A Survey on Finding Image Similarity and Retrieval of Near Duplicate Images

Bedade Jayshree[1] Prof.N.L.Bhale[2]

[1] ME Student, Department of Computer Engineering, *Mcoerc Nasik, Maharashtra, India*
[2] *HOD,* Department of Information Technology, *Mcoerc Nasik, Maharashtra, India*

ABSTRACT

The image can be modified using some transformation and scaling on original images which forms near duplicate images. An image can be represented by their size or length and the number of patches in the image varies with respect to the length. The adjacent and visually same pixels are clustered or grouped and then the resulting cluster is considered as a patch. Image representation and image similarity measurement are two major issues in image matching. The proposed method extracts patches from given image and represents by changeable length signature. The signature is further verify and validated in a near duplicate natural image detection which makes a decision about whether two images are near duplicates or not. The near duplicate image retrieval aims at retrieving relevant or same images from image database which are similar to query image. The similarity between two images is finding by Using Earth Move Distance Algorithm.

1. INTRODUCTION

An image is representing by the variable length signature. The length of the signature is indifferent as per number of patches in the image .Images having indifferent features then shows them as vector. NEAR-DUPLICATE images are form by taking separate photos of the same object. They can be created by changing the original images using some transformations such as image rotation and scaling. To calculate the similarities between two images we used earth mover distance algorithm. In many applications such as postal automation, copyright protection, etc. This paper proposed the system in which we retrieve duplicate image and finding the near duplicate images .Also finding similarities and dissimilarities .Now a day Near duplicate image detection and image retrieval is an problem when we copyright, or in postal automation so this problem solve by the given paper. Different Basic things are used to represent images are raw, pixel key point and so on. Patch composed of pixels which mostly similar. Probabilistic centre symmetric local binary method is used for describe emergence of each image piece and also represent the patch appearance. In this we also express the relationship between patches. To work out similarities Earth Movers Distance algorithm is employed.

2. LITERATURE SURVEY

Bin Wang, Zhiwei Li,[1] Discussed duplicate detection algorithm for detecting visually duplicate images in a set of images .In lot of applications finding visually identical images in large image collections is important. Firstly the k-bit hash code for each image is calculated i.e. each image is converted to a k-bit hash code according to its content and then conduct the duplicate image detection with only the hash codes.

Yan ke, Rahul Sukthankar [2], Larry Huston propose a system for near duplicate detection and sub image retrieval. The near duplicate detection and sub image retrieval problem solve by using robust interest point detection (DoG detector), local descriptor (PCA-SIFT), and efficient similarity search of high dimensional data (LSH). Parts based representation of images using distinctive local descriptors and give high quality matches under several transformations. Locality sensitive hashing used to index the local descriptors. The limitation of system and a potential drawback in using a parts-based approach is that system needs to query hundreds to thousands of features at a time which could be slow.

Narendra Ahuja [4] presents an approach to region-based hierarchical image matching .The two images are given and the goal is to identify the largest part in image 1 and its match in image 2 having the maximum similarity measure defined in terms of geometric and photometric properties of regions(e.g., area, boundary shape, and color), as well as region topology (e.g., recursive embedding of regions). To this end, each image is represented by a tree of recursively embedded regions, obtained by a multiscale segmentation algorithm.

Pattern recognition technology has achieved great progress in many applications, among which postal automation is one of its representative and successful practical areas. In this Yue Lu present the applications of pattern recognition technology to postal automation in China, in particular, give the details of image acquisition, postcode and address segmentation and recognition which are key technologies in automatic letter sorting machines.

James Philbin [3] proposes and compares two novel schemes for near duplicate image and video-shot detection. The first approach is based on global hierarchical color histograms, using Locality Sensitive Hashing for fast retrieval. The second approach uses local feature descriptors (SIFT) and for retrieval exploits techniques used in the information retrieval community to compute approximate set intersections between documents using a min-Hash algorithm. The requirements for near-duplicate images vary according to the application, and address two types of near duplicate definition: (i) being perceptually identical and (ii) being images of the same 3D scene.

3. SYSTEM FLOW

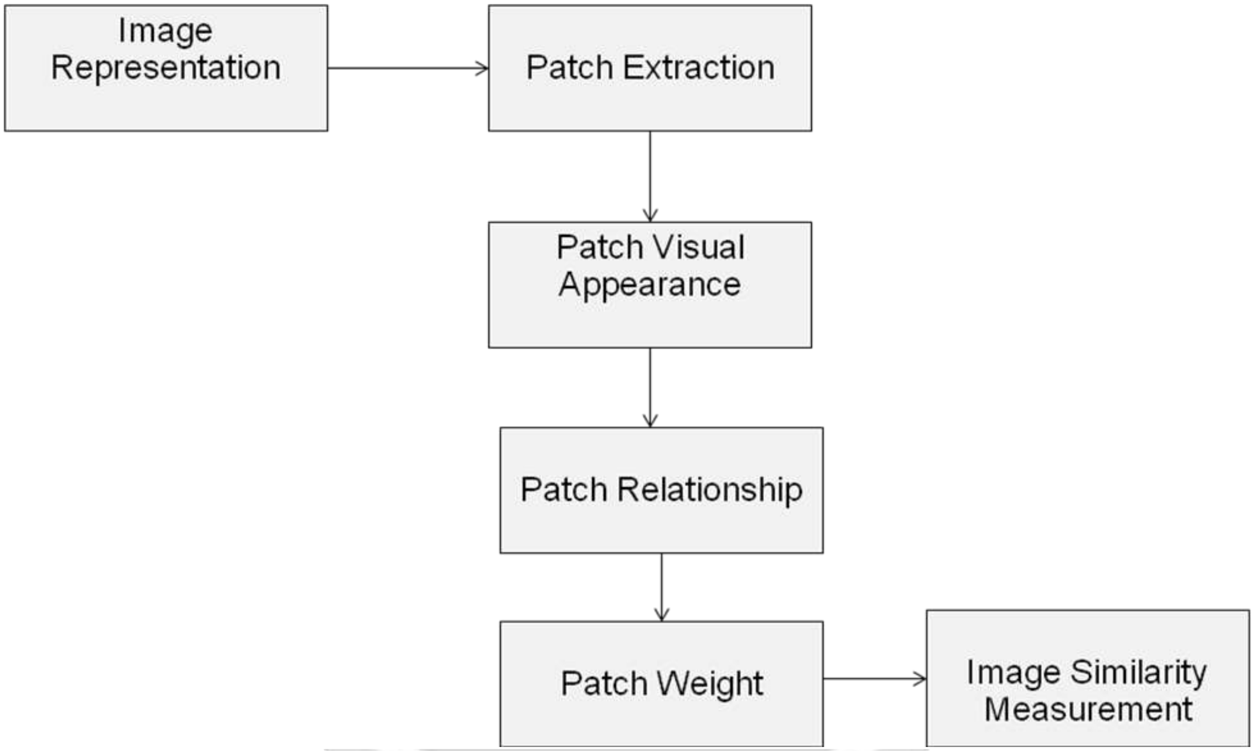

Fig.System Flow

A) Patch Extraction:

Extracting stable patches from an image is a challenging issue. From time to time, even small variations of the images cause to be different patches, which will hamper the subsequent matching process. Pixel intensity was employed in and for patch extraction, while the spatial aspect of the pixel was totally ignored. on the other hand, the spatial aspect was the major concern in the approaches . In this paper take into account both the spatial aspect and pixel intensity for patch extraction. Specifically, we formulate patch extraction in terms of clustering. The pixels which are spatially nearby and similar in intensity are clustered, and then each resulting cluster is consider as a patch. No heuristics about the number of clusters are required in advance. Additionally, the outliers that form isolated islands in the image which do not belong to any cluster can be identified.

B) Patch Visual Appearance:

To describe patch visual appearance, good strength to image orientation, illumination and scale variations is highly desired. In our work, we propose a patch visual appearance descriptor, viz. Probabilistic Center-symmetric Local Binary Pattern (PCSLBP), which is an improvement of Center symmetric Local Binary Pattern (CSLBP).This first give a concise introduction to CSLBP and then elaborate on PCSLBP.

C) Patch Relationship:

Beyond the patch visual appearance, we grab the relationships between the patches in the image. More specifically, we compute the Euclidean distance between the gravity centers of two patches. The possible maximal distance between two patches in the image is quantized into L levels. Then a histogram HRi of L bins is related with patch Oi, with the lth(1< l < L) bin representing the number of patches in the image whose distances from Oi fall into the lth level.

D) Patch Weight:

We allot a weight ωi to every patch Oi in the image to indicate its contribution giving in identifying the image. It is defined to be proportional to the size of the patch, simply based on the supposition that the bigger a patch is, the more important it is for the image. Formally, ωi is computed according to

$$\omega_i = \frac{size(\mathcal{O}_i)}{size(I)},$$

E) Image Similarity Measurement

With every image represented by a signature, computing the similarity between two images transforms to the similarity computation between two signatures. Since the signature is of changeable length, the normally used similarity measures based on the fixed-dimension vectors are not relevant. Earth Move’s Distance Algorithm is used to finding similarity measurement.

4. DATASET

Following are the datasets for near duplicate image matching:

A. Esri Data Set

It includes 17,156 envelope images taking from real-life Chinese handwritten mail pieces, along with which 600 images are arbitrarily taken elsewhere to construct the query set, and the remaining 16,556 images are unrelated ones. Afterwards, dissimilar variations are apply to each query image for creating its near-duplicates, which entail rotating the image from 10° to 180° by step size of 10°, scaling the image with dissimilar ratios from 20% to 90% by step size of 10%, and ever-increasing/decrease the brightness of the image from 20% to 90% by step size of 10% as well.

B. Euw2 Data Set

This data set involves 623 typewritten document images scanned from technical journals published in English. The setup of this data set is similar to that of the eSRI data set, with a query set of 320 randomly selected images.

C. PRI Data Set

The near-duplicate images in eSRI and eUW2 data sets given above are generated by hand, the motivebehind which is to test the performance of the proposed method with respect to different levels of image variations. Like eSRI data set, PRI comes from envelope images as well. However, rather than creating the near-duplicates by hand, the near-duplicate images in the PRI data set are obtained by taking independent pictures of a same envelope under different conditions. So they differ in several aspects.

4. PERFORMANCE METRICS

The performance of the face matching system can be calculated using verification time and accuracy.

A. Verification Time

Verification time is the period of time that a system takes to make decision.

B. Matching Accuracy

Matching accuracy is defined as below:

$$\text{Matching Accuracy} = \frac{\text{Number of correct matches} \times 100}{\text{Total Number of testing images}}$$

5 CONCLUSIONS

In this paper, I conclude that there are some ways for near duplicate image similarity. Variable length signature is proposed for near duplicate image matching. Patches are used for image representation. On basis of Different patches length get varies. Probabilistic center symmetric local binary pattern method is used for extract appearance of patch in image. To calculate the similarity between two images earth mover’s distance is used. Also the Clustering process is used find out similarities.

6. REFERENCES

[1] B. Wang, Z. Li, M. Li, and W.-Y. Ma, “Large-scale duplicate detection for web image search,” in Proc. IEEE Int. Conf. Multimedia Expo, Jul. 2006, pp. 353–356.

[2] Y. Ke, R. Sukthankar, and L. Huston, “Efficient near-duplicate detection and sub-image retrieval,” in Proc. ACM Int. Conf. Multimedia, 2004, pp. 869–876.

[3] O. Chum, J. Philbin, and A. Zisserman, “Near duplicate image detection: Min-hash and tf-idf weighting,” in Proc. Brit. Mach. Vis. Conf., 2008, pp. 493–502.

[4] S. Todorovic and N. Ahuja, “Region-based hierarchical image matching,” Int. J. Compute. Vis., vol. 78, no. 1, pp. 47–66, 2008.

[5] G. Meng, N. Zheng, Y. Zhang, and Y. Song, “Document images retrieval based on multiple features Combination,” in Proc. 9th Int. Conf. Document Anal. Recognit., Sep. 2007, pp. 143–147.

[6] S. Aksoy and R. M. Haralick, “Probabilistic vs. geometric similarity measures for image retrieval,” in Proc. IEEE Int. Conf. Compute. Vis. Pattern Recognition., Jun. 2000, pp. 357–362.