ABSTRACT

Using some transformation on original images, images can be modified which forms near duplicate images. Using the size and length of image, image represented and the length is variable with number of patches in the image. Spatially adjacent and similar pixels are used to form clusters. Patch means a form of clusters. Image representation and image similarity measurement are two major issues in image matching. The proposed method extracts patches from given image and represents by variable length signature. In the detection of duplicate natural images the signatures are use. Then take a decision for images are duplicates or not. The retrieval of near duplicate images means a query image and datasets are given then the near duplicate images are retrieving. Similarity is computed between two images, which can be done on variable length signature to improve the effectiveness.

Keywords
Image retrieval, Near duplicate image, Similarity matching, Variable length signature.

1. INTRODUCTION

The original images are modified using some transformations like scaling, rotation which creates near duplicate images. The photograph of original image are altered slightly that matching is called detection of near duplicate image. The images which are close to an image according to some measures then also called near duplicate image. Retrieval of subimages means match an original image with its small portion. The near duplicate image matching has different applications like postal automation, and copyright. There are different algorithms for near duplicate image matching. Also document images are also retrieving. Document image retrieval means finding the relevant or similar document images. Fordocument image retrieval feature extraction is important. When a query shot is given then duplicate shots are find called near duplicate shot detection. The near duplicate images have some common transformations like changing contrast, saturation, scaling, cropping and framing. For image matching two common issues are important i.e. image representation and image similarity. The different similarity measures are given for image retrieval. By computing distance between features vectors similarity is measure for image retrieval. The probabilistic method is better than geometric similarity measures.

2. LITERATURE SURVEY

For the image representations different key points, intersection points, pixels, patches are used. Also for computation of similarity different mathematical tools are used like Euclidean distance.

Bin Wang, Zhiwei Li, Discussed duplicatedetection algorithm. Duplicate images can bedetected in dataset i.e. large. Identical images can be computed in large dataset of images important in many applications. Firstly they convert the image to a k-bit hash code and then conducting the detection of duplicate images with hash codes. By using hash code the representation form of image is compact.

So for further processing the clustering is required.

Yan ke, Rahul Sukthankar propose a system for detecting duplicate images and retrieving some near parts of
images. The detection of duplicate images and sub parts of images are retrieve using some descriptors and interest points for problem solving. Distinctive local descriptors used for parts based representation of images and also provides high quality matches for several transformations. Locality sensitive hashing is used for indexing the local descriptors. Hundreds of features need to query, so for parts based approach which is slow i.e. the limitation of the system.

So to overcome the limitation of sub part retrieval system Narendra Ahuja proposes system for matching region based images. Two images are given in which the part in image one is consider for matching the parts in other images. There are different properties for regions are area, shape, boundary, and colour. Using trees, images are represented and multiscale segmentation algorithm is used for computation.

Pattern recognition technology has great progress postal automation. They give the details of image acquisition. Postcode, address segmentation and recognition are key technologies in automatic letter sorting machines. James Philbin proposes system for detection of video shot and for duplicate images. For detection, two novel schemes are used. Global Hierarchical Colour Histograms is the first for retrieval fastly. Min hash algorithm in the second scheme. In this no extra computation effort require for similarity measures.

3. PROPOSED WORK

The proposed system details are discussed in this section. Figure 2 shows Block diagram which shows the overall system. Images are the system inputs. We have to verify the images which are near duplicates or not. Providing some thresholds to some similarity measures near duplicate images are verified. The functions of proposed system for each block are discussed.

1) Pre-processing:

Because of the different image formation characteristics, images differ in appearance. In this the images are scan. Also the image parameters are calculated. The images are converted into gray scale images.

2) Patch Extraction:

To extract the patches from images is a difficult task. For extraction of patch Pixel intensity is employed. Spatially adjacent and similar pixels are used to form clusters. Patch means a form of clusters.

3) Patch Visual Appearance:

Appearance of patch is computed by using different descriptors. A new visual descriptor is used for visual appearance of patch.

4) Patch Relationship:

Describes the relationships between patches and Compute the distance between each patch.

5) Patch Weight:

For each patch a weight is assign in the image. Patch weight is a ratio of sizes of patch and entire image.

6) Content Measurement:

The image signature is generated and similarity is calculated.

7) Distance Matching:

To compute similarity Earth mover’s distance is used. Similarity results are given.

Fig.2. System Block Diagram
4. EXPERIMENTAL SETUP
4.1 Datasets
eSRI: This dataset has 17,156 envelope images and in this dataset for construction of query set 600 images are randomly taken.

eUW2: It includes 623 typewritten document images.

PRI: PRI dataset includes envelop images and some photos of envelop images with different conditions.

4.2 Performance Measures
The performance will be evaluated by using recall and precision computation. For query q the relevant images in the database is denoted as R(q), and the result of retrieval is denoted as Q(q). The image which is relevant but is not retrieved from the database is denoted by N(q). The precision of the retrieval is defined as the fraction of the retrieved images that are indeed relevant for the query.

Precision = \frac{R(q)}{Q(q)}

The recall is the fraction of relevant images that is returned by the query.

Recall = \frac{R(q)}{R(q) + N(q)}

4.3 Results
The proposed system uses variable length signatures for matching near duplicate images. The average numbers of patches are calculated. Then signature value is generated for the images. Similarity is computed between two signatures and has data sets. For patch appearance new descriptors are used.

4.4 Performance Analysis
The computation time required for signature generation and similarity computation is calculated which is better.

4.5 Comparison with State-of-the-Art Methods
To verify the validity of the proposed approach in detecting near-duplicate images, we compare it with the three aforementioned methods. From the EER comparison and the ROC curves, it is obvious that the proposed approach achieves the best performance overall on the two data sets.

5. SUMMARY AND CONCLUSION
There are different ways for matching near duplicate images. For matching near duplicate images variable length signature is proposed. For representation of image
patches are used. The signature length varies with number of patches. Probabilistic Centre Symmetric Local Binary Pattern is used for patch appearance in the image. In two images the similarity is computed by using earth mover’s distance issued.

6. ACKNOWLEDGEMENTS
I would like to express my special thanks to all those people who have helped me to complete this work. I am very grateful to my guide, Prof. J. R. Mankar, Department of Computer Engineering, KKWIEER, Nasik.

7. REFERENCES