Matching and Retrieval of Near Duplicate Images

Vishakha B. Pawar
Department of Computer Engineering
KKWIEER,
Nashik, India
SavitribaiPhule Pune University,
Pun

J. R. Mankar
Department of Computer Engineering
KKWIEER,
Nashik, India
SavitribaiPhule Pune University,
Pune

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 whichcreates near duplicate images. The photographof original image are altered slightly thatmatching is called detection of near duplicateimage. The images which are close to an imageaccording to some measures then also calledas near duplicate image. Retrieval of subimages means match an original image withits small portion. The near duplicate imagematching has different applications like postalautomation, and copyright. There are differentalgorithms for near duplicate image matching.Also document images are also retrieving.Document image retrieval means finding therelevant or similar document images. Fordocument image retrieval feature extractionis important. When a query shot is giventhen near duplicate shots are find called nearduplicate shot detection. The near duplicateimages have some common transformationslike changing contrast, saturation, scaling,cropping and framing. For image matchingtwo common issues are important i.e. imagerepresentation and image similarity. Thedifferent similarity measures are given forimage retrieval. By computing distancesbetween features vectors similarity is measurefor image retrieval. The probabilistic methodis better than geometric similarity measures.

Fig.1 Examples of Near duplicate image pairs due to lighting, viewpoint, colour, and lens variations

Near-duplicate image matching is doneby usingvariable length signature. Signaturelength can be varying with number of patches.Representation of an image can be done byimage. For appearance of patch in each image,a new visual descriptor is used i.e. ProbabilisticCentre Symmetric Local Binary Pattern. Thespatial relationships between the patches arecalculated, beyond each individual patch. Thesimilarity between two images is computed.There are two different applications of imagesignatures are near duplicate natural imagedetection and near duplicate document imageretrieval.

Paper organization includes basicintroduction and proposed system description.Second section includes the literature survey.Mathematical model and proposed systemblock diagram are in the section third. Sectionfour includes the results and discussion.

2. LITERATURE SURVEY
For the image representations different keypoints, intersection points, pixels, patchesare used. Also for computation of similaritydifferent mathematical tools are used likeEuclidean distance.

Bin Wang, Zhiwei Li, Discussed duplicatedetection algorithm. Duplicate images can bedetected in dataset i.e. large. Identical imagescan be computed in large dataset of imagesis important in many applications. Firstly thek-bit hash code for each image is calculated

i.e. converting each image to a k-bit hash codeand then conducting the detection of duplicateimages with hash codes. By using hash codesthe representation form of image is compact.

So for further processing the clustering isrequired.

Yan ke, Rahul Sukthankar propose a systemfor detecting duplicate images and retrievingsome near parts of

images. The detection ofduplicate images and sub parts of images areretrieve using some descriptors and interest

points for problem solving. Distinctive localdescriptors used for parts based representationof images and also provides high qualitymatches for several transformations. Localitysensitive hashing is used for indexing thelocal descriptors. Hundreds of features need to query, so for parts based approach which isslow i.e. the limitation of the system.

So to overcome the limitation of sub partretrieval system NarendraAhuja proposesa system for matching region based images.Two images are given in which the part inimage one is consider for matching the partsin other images. There are different propertiesfor regions are area, shape, boundary, andcolour. Using trees, images are representedand multiscale segmentation algorithm is usedfor computation.

Pattern recognition technology has greatprogress postal automation. They give thedetails of image acquisition. Postcode,address segmentation and recognition arekey technologies in automatic letter sortingmachines.James Philbin proposes system for detectionof video shot and for duplicate images. Fordetection, two novel schemes are used. GlobalHierarchical Colour Histograms is the first forretrieve fastly. Min hash algorithm in the secondscheme. In this no extra computation effortrequire for similarity measures.

3. PROPOSED WORK

The proposed system details are discus in thissection.
Figure 2 shows Block diagram which showsthe overall system. Images are the systeminputs. We have to verify the images whichare near duplicates or not. Providing somethresholds to some similarity measures nearduplicate images are verified. The functions ofproposed 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 imageparameters are calculated.The images areconverted into gray scale images.

2) Patch Extraction:

To extract the patches from images is adifficult task. For extraction of patch Pixelintensity is employed. Spatially adjacent andsimilar pixels are used to form clusters. Patchmeans a form of clusters.

3) Patch Visual Appearance:

Appearance of patch is computed by usingdifferent descriptors. A new visual descriptoris use for visual appearance of patch.

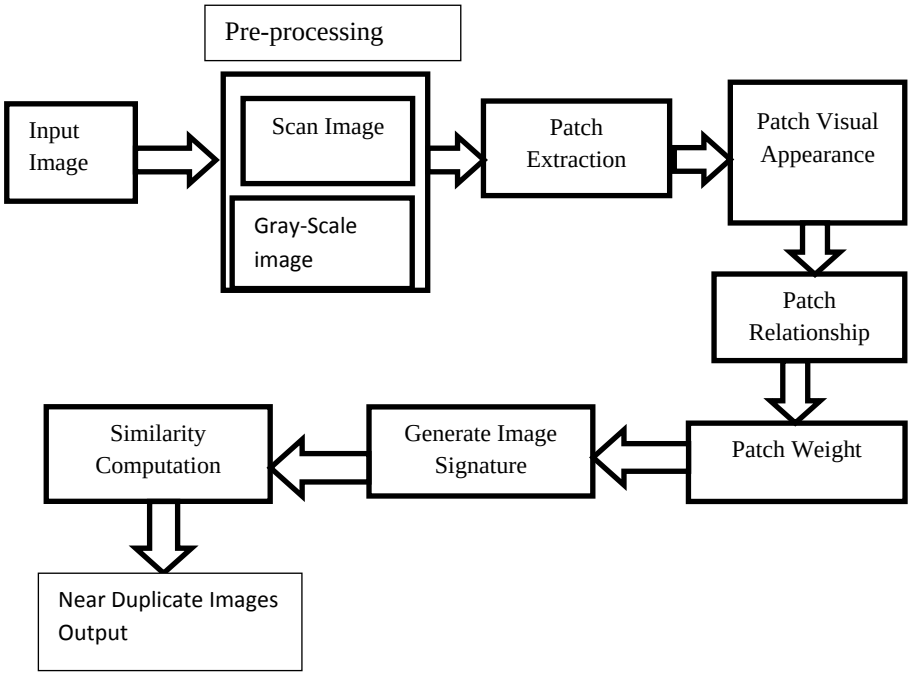

Fig.2. System Block Diagram

4) Patch Relationship:

Describes the relationships between patchesand Compute the distance between each patch.

5) Patch Weight:

For each patch a weight is assign in theimage. Patch weight is a ratio of sizes of patchand entire image.

6) Content Measurement:

The image signature is generated andsimilarity iscalculated.

7) Distance Matching:

To compute similarity Earth movers distance is used. Similarity results are given.

4. EXPERIMENTAL SETUP

4.1 Datasets

eSRI:

This dataset has 17,156 envelope images andin this dataset for construction of query set600 images are randomly taken.

Fig.3. a sample pair of near duplicate images on PRI dataset

eUW2:

It includes 623 typewritten documentimages.

PRI:

PRI dataset includes envelop images andsome photos of envelop images with differentconditions.

4.2 Performance Measures

The performance will be evaluated by usingrecall and precision computation. For query qthe relevant images in the database is denotedas R(q), and the result of retrieval is denotedas Q(q). The image which is relevant but isnot retrieved from the database is denoted byN (q). The precision of the retrieval is definedas the fraction of the retrieved images that areindeed relevant for the query.

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

The recall is the fraction of relevant imagesthat is returned by the query.

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

4.3 Results

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

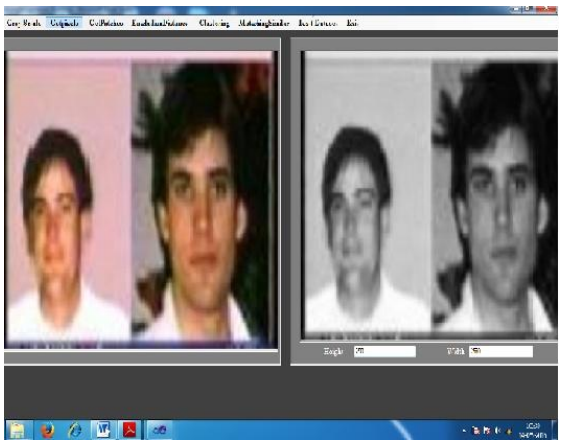

Fig.4. Gray-scale Image

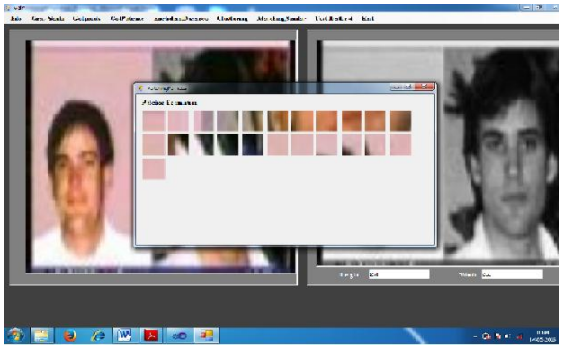

Fig.5. Patch Extraction

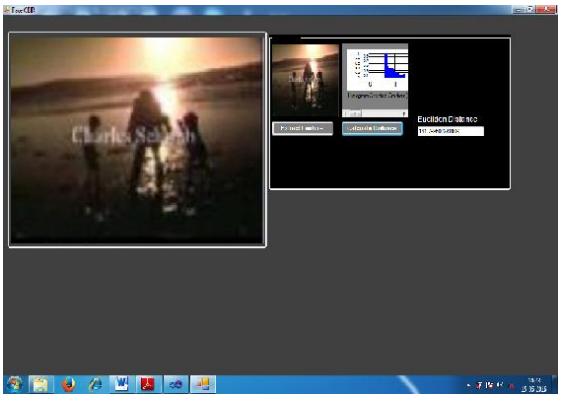

Fig.6. Feature Extraction and calculate distance

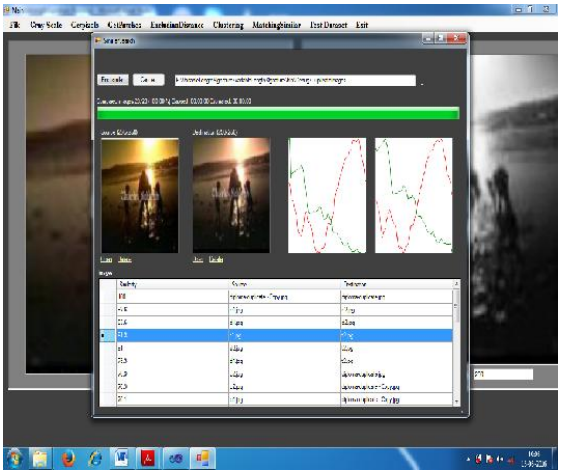

Fig.7. Similarity Matching

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 nearduplicate images. For matching near duplicateimages variable length signature is proposed.For representation of image

patches are used.The signature length varies with number ofpatches. Probabilistic Centre Symmetric LocalBinary Pattern is used for patch appearancein the image. In two images the similarity iscomputed by using earth movers distance isused.

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

[1] Li Liu, Yue Lu, Senior Member, IEEE, and Ching Y. Suen,Fellow,Variable-Length Signature for Near-Duplicate ImageMatching, IEEE NO. 4, APRIL 2015.

[2] F. Zou et al., Nonnegative sparse coding induced hashing forimage copy detection, Neurocomputing, vol. 105, no. 1, pp.81-89, 2013.

[3] G.-H. Liu and J.-Y.Yang,Content-based image retrieval usingcolor difference histogram, Pattern Recognit., vol. 46, no. 1,pp. 188-198, 2013.

[4] Y. Lu, X. Tu, S. Lu, and P. S. P. Wang, Application of patternrecognition technology to postal automation in China,IEEEPattern Recognition and Machine Vision-in Honor andMemory of Professor King-Sun Fu. Copenhagen, Denmark: River Pub.Co., Mar. 2010, pp. 367-381.

[5] S.Todorovic and N. Ahuja, Region-based hierarchical imagematching," Int. J.Comput.Vis., vol. 78, no. 1, pp. 47-66, 2008.

[6] O. Chum, J. Philbin, M. Isard, and A. Zisserman, Scalablenear identical image and shot detection,in Proc. 6th ACM Int.Conf. Image Video Retr., 2007, pp.549-556.

[7] B. Wang, Z. Li, M. Li, and W.-Y.Ma,Large-scale duplicatedetection for web image search," in Proc. IEEE Int. Conf.Multimedia Expo, Jul. 2006, pp. 353-356.

[8] D.-Q. Zhang and S.-F.Chang,Detecting image near-duplicateby stochastic attributed relational graph matching with learning,in Proc. ACM Int. Conf. Multimedia,2004, pp. 877-884.

[9] Y. Meng, E. Chang, and B. Li,Enhancing DPF for near-replicaimage recognition,in Proc. Int. Conf. Comput.Vis. Pattern Recognit., Jun. 2003, pp. II-416-II-423.

[10] C. Kim,Content-based image copy detection,Signal Process.,Image Commun.,vol. 18, no. 3, pp. 169-184, 2003.