 106-110, oct 2010.

[13] Lim Sunghyun and Lee Chulhee, "Principal component analysis for compression of hyperspectral images" published in Geoscience and remote sensing symposium, IEEE conference, vol. 1, pp. 97-99, july 2001.

[14] Jian Yang, David Zhang, Alejandro F. Frangi, and Jing-yu Yang, "Two-Dimensional PCA: A New Approach to Appearance-Based Face Representation and Recognition" IEEE Transaction on Pattern Analysis and Machine Intelligence, vol. 26, no. 1, jan 2004.

[15] Pan, X. Z. & Wang, H. M. (2012), "The Detection Method of Image Regional Forgery Based DWT and 2DIMPCA", Advanced Materials Research, 532, 692-696.

[16] Mahdian, B., & Saic, S. (2007), "Detection of copy–move forgery using a method based on blur moment invariants" Forensic Science International, 171(2), 180-189.

[17] Wang, J.-W., Liu, G.-J., Zhang, Z., Dai, Y., & Wang, Z. (2009), "Fast and robust forensics for image regionduplication forgery", Acta Automatica Sinica, 35(12), 1488-1495.

[18] Mohamadian, Z., & Pouyan, A. A. (2013), "Detection of Duplication Forgery in Digital Images in Uniform and Non-uniform Regions", Paper presented at the UKSim.

[19] Popescu, A. C., & Farid, H. (2004), "Exposing digital forgeries by detecting duplicated image regions", Dept. Comput. Sci., Dartmouth College, Tech. Rep. TR2004-515.

[20] Ting, Z., & Rang-ding, W. (2009). Copy-move forgery detection based on SVD in digital image. Paper presented at the Image and Signal Processing, 2009. CISP'09. 2nd International Congress on.

[21] Bashar, M., Noda, K., Ohnishi, N., & Mori, K. (2010), "Exploring duplicated regions in natural images", IEEE Transactions on Image Processing,(99), 1.

[22] Zimba, M., & Xingming, S. (2011), "DWT-PCA(EVD) Based Copy-move Image Forgery Detection", International Journal of Digital Content Technology and its Applications, 5(1).

[23] Luo, W., Huang, J., & Qiu, G. (2006), "Robust detection of region-duplication forgery in digital image", Paper presented at the Pattern Recognition, 2006. ICPR 2006. 18th International Conference on.

[24] Bravo-Solorio, S., & Nandi, A. K. (2011), "Automated detection and localisation of duplicated regions affected by reflection, rotation and scaling in image forensics", Signal Processing, 91(8), 1759-1770.

[25] Lin, H.-J., Wang, C.-W., & Kao, Y.-T. (2009), "Fast copy-move forgery detection", WSEAS Transactions on Signal Processing, 5(5), 188-197.

[26] Wang, J., Liu, G., Li, H., Dai, Y., & Wang, Z. (2009), "Detection of image region duplication forgery using model with circle block", Paper presented at the Multimedia Information Networking and Security, 2009. MINES'09. International Conference on.

[27] Sridevi, M., Mala, C., & Sandeep, S. (2012), "Copy–move image forgery detection", Computer Science & Information Technology (CS & IT), 52, 19-29.

[28] Fridrich, A. J., Soukal, B. D., & Lukáš, A. J. (2003), "Detection of copy-move forgery in digital images" Paper presented at the in Proceedings of Digital Forensic Research Workshop.

[29] Zhang, J., Feng, Z., & Su, Y. (2008), "A new approach for detecting copy-move forgery in digital images", Paper presented at the Communication Systems, 2008. ICCS 2008. 11th IEEE Singapore International Conference on.

[30] Bayram, S., Sencar, H. T., & Memon, N. (2009), "An efficient and robust method for detecting copy-move forgery", Paper presented at the Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on.

[31] Li, L., Li, S., & Wang, J. (2012), "Copy-move forgery detection based on PHT", Paper presented at the Information and Communication Technologies (WICT), 2012 World Congress on.

[32] Muhammad, G., Hussain, M., Khawaji, K., & Bebis, G. (2011a), "Blind copy move image forgery detection using dyadic undecimated wavelet transform", Paper presented at the Digital Signal Processing (DSP).

[33] Ghorbani, M., Firouzmand, M., & Faraahi, A. (2011), "DWT-DCT (QCD) based copy-move image forgery detection", Paper presented at the Systems, Signals and Image Processing (IWSSIP), 2011 18th International Conference on.

[34] Li, L., Li, S., Zhu, H., Chu, S.-C., Roddick, J. F., & Pan, J.-S. (2013), "An Efficient Scheme for Detecting Copy-move Forged Images by Local Binary Patterns", Journal of Information Hiding and Multimedia Signal Processing, 4(1), 46-56.

[35] Huang, H., Guo, W., & Zhang, Y. (2008), "Detection of copy-move forgery in digital images using SIFT algorithm", Paper presented at the Computational Intelligence and Industrial Application, 2008. PACIIA'08. Pacific-Asia Workshop on.

[36] Bo, X., Junwen, W., Guangjie, L., & Yuewei, D. (2010), "Image copy-move forgery detection based on SURF", Paper presented at the Multimedia Information Networking and Security (MINES), 2010 International Conference on.

[37] Zheng, J., Haoa, W., & Zhub, W. (2012), "Detection of Copy-move Forgery Based on Keypoints' Positional Relationship", Journal of Information and Computational Science, 1(3), 53-60.

[38] Image Signature Tools ISO/IEC Std. ISO/IEC 15938-3:2002/Amd. 3:2009(E) [Online]. Available:http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=50550.

[39] Kusum, Pawanesh Abrol and Devanand, "Digital Tampering Detection Techniques: A Review", BIJIT - BVICAM's International Journal of Information Technology, vol. 1 no. 2, ISSN 0973-5658.