A Technique for Searching Duplicate Images in Large Scale Database

Dr. Kamini Solanki[1]
Parul Institute of Computer Application, Parul University[1]
Kamini.solanki@paruluniversity.ac.in[1]

Abstract - A technique which is use for searching duplicate image from large scale database is the active research now days. The photo de-duplication exercise will carried out in a large image database. De-duplication is carried out in different ways, which is based on biometric features and clusters. Features of the images are extracted using PCA Algorithm and then it divides into the cluster for fastest searching of duplicated images then comparison is done between input and database images. If images already exist in database, it will find the duplicate images from large scale database. Principal component analysis is used for better recognition result. The feature extraction is an essential step for image analysis, object representation, visualization, and many other image processing tasks. PCA is used for dimension reduction.

Index Terms— Principal component analysis (PCA), Cluster, Datamining

1. INTRODUCTION

Duplicate images introduce problems of redundancy in large image collections. The problem is acute on the web, where appropriation of images without acknowledgment of source is prevalent. In this paper, we present an effective features extraction and clustering approach for near duplicate images, using PCA techniques from large scale database.

There are three types of data are stores in Database.

1.1 Structured Data

This data format contains high degree of the organization. It is usually text file, relational databases which has rows and column that can be easily processed. It has advantage easy to entered, stored, analyzed and processed.

1.2 Unstructured Data

Unstructured data has no pre-defines structure or cannot be organized in predefined manner. It is not easy to understand. We have to process the data and understand it. It has majority machine generated data like images, photographs, videos, audios etc.

1.3Semi Structured Data

It has information which it cannot be stored in relational database but it has some organizational properties which make it easy to analyze. Semi structured data can be stored in relational database using some processes.

2. DATA MINING FUNCTIONALITIES

Data mining functionalities include classification, clustering, association analysis, time series analysis, and outlier analysis [20].

- Classification is the process of finding a set of models or functions that describe and distinguish data classes or concepts, for the purpose of predicting the class of objects whose class label is unknown.
- Clustering analyzes data objects without consulting a known class model.
- Association analysis is the discovery of association rules displaying attribute-value conditions that frequently occur together in a given set of data.
- Time series analysis comprises methods and techniques for analyzing time series data in order to extract meaningful statistics and other characteristics of the data.
- Outlier analysis describes and models regularities or trends for objects whose behavior changes over time.

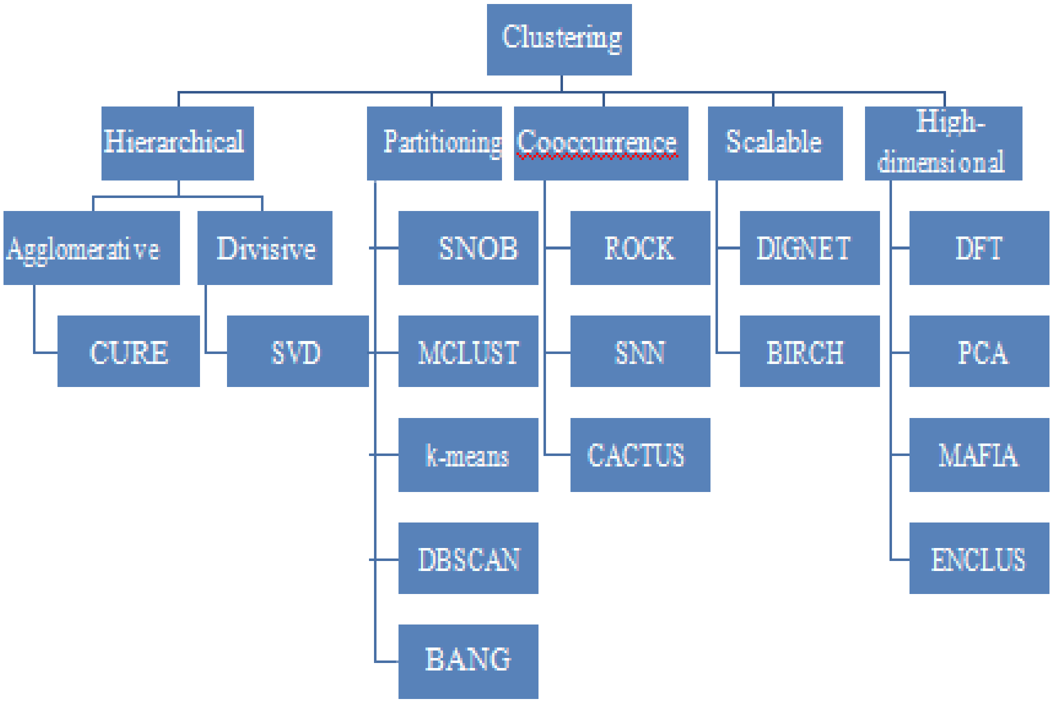


Fig 1. The research structure of clustering.

Clustering. Clustering algorithms [2] divide data into meaningful groups (see Figure 3) so that patterns in the same group are similar in some sense and patterns in dif ferent group are dissimilar in the same sense. Searching for clusters involves unsupervised learning [2]. In information retrieval, for example, the search engine clusters billions of web pages into different groups, such as news, reviews, videos, and audios. One straightforward example of clustering problem is to divide points into different groups [16].

- Hierarchical clustering method combines data objects into subgroups; those subgroups merge into larger and high level groups and so forth and form a hierarchy tree. Hierarchical clustering methods have two classifications, agglomerative (bottom-up) and divisive (top-down) approaches. The agglomerative clustering starts with one-point clusters and recur-sively merges two or more of the clusters. T he divi-sive clustering in contrast is a top-down strategy; it starts with a single cluster containing all data points and recursively splits that cluster into appro-priate subclusters [3, 4]. CURE (Clustering Using Representatives) [5, 6] and SVD (Singular Value Decomposition) [7] are typical research.
- Partitioning algorithms discover clusters either by iteratively relocating points between subsets or by identifying areas heavily populated with data. The related research includes SNOB [8], MCLUST [9], k-medoids, and k-means related research [10, 11]. Density-based partitioning methods attempt to discover low-dimensional data, which is dense-connected, known as spatial data. The related research includes DBSCAN (Density Based Spatial Clustering of Applications with Noise) [12, 13]. Grid based par-titioning algorithms use hierarchical agglomerationas one phase of processing and perform space seg-mentation and then aggregate appropriate segments; researches include BANG [14].

- In order to handle categorical data, researchers change data clustering to preclustering of items or categorical attribute values; typical research includes ROCK [15].

- Scalable clustering research images scalability prob-lems for computing time and memory requirements, including DIGNET [16] and BIRCH [17].
- High dimensionality data clustering methods are designed to handle data with hundreds of attributes, including DFT [18] and MAFIA [19].

3. PCA ALGORITHM

An Efficient method for image recognition is Principal Component Analysis (PCA). The PCA has been extensively employed for image recognition algorithms. It is one of the most popular representation methods for an image. It not only reduces the dimensionality of the image, but also retains some of the variations in the image data. The system functions by projecting image onto a feature space that spans the significant variations among known images. The significant features are known as “Eigen images”, because they are the eigenvectors (Principal Component) of the set of images. The Eigen Object Recognizer class applies PCA on each image, the results of which will be an array of Eigen values. To perform PCA several steps are undertaken: [1]

Stage 1: Subtract the Mean of the data from each variable (our adjusted data) subtraction of the overall mean from each of our values as for covariance we need at least two dimensions of data. It is in fact the subtraction of the mean of each row from each element in that row.

Stage 2: Calculate and form a covariance Matrix

Stage 3: Calculate Eigenvectors and Eigen values from the covariance Matrix Eigen values are a product of multiplying matrices however they are as special case. Eigen values are found by multiples of the covariance matrix by a vector in two dimensional space (i.e. a Eigenvector). This makes the covariance matrix the equivalent of a transformation matrix.

Stage 4: Chose a Feature Vector (a fancy name for a matrix of vectors) Once Eigenvectors are found from the covariance matrix, the next step is to order them by Eigen value, highest to lowest. This gives you the components in order of significance.

Stage 5: Multiply the transposed Feature Vectors by the transposed adjusted data The final stage in PCA is to take the transpose of the feature vector matrix and multiply it with the transposed adjusted data set (the adjusted data set is from Stage 1 where the mean was subtracted from the data).

4. DATABASE

The FEI image database is a Brazilian image database that contains a set of image images taken between June 2005 and March 2006 at the Artificial Intelligence Laboratory of FEI in São Bernardo do Campo, São Paulo, Brazil. There are 14 images for each of 200 individuals, a total of 2800 images. All images are colorful and taken against a white homogenous background in an upright frontal position with profile rotation of up to about 180 degrees. Scale might vary about 10% and the original size of each image is 640x480 pixels. All images are mainly represented by students and staff at FEI, between 19 and 40 years old with distinct appearance, hairstyle, and adorns. The numbers of male and female subjects are exactly the same and equal to 100. Figure 1 shows some examples of image variations from the FEI image database [21].

5. OBJECTIVES OF PROPOSED ALGORITHM

- Best features extraction of image using PCA
- Trying to find the same image within the large database. In this approach, the system returns the image which has nearest distant between the input image and database image using Ecudian distance measurement method.
- Clustering of the images based on the features for fastest searching of duplicate image.
- If image is new then store into database. If image is already exist in database then store only reference of that image so it will save the memory.

6. PROPOSED ALGORITHM

Algorithm 1: Store images into large scale database.

Step 1: Input colored images.
Step 2: Convert colored image into grayscale.
Step 3: Find the mean of the image.
Step 4: Subtract the mean from each column in the grayscale image.
Step 5: Execute PCA on result of step 4.
Step 6: Store images into database.

Algorithm 2: search images from large scale database.

Step 1: Input colored images for searching.
Step 2: Convert colored image into grayscale.
Step 3: Find the mean of the image.
Step 4: Subtract the mean from each column in the grayscale image.
Step 5: Execute PCA on result of step 4.

Step 6: Create the cluster of images using datamining technique that have similar or nearest decimal values.
Step 7: search the cluster that have similar decimal values of input image's decimal values.
Step 8: search image values which is nearest to that input image's decimal value in the cluster.
Step 9: Retrieved image from user database which has a minimum distance between input image and Database images using Euclidean distance measurement method.

7. PCA ON FEI DATABASE

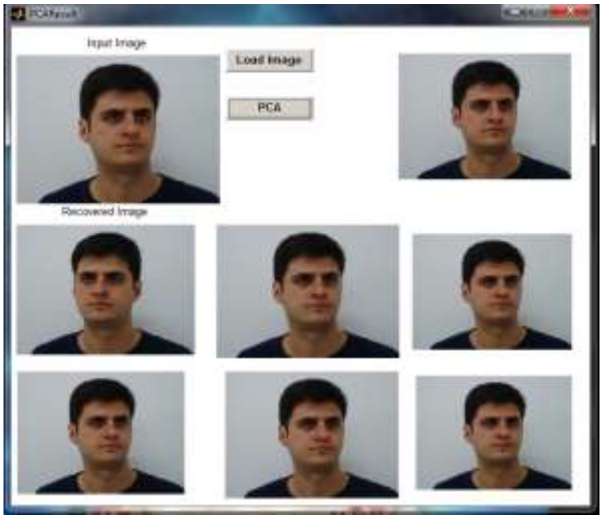
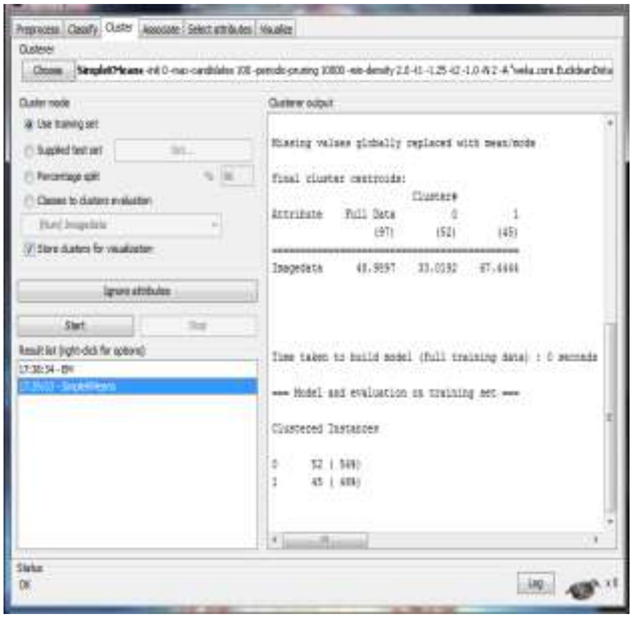


Fig. 2. PCA on FEI Database



SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 2 -A "weka.core.EuclideanDista

◉ Use training set
○ Supplied test set
○ Percentage split % 66
○ Classes to clusters evaluation
(Num) Imagedata
☑ Store clusters for visualization

17:30:34 - EM
17:35:03 - SimpleKMeans

```
Missing values globally replaced with mean/mode

Final cluster centroids:
                              Cluster#
Attribute     Full Data          0          1
                   (97)       (52)       (45)
=================================================
Imagedata       48.9897    33.0192    67.4444




Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       52 ( 54%)
1       45 ( 46%)
```

Status
OK

Fig. 3. Cluster in Weka

8. TESTING PARAMETERS

- Matlab R2012b is used for coding. Matlab is a programming language developed by MathWorks. Matlab stands for MATrix LABoratory. Matlab was originally written for the purpose of providing easy access. It started out as a matrix programming language. It is the high-level language and provides interactive environment to the developers. It is used in various disciplines including signal and image processing, communications, control systems etc. It integrates computation, visualization, and programming environment. It is also a modern programming language environment: it has a sophisticated data structure, containing inbuilt editing and debugging tools and support object-oriented programming. So it is an excellent tool for research work. Colored images are converted into the grayscale images because of better image processing.
- Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes. Found only on the islands of New Zealand, the Weka is a flightless bird with an inquisitive

nature. The name is pronounced like this, and the bird sounds like this. Weka is open source software issued under the GNU General Public License. We have put together several free online courses that teach machine learning and data mining using Weka.

9. CONCLUSION

The primary contribution of the paper is to find the effective similarity search in high dimensional data. By using this PCA and clustering methods, we accurately solve the near- duplication problem. PCA is used for global featured based approach. This approach will find the images which have highest features match from large scale database.

REFERENCES

[1] Dr. Prashant P. Pittalia ,Mrs. Kamini H. Solanki. An Invention Approach to 3D Image Recognition using Combination of 2D Texture Data and 3D Shape Data. International Journal of Application or Innovation in Engineering & Management (IJAIEM) Volume 2, Issue 11, November 2013.

[2] A. K. Jain and R. C. Dubes, Algorithms for Clustering Data, Prentice Hall, Englewood Cliffs, NJ, USA, 1988.

[3] K. Srivastava, R. Shah, D. Valia, and H. Swaminarayan, "Data mining using hierarchical agglomerative clustering algorithm in distributed cloud computing environment," International Journal of Computer Theory and Engineering, vol. 5, no. 3, pp. 520–522, 2013.

[4] P. Berkhin, "A survey of clustering data mining techniques," in Grouping Multidimensional Data, pp. 25–71, Springer, Berlin, Germany, 2006.

[5] S. Guha, R. Rastogi, and K. Shim, "CURE: an efficient clustering algorithm for large databases," ACM SIGMOD Record, vol. 27, no. 2, pp. 73–84, 1998.

[6] S. Guha, R. Rastogi, and K. Shim, "CURE: an efficient clustering algorithm for large databases," Information Systems, vol. 26, no. 1, pp. 35–58, 2001.

[7] M. W. Berry and M. Browne, Understanding Search Engines: Mathematical Modeling and Text Retrieval, vol. 17, SIAM, 2005.

[8] C. S. Wallace and D. L. Dowe, "Intrinsic classification by MML-the Snob program," in Proceedings of the 7th Australian Joint Conference on Artificial Intelligence, pp. 37–44, World Scientific, 1994.

[9] C. Fraley and A. E. Raftery, "MCLUST version 3: an R package for normal mixture modeling and model-based clustering," DTIC Document, 2006.

[10] A. Broder, L. Garcia-Pueyo, V. Josifovski, S. Vassilvitskii, and S. Venkatesan, "Scalable K-Means by ranked retrieval," in Proceedings of the 7th ACM International Conference on Web Search and Data Mining, pp. 233–242, Feburary 2014.

[11] Q. Li, P. Wang, W. Wang, H. Hu, Z. Li, and J. Li, "An efficient K-means clustering algorithm on MapReduce," in Proceedings of the 19th International Conference on Database Systems for Advanced Applications (DASFAA '14), Bali, Indonesia, April 2014, vol. 8421 of Lecture Notes in Computer Science, pp. 357– 371, Springer International Publishing, 2014.

[12] J. Agrawal, S. Soni, S. Sharma, and S. Agrawal, "Modification of density based spatial clustering algorithm for large database using naive's bayes' theorem," in Proceedings of the 4th Inter-national Conference on Communication Systems and Network Technologies (CSNT '14), pp. 419–423, Bhopal, India, April 2014.

[13] M. Ester, H. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD '96), pp. 226–231, Portland, Ore, USA, 1996.

[14] E. Schikuta and M. Erhart, "The BANG-clustering system: grid-based data analysis," in Advances in Intelligent Data Analysis Reasoning about Data, vol. 1280 of Lecture Notes in Computer Science, pp. 513–524, Springer, Berlin, Germany, 1997.

[15] S. Guha, R. Rastogi, and K. Shim, "ROCK: a robust clustering algorithm for categorical attributes," in Proceedings of the 15th International Conference on Data Engineering (ICD '99), pp. 512–521, March 1999.

[16] S. C. A. T homopoulos, D. K. Bougoulias, and C.-D. Wann, "Dignet: an unsupervised-learning clustering algorithm for clustering and data fusion," IEEE Transactions on Aerospace and Electronic Systems, vol. 31, no. 1, pp. 21–38, 1995.

[17] T. Zhang, R. Ramakrishnan, and M. Livny, "BIRCH: a new data clustering algorithm and its

applications," Data Mining and Knowledge Discovery, vol. 1, no. 2, pp. 141–182, 1997.

[18] E. Keogh, K. Chakrabarti, M. Pazzani, and S. Mehrotra, "Dimensionality reduction for fast similarity search in large time series databases," Knowledge and Information Systems, vol. 3, no. 3, pp. 263–286, 2001.

[19] H. S. Nagesh, S. Goil, and A. N. Choudhary, "Adaptive grids for clustering massive data sets," in Proceedings of the 1st SIAM International Conference on Data Mining (SDM '01), pp. 1–17, Chicago, Ill, USA, April 2001.

[20] Data Mining for the Internet of Things: Literature Review and Challenges, Hindawi Publishing Corporation , International Journal of Distributed Sensor Networks, Volume 2015, Article ID 431047, 14 pages, http://dx.doi.org/10.1155/2015/431047

[21] http://fei.edu.br/~cet/imagedatabase.html

[22] S. Ansari, S. Chetlur, S. Prabhu, G. N. Kini, G. Hegde, and Y. Hyder, "An overview of clustering analysis techniques used in data mining," International Journal of Emerging Technology and Advanced Engineering, vol. 3, no. 12, pp. 284–286, 2013.

[23] H. C. Koh, W. C. Tan, and C. P. Goh, "A two-step method to construct credit scoring models with data mining techniques," International Journal of Business and Information, vol. 1, no. 1, 96–118, 2006.

[24] N. C. Hsieh and L. P. Hung, "A data driven ensemble classifier for credit scoring analysis," Expert Systems with Applications, vol. 37, no. 1, pp. 534–545, 2010.

[25] E. Kambal, I. Osman, M. Taha, N. Mohammed, and S. Mohammed, "Credit scoring using data mining techniques with particular reference to Sudanese banks," in Proceedings of the 1st IEEE International Conference on Computing, Electrical and Electronics Engineering (ICCEEE '13), pp. 378–383, August 2013.

[26] Q. Liu, J. Wan, and K. Zhou, "Cloud manufacturing service system for industrial-cluster-oriented application," Journal of Internet Technology, vol. 15, no. 3, pp. 373–380, 2014.

[27] D. Maaß, M. Spruit, and P. de Waal, "Improving short-term demand forecasting for short-lifecycle consumer products with data mining techniques," Decision Analytics, vol. 1, no. 1, pp. 1–17, 2014.

[28] X. F. Du, S. C. H. Leung, J. L. Zhang, and K. K. Lai, "Demand forecasting of perishable farm products using support vector machine," International Journal of Systems Science, vol. 44, no. 3, pp. 556–567, 2013.

[29] C.-J. Lu and Y.-W. Wang, "Combining independent component analysis and growing hierarchical self-organizing maps with support vector regression in product demand forecasting," International Journal of Production Economics, vol. 128, no. 2, 603–613, 2010.

[30] H. Lee, S. G. Kim, H.-W. Park, and P. Kang, "Pre-launch new product demand forecasting using the Bass model: a statistical and machine learning-based approach," Technological Forecasting and Social Change, vol. 86, pp. 49–64, 2013.

[31] M. Chen, S. Gonzalez, V. Leung, Q. Zhang, and M. Li, "A 2G-RFID-based e-healthcare system," IEEE Wireless Communica-tions, vol. 17, no. 1, pp. 37–43, 2010.

[32] J. Liu, J. Wan, S. He, and Y. Zhang, "E-healthcare supported by big data," ZTE Communications, vol. 12, no. 3, pp. 46–52, 2014.

[33] M. Chen, Y. Ma, J. Wang, D. O. Mau, and E. Song, "Enabling comfortable sports therapy for patient: a novel lightweight durable and portable ECG monitoring system," in Proceedings of the IEEE 15th International Conference on e-Health Networking, Applications and Services (Healthcom '13), pp. 271–273, IEEE, Lisbon, Portugal, October 2013.

[34] J. Liu, Q. Wang, J. Wan, J. Xiong, and B. Zeng, "Towards key issues of disaster aid based on Wireless Body Area Networks," KSII Transactions on Internet and Information Systems, vol. 7, no. 5, pp. 1014–1035, 2013.

[35] M. Chen, "NDNC-BAN: supporting rich media healthcare services via named data networking in cloud-assisted wireless body area networks," Information Sciences, vol. 284, no. 10, pp. 142–156, 2014.

[36] M. Chen, D. O. Mau, X. Wang, and H. Wang, "T he virtue of sharing: efficient content delivery in wireless body area networks for ubiquitous healthcare," in Proceedings of the IEEE 15th International Conference on e-Health Networking, Applications & Services (Healthcom '13), pp. 669–673, Lisbon, Portugal, October 2013.

[37] J. Wan, C. Zou, S. Ullah, C.-F. Lai, M. Zhou, and X. Wang, "Cloud-enabled wireless body area networks for pervasive healthcare," IEEE Network, vol. 27, no. 5, pp. 56–61, 2013.

[38] Neel Borkar and Sonia Kuwelkar, Face Recognition Sysyem Using PCA,LDA & Jacobi Method, European Journal of Advances in Engineering and Technology, 2017, 4(5):326-331, ISSN:2394-658X.

[39] Muhammad Sharif, Farah Naz, Mussarat Yasmin, Muhammad Alyas Shahid and Amjad Rehman, Face Recognition: A Survey, Journal of Engineering Science and Technology Review 10(2)(2017)166-177 www.jestr.org , March 2017.

[40] Bruce Poon, M. Ashraful Amin and Hong Yan, PCA Based Human Face Recognition with Improved Method for Distorted Images due to Facial Makeup, Proceeding of the International MultiConference of Engineers and Computer Scientists 2017 Vol I, IMECS 2017, March 15-17, 2017, Hong Kong. ISBN: 978-988-14047-3-2, ISSN: 2078-0958(Print): ISSN: 2078-0966(Online).

[41] Riddhi A. Vyas, Dr. S.M.Shah, Comparison of PCA and LDA Techniques for Face Recognition Feature Based Extraction with Accuracy Enhancement, International Research Journal of Engineering and Technology (IRJET) www.irjet.net, Vol. 04, Issue. 06 June – 2017, e-ISSN: 2395-0056, p-ISSN:2395-0072.