A Hybrid Model for Near-Duplicate Image Detection in MapReduce Environment

Nadiah Yusof, Amirah Ismail, Nazatul Aini Abd Majid

Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia (UKM) Bangi, Selangor Darul Ehsan, Malaysia

Abstract **– It has been proven that the large-scale image dataset is strictly complex in content-based image retrieval (CBIR) as the present strategies in CBIR might have difficulties in processing it. Other than this, near-duplicate images would possibly consume space, in which as an alternative can be used for storing other or unique images. In order to solve these problems, MapReduce has been used for speed-up filtering near-duplicate images. However, there is still a lack of accuracy in detecting near-duplicate images. Hence, this study has discovered that image features extraction by means of Principal Component Analysis (PCA) technique, which is primarily based on the matrix of image representation that will expand the similarity of detection. There is a need whereby PCA approach requires to be enhanced resulting from the lack of the extraction of features in Songket motives images. Therefore, this study proposes a new hybrid model that will integrate PCA with MapReduce for image feature extraction and clustering the large-scale image dataset in the cloud environment. In view of this, the present study employs the use of a qualitative experimental design model and goes through three main phases iteration: firstly, is the analysis and design phase, secondly is a development phase and lastly is testing and evaluation phase. However, this study focuses only on the analysis and design phase. The outcomes process of the empirical phase is followed by designing the algorithm and model according to the result of literature reviews. The expected results of this study is a proposed model and extract principal component elements of the large-scale image dataset using PCA, as well as boosting up time in filtering the images through MapReduce environment.**

Keywords **– Image, Image Retrieval, Near-Duplicate, PCA, Geometric, MapReduce**

DOI: 10.18421/TEM84-21
https://dx.doi.org/10.18421/TEM84-21

Corresponding author: Amirah Ismail,
Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia (UKM) Bangi, Selango, Malaysia
Email: amirahismail@ukm.edu.my

Received: 16 August 2019.
Revised: 02 November 2019.
Accepted: 09 November 2019.
Published: 30 November 2019.

1. Introduction

Detecting near-duplicate images or pictures in a large-scale image dataset within the cloud environment is to allow semantic computation [1],[2],[3],[4] and to detect images that are copied illegally on the web [5]. Nevertheless, to develop a near-duplicate image detection task in large-scale image dataset requires strenuous effort. This is due to the computation complexity in which majority of clustering algorithm is 0 (n2) [6],[7],[8] or even more. As a result, they cannot directly be used from large-scale image datasets for the identification of near-duplicate images [4]. Also, previous studies have shown that most clustering algorithms may not be able to support parallel computing because they are mostly data dependent, and the features applied in CBIR algorithm are somewhat in a large dimension which may result in an increase in computational complexity, thereby causing difficulties to use newly-developed solutions to new image dataset of various scenarios [4], [9].

Over the years, there have been few near-duplicate image detection techniques that have been developed for large-scale image dataset, including MapReduce clustering which utilizes locality constrain linear coding and integrated into maxIDF cut model [4] with min-hash technique and tf-idf matching multi clustering technique [10]. However, this research focuses more towards local features of image extraction. According to [11], focusing on one feature extraction, either by local or global, it will increase below required percentage compared to combined local and global image features extraction. As stated by [12], global and local images feature extraction will be interpreted differently. In terms of global descriptors, they are used to discover seed clusters with high precision, while local descriptors grow the seeds to cover good recall [4].

In view of this, this study's aim is to propose a model of near-duplicate image detection using PCA technique and applying it inside MapReduce environment to extract and detect near-duplicate image based on local and global features extraction of image. In order to achieve this, PCA technique is used to improve feature extraction based on matrix inside image [13]. Also, the implementation of MapReduce is to cluster and reduce the computational cost [14],[15],[16] and identify the near-duplicate images effectively. This is expected to achieve a lower complexity on image clustering and effectively scale with the sizes of image datasets [17].

The following Section 2 of this study will review previous literatures related to this study in order to understand all studies that has been done to support this research. Section 3, on the other hand, is the technique area that explains the methods used to design and develop this research model, while Section 4 explains the proposed and inner factors of the model in this study. Conclusion is thereby the last section of this study.

2. Research Background

Generally, researches in near-duplicate image detection is categorized into three focuses. These are detection technique, clustering of images and applying both similarity detection and clustering of images [10], [18].

Similarity measurement of features extraction has received much concentration for improvement by many researchers. [10] proposed one of the most popular technique called Bag of Words model, in order to extract local features and to use min-hash technique advantages. This is because, this technique employs the use of Jaccard coefficient as the similarity detection [19], [20]. Although, these approaches have displayed that they can be used on similarity measurement of near-duplicate of image detection, but it is still difficult to directly use them to solve the problem on a large-scale image clustering [4].

Recently, there have also been researches which focused on enhancing clustering of near-duplicates image part. They had developed an interesting algorithm to combine the local and global features in order to detect the near-duplicate images. However, there is a limitation inherent by global features in this technique [4].

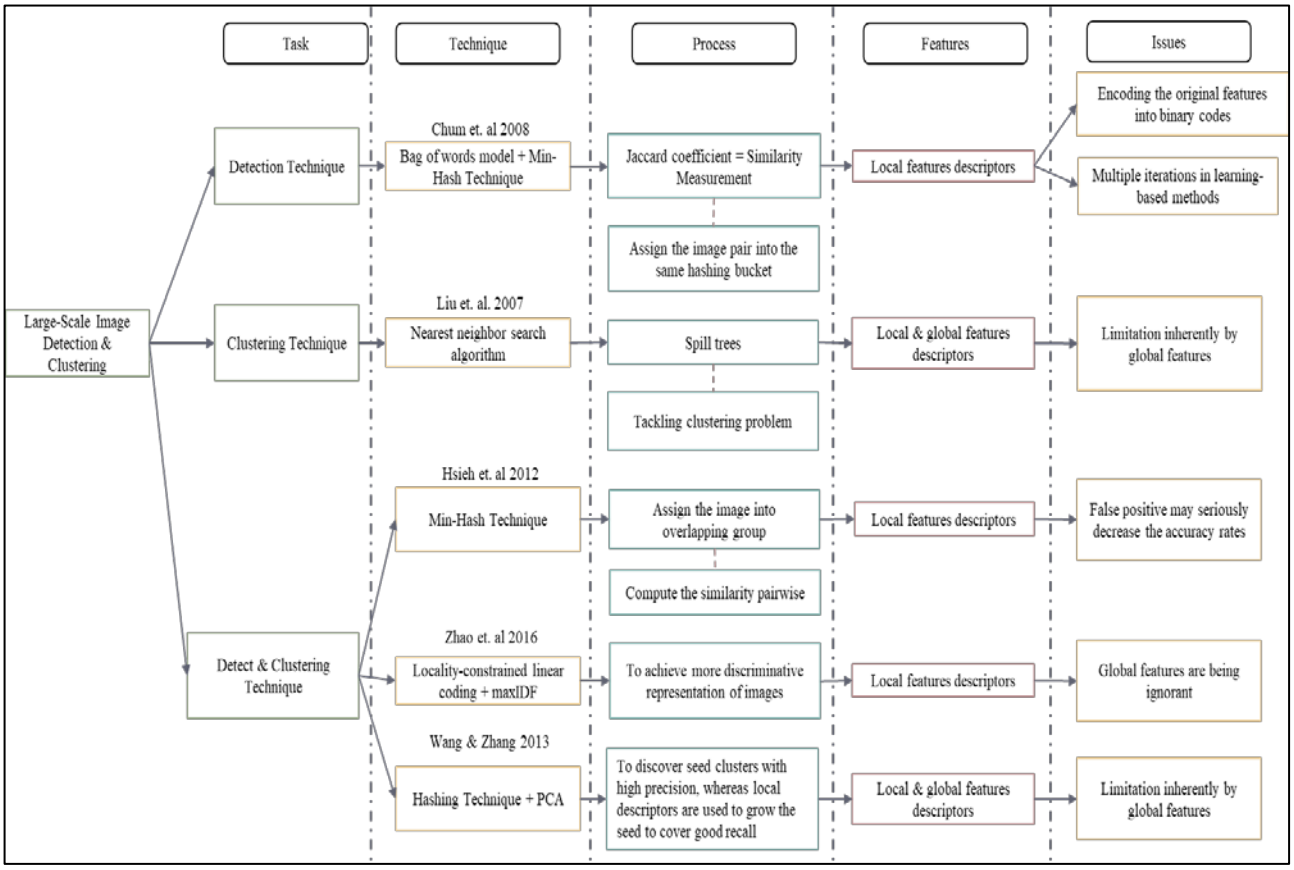

Figure 1. A summary of large scale near-duplicate image detection with identified issues.

Figure 1 briefly shows the several issues related to near-duplicate image detection in a large-scale image dataset. Additionally, some other researchers have focused on improving clustering and speed detection of images [4]. In this research, locality constrained linear coding and integrated with the maxIDF cut model will be used to tackle the problem of near-duplicate image identification and clustering. This is because of that the present study focuses more on the detection of local features extraction image representation. This is in line with the study of [11] who mentioned that similarity is approximately up to 54% detection if the focus is only on one side of image features extraction either local or global. However, the similarity will increase up to 65.5% by combining the two types of features extraction.

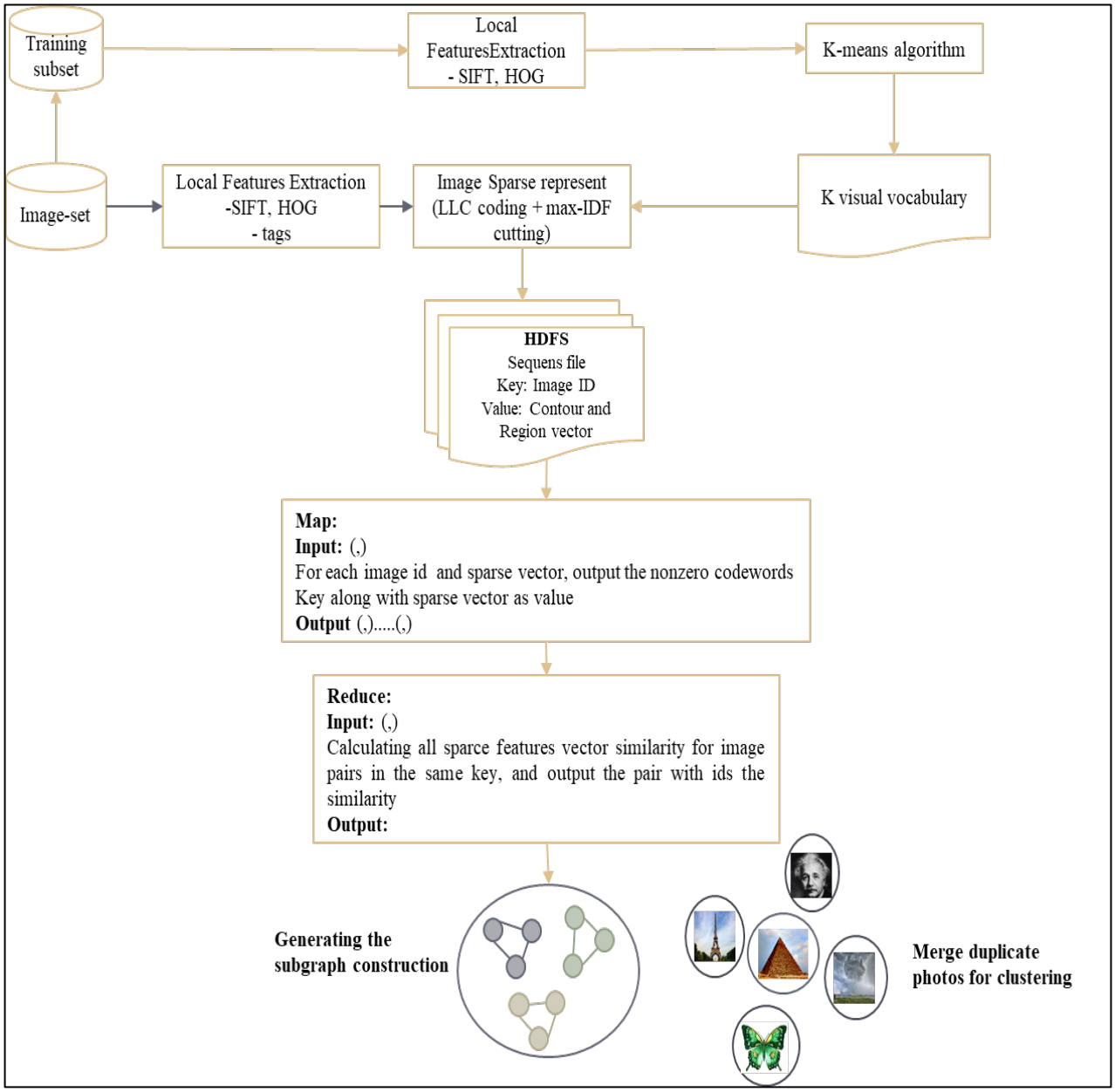

Figure 2. Hybrid model near-duplicate image detection using LLC + max-idf into MapReduce Environment. redrawn from [1]

The framework model as illustrated in Figure 2 represents the images which apply the use of LLC and max-IDF model, including the application of sparse features to segregate the image that is set into the overlapping subspaces bucket[4],[1]. This is because the model process examines near-duplicate image in each sparse feature vector of images are to combine with the similarity of image pair that will generate the near-duplicate cluster which is then followed by mapping and reduce the function through the use of the MapReduce framework. Therefore, after accepting features extraction phase direction, the activity will then relate with Map function processes otherwise known as <key, value>. The reduce function will collect and processes <key, value>, which is the pair that comes from the Map

function including the same key. Although, this framework model output is a generated sub-graph construction, which can merge duplicate photos for the same clustering in various buckets of images in line with their unique code or category [1].

This study adopts the use of Zhao framework model [1] which is enhanced through the use of an existing method (PCA technique) to extract local and global image features. The input of images, features extraction through PCA [21],[22],[23] Mapping and reduce function by MapReduce are the basic elements of this model [24]. The result of clustered images is then sent to the user to make the choice of whether the result should be stored or retrieved.

The result of the previous literature reviews detects the problem of both structures (local and global) features extraction in images that will be solved by employing the use of PCA technique. The reason is that, it is suitable to use on the inside of the image feature extraction due to the fact that PCA technique calculates both sides of the image structures based on the local and global features inside the images. [9] also stated that it can be used inside the framework model of the MapReduce.

3. Method

This method goes through three main phases iteration: firstly, is the analysis and design phase, secondly is a development phase and lastly is testing and evaluation phase. Figure 3 shows the interconnection between the three phases that have been implemented in this paper.

The Analysis phase is divided into two main categories of the process. The process is followed by reading 400 and above the article including books, in order to structure and prepared a systematic literature review on a variety of techniques have been applied and proposed in near-duplicate image detecting and clustering. The advantages of Prepared a systematic literature review can help finding the new significant contribution that can contribute into the body of knowledge in image retrieval research area. It is also important to look at the advantages and disadvantages of each proposed technique, and what techniques are appropriate to apply to the cultural heritage domain and various other domains in general. As can be seen, the Songket motif's image structure is more geometric structure, so the hybrid of the two techniques is assumed to coincide with the image structure described earlier.

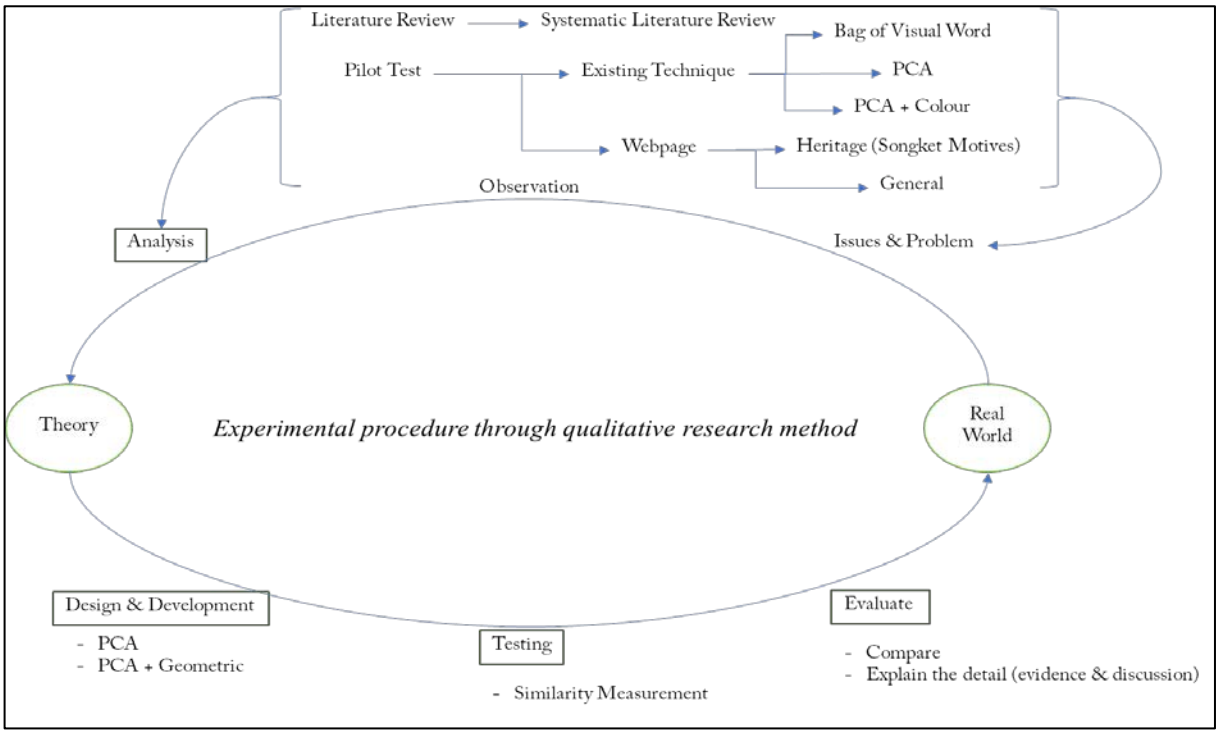

Figure 3. the focus of the study was conducted to examine identifiable gaps

4. Model

This section describes in detail what is related to the adaptation of the new improvement model as suggested by [4]. In this model, there are associated factors covered by a group of large-scale images dataset. To demonstrate that, this study adopts the use of images in the heritage area as a dataset, due to the fact that this domain possesses historical values that depict Malaysian identity [25], [26]. In this model, the procedure for features extraction is

conducted via the use of PCA techniques in defining the principal component (PC) of the images including differentiating between local and global features inside the image. Eq. 1 shows the algorithm that is used to calculate the PC inside the image.

$$x = TP' + E = \Sigma \sum_{r=1}^{R} t_r p_r + E \quad (1)$$

The eq. 1 explanation shows that T is a score symbol in the image, in which P is an eigenvector for the images, while E, on the other hand, is a noise inside the image. Also, the TP symbol is the main vicinity or principal component that is inside the image. This principal component is the first border in a massive dataset storage that will detect the similarity of near-duplicate image [27].

Additionally, the PCA technique is normally employed to extract Songket motives images that are fundamentally based on the main features inside the images which is otherwise known as the principal component (PC). Figure 4 illustrates how PCA works to extract the features of Songket motives. PCA will reduce the matrix dimension of the images. This is because the extraction of PC is left in PCA, in order to get the main matrix structure of the image framework so as to detect near-duplicate images. Furthermore, the MapReduce framework is applied to mapping and reduce similarity image, immediately after all the PC features extraction has been carried out.

The result of preliminary testing using the PCA technique on two sample images of Songket motives is shown in Figure 5 Through this outcome, it is seen that major framework of the image in a sequence is easily extracted and located in every part.

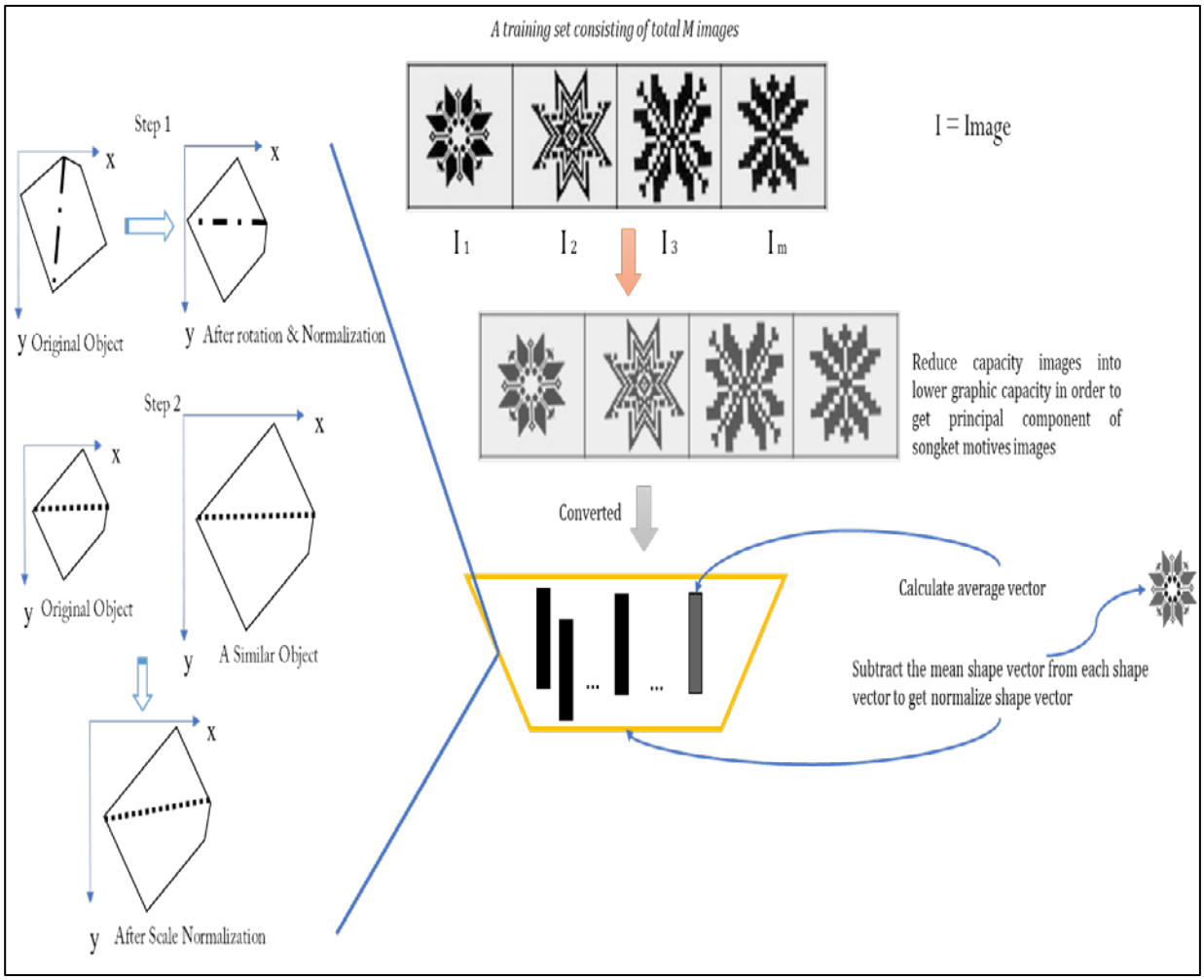

Figure 4. Extraction of PCs inside the images of songket motives

However, it is important to note that this result is shown without considering the implementation of MapReduce. Therefore, it is recommended for future study to combine the usage of PCA technique into MapReduce to increase similarity detection on large-scale image detection in cloud environment.

The proposed near-duplicate image detection model of this study is shown in Figure 6 in line with all elements that have been mentioned from the previous section. As a result, this model is divided into seven-step processes, in which the first step is an input of large-scale image dataset which desires to be

converted into data to ensure the computation of the PCA. Secondly, the use of linear aggregate was employed to represent vector image of principal component and extraction of the element's images into a matrix, in order to divide image dataset into the similarity matrix of image category. While the third step which is shown in this figure is to send outcomes of features extraction to HDFS.

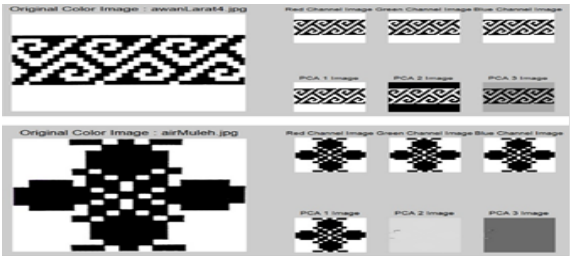

Figure 5. Result of PCA algorithm on two sample images.

Fourthly, the step taken in this model is HDFS which sends the information at once to Map function. The fifth step involves the map function to categorize the data in line with equal matrix, that is equal to the equal <key, value=""> element inside the data. The sixth step is taken in this study, which on the other hand, is to reduce function, in order to calculate all matrix features vector similarity for image pair in the identical <key, value=""> element. This is because the similarity step of image pair needs to merge near-duplicate images in the same category, so as to enable user to store or retrieve near-duplicate of images.

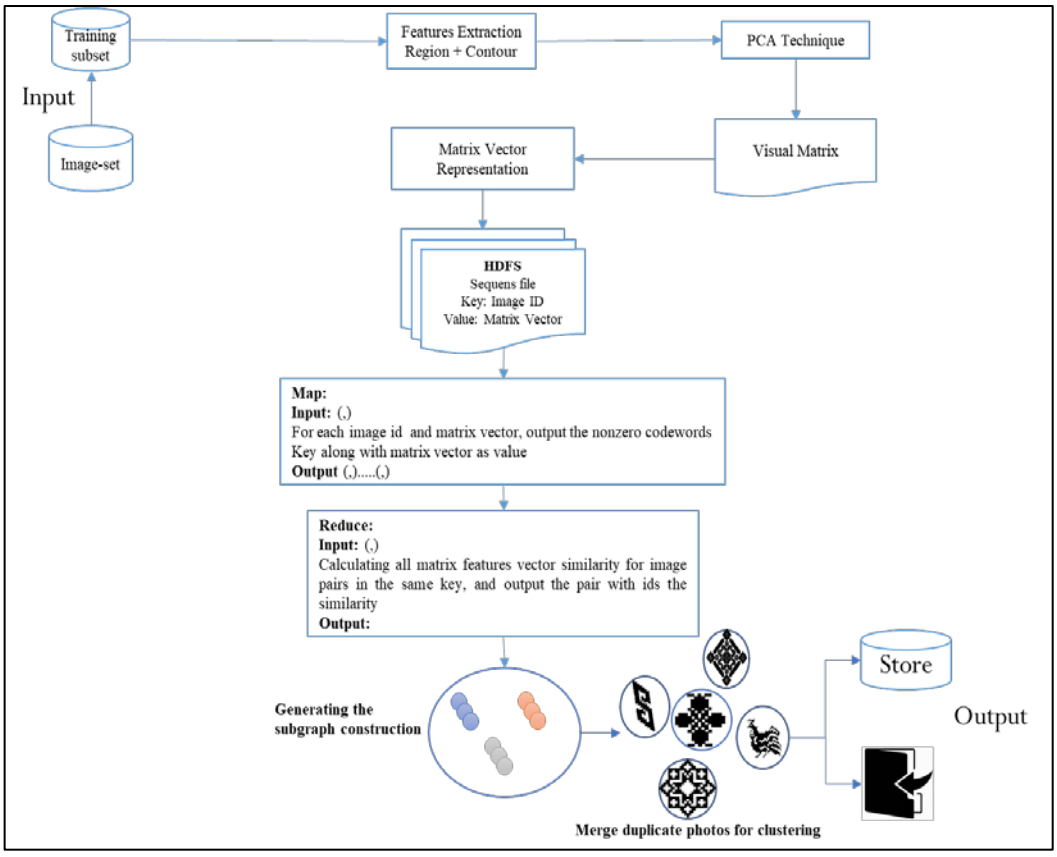

Figure 6. The proposed model of near-duplicate image detection in cloud environment

5. Conclusion

The detection of near-duplicate images and clustering the image is in line with their category which has comparable features that could assist the system and as a result. It will increase the percentage of recall and precision incomparable image retrieval and detection. The reason can be explained in the following manner: the way to increase the percentage of similar image detection is to focus on extra detail of two areas (local and global) image representation. This can be done through utilizing the hybrid PCA technique which is to be embed into the MapReduce environment model. This solution is expected to cater for the near-duplicate image detection issues that involve largely on local and global features extraction.

Acknowledgements

This research was supported by GUP-2017-077. Special thanks to our associates who gave the skill and knowledge that enormously assist in the exploration of this research.

References

[1] Tzeng, J. (2013). Split-and-combine singular value decomposition for large-scale matrix. *Journal of Applied Mathematics*, *2013*.

[2] Hackel, T., Wegner, J. D., & Schindler, K. (2017). Joint classification and contour extraction of large 3D point clouds. *ISPRS Journal of Photogrammetry and Remote Sensing*, *130*, 231-245.

[3] Asim, M. N., Wasim, M., Khan, M. U. G., Mahmood, N., & Mahmood, W. (2019). The Use of Ontology in Retrieval: A Study on Textual, Multilingual, and Multimedia Retrieval. *IEEE Access*, *7*, 21662-21686.

[4] Zhao, W., Luo, H., Peng, J., & Fan, J. (2017). MapReduce-based clustering for near-duplicate image identification. *Multimedia Tools and Applications*, *76*(22), 23291-23307.

[5] Foo, J. J., Sinha, R., & Zobel, J. (2007, July). SICO: a system for detection of near-duplicate images during search. In *2007 IEEE International Conference on Multimedia and Expo* (pp. 595-598). IEEE.

[6] Zhu, X., Li, X., Zhang, S., Xu, Z., Yu, L., & Wang, C. (2017). Graph PCA hashing for similarity search. *IEEE Transactions on Multimedia*, *19*(9), 2033-2044.

[7] Sokic, E., & Konjicija, S. (2016). Phase preserving Fourier descriptor for shape-based image retrieval. *Signal Processing: Image Communication*, *40*, 82-96.

[8] Jenni, K., Mandala, S., & Sunar, M. S. (2015). Content based image retrieval using colour strings comparison. *Procedia Computer Science*, *50*, 374-379.

[9] Wang, X. J., Zhang, L., & Liu, C. (2013). Duplicate discovery on 2 billion internet images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (pp. 429-436).

[10] Chum, O., Philbin, J., & Zisserman, A. (2008, September). Near duplicate image detection: min-hash and tf-idf weighting. In *Bmvc* (Vol. 810, pp. 812-815).

[11] Lisin, D. A., Mattar, M. A., Blaschko, M. B., Learned-Miller, E. G., & Benfield, M. C. (2005, September). Combining local and global image features for object class recognition. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)-Workshops* (pp. 47-47). IEEE.

[12] Wang, J., Tang, J., & Jiang, Y. G. (2013, October). Strong geometrical consistency in large scale partial-duplicate image search. In *Proceedings of the 21st ACM international conference on Multimedia* (pp. 633-636). ACM.

[13] G. R. Naik.(2017). *Advances in Principal Component Analysis (Research and Development)*. Kingswood: Springer.

[14] Li, R., Hu, H., Li, H., Wu, Y., & Yang, J. (2016). MapReduce parallel programming model: a state-of-the-art survey. *International Journal of Parallel Programming*, *44*(4), 832-866.

[15] Deshmukh, A. S., & Lambhate, P. D. (2016). A methodological survey on mapreduce for identification of duplicate images. *Int J Sci Res (IJSR)*, *5*(1), 206-210.

[16] Cai, F., & Chen, H. (2013, July). A MapReduce scheme for image feature extraction and its application to man-made object detection. In *Fifth International Conference on Digital Image Processing (ICDIP 2013)* (Vol. 8878, p. 88782D). International Society for Optics and Photonics.

[17] Broder, A. Z., Glassman, S. C., Manasse, M. S., & Zweig, G. (1997). Syntactic clustering of the web. *Computer networks and ISDN systems*, *29*(8-13), 1157-1166.

[18] W. Dong *et al.*, "High-Confidence Near-Duplicate Image Detection," in *Proceeding ICMR '12 Proceedings of the 2nd ACM International Conference on Multimedia Retrieval*, 2012, pp. 3304--3311.

[19] Hassanian-esfahani, R., & Kargar, M. J. (2018). Sectional minhash for near-duplicate detection. *Expert Systems with Applications*, *99*, 203-212.

[20] Ahasanuzzaman, M., Asaduzzaman, M., Roy, C. K., & Schneider, K. A. (2016, May). Mining duplicate questions in stack overflow. In *Proceedings of the 13th International Conference on Mining Software Repositories* (pp. 402-412). ACM.

[21] Sanguansat, P. (Ed.). (2012). *Principal Component Analysis: Multidisciplinary Applications*. BoD–Books on Demand.

[22] Kim, K. I., Jung, K., & Kim, H. J. (2002). Face recognition using kernel principal component analysis. *IEEE signal processing letters*, *9*(2), 40-42.

[23] Xu, J. L., & Gowen, A. A. (2019). Spatial-spectral analysis method using texture features combined with PCA for information extraction in hyperspectral images. *Journal of Chemometrics*, e3132.

[24] Wang, H., Zhu, F., Xiao, B., Wang, L., & Jiang, Y. G. (2015). GPU-based MapReduce for large-scale near-duplicate video retrieval. *Multimedia Tools and Applications*, *74*(23), 10515-10534.

[25] Jamil, N., Bakar, Z. A., & Sembok, T. T. (2006, July). Image retrieval of songket motifs using simple shape descriptors. In *Geometric Modeling and Imaging--New Trends (GMAI'06)* (pp. 171-176). IEEE.

[26] Jamil, N., & Bakar, Z. A. (2006, July). Shape-based image retrieval of songket motifs. In *19th Annual Conference of the NACCQ* (pp. 213-219).

[27] Davies, T., & Fearn, T. (2004). Back to basics: the principles of principal component analysis. *Spectroscopy Europe*, *16*(6), 20.