



Hybrid Method for Detecting Duplicate Image by Using Image Retrieval Technique in Data Mining

Dr. S. Thavamani

Associate Professor, Department of Computer Application, Sri Ramakrishna College of Arts and Science [Formerly SNR SONS College], Coimbatore-6, Tamilnadu

ARTICLE INFO	ABSTRACT
Published Online: 09 April 2021	Duplicated images cause several problems in online sites, so these demand special attention. To address the disadvantages of frames copy detection, the Hybrid Method of Detecting Duplicate Image by Using Image Retrieval Technique in Data Mining was proposed. We use the new method of eliminating duplicates in this example. To address the disadvantages of frames copy detection, the Hybrid Method of Detecting Duplicate Image by Using Image Retrieval Technique in Data Mining was proposed. The new method of eliminating duplicates in this example has proposed. Using this method, you can get rid of frames that aren't relevant to the video. This makes for more precise and faster video retrieval with fewer duplicates. As a back end, this technique is implemented in C# and SQL. The findings are put to the test and compared to the current SIFT process. The results showed that the output improved accuracy while reducing storage space, computational time, and memory use.
Corresponding Author: Dr. S. Thavamani	
KEYWORDS: D-SIFT Method, Segmentation, Frames, Storage Space	

1. INTRODUCTION

This section has focused few of the data mining techniques which was helpful to image retrieval techniques to detect the duplicate image process.

A. *Data Mining*: Data mining refers to the extracting or “mining” knowledge from large amount of data. Data Mining

is also the research for valuable information in large volumes of data. In this proposed method we implement our method under web mining. Under web mining there are three concept among them, like Web usage mining, Web content mining, Web structure mining. [1]

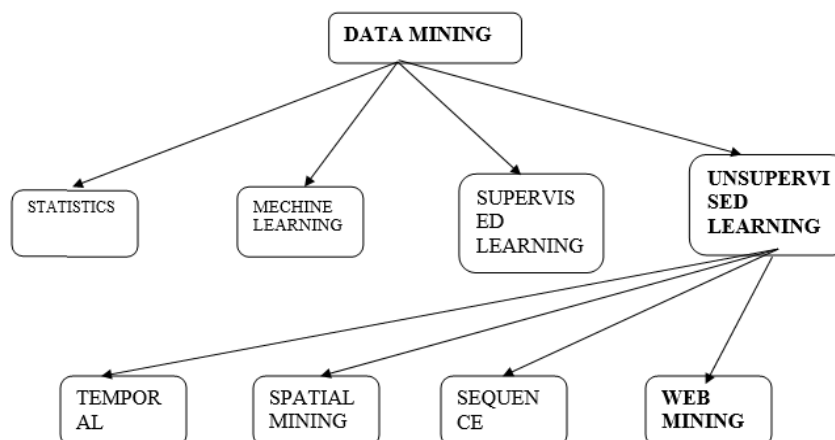


Figure-1 Overall View of Data mining

“Hybrid Method for Detecting Duplicate Image by Using Image Retrieval Technique in Data Mining”

B. Web Mining and Web Content Mining: Web mining is the use of Data Mining techniques to discover and extract information from Web documents and services automatically. The aim of web mining is to extract valuable knowledge from the World Wide Web and its use trends. This technology has adopted here to extract images over internet. The application of extracting valuable knowledge from the content of web documents is known as web content mining. [2] Text, image, audio, video, and other types of data make

up web material. A web page's content data is a set of information. It has the potential to provide useful and interesting trends regarding user requirements. Text mining, machine learning, and natural language processing are all applied to text documents. Text mining is another term for this form of mining. According to the content of the input, this method of mining scans and mines text, photographs, and groups of web pages. [3][4].

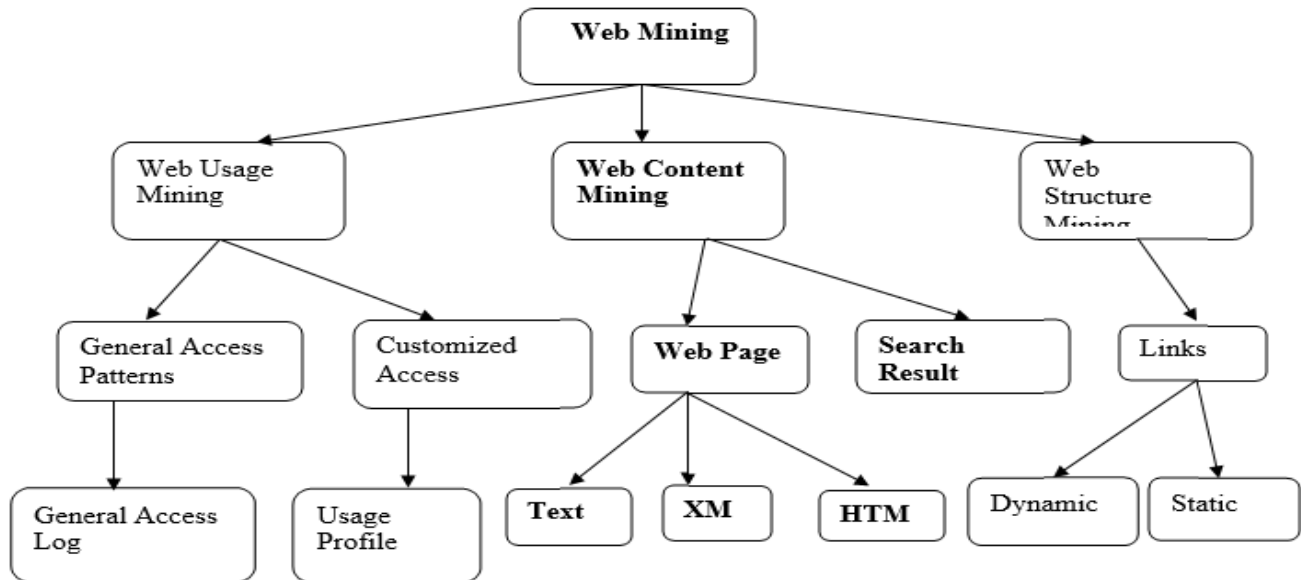


Figure-2 Web mining process

C. Image Retrieval method: Image mining requires that images be retrieved according to some requirement description. The requirement description can be classified into three levels of increasing complexity.

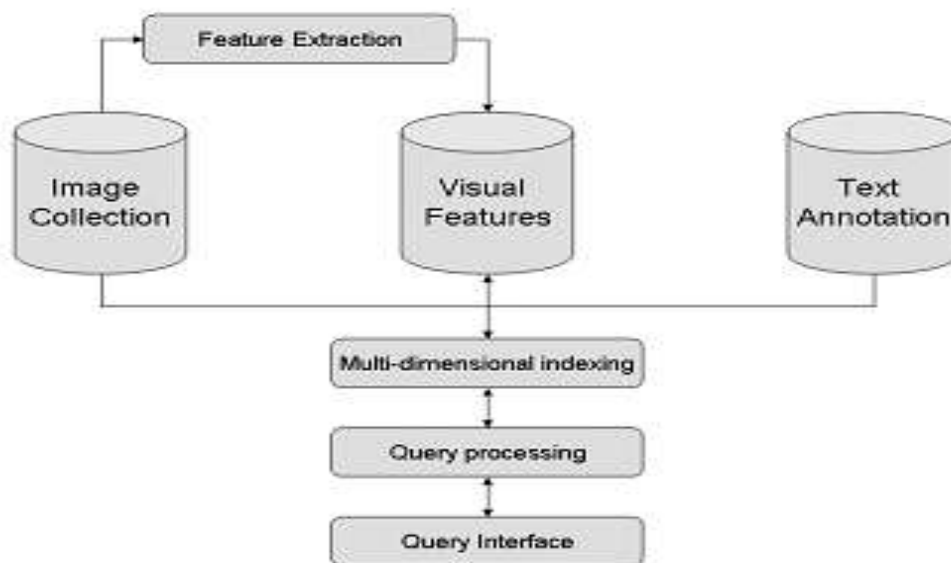


Figure-3 Process of Image Retrieval

(a) Level 1 comprises image retrieval by primary features such as color, texture and shape.

(b) Level 2 comprises image retrieval by derived or logical features like objects of a given type or individual objects or persons.

(c) Level 3 comprises image retrieval by obtaining attributes, involving a significant amount of high level reasoning about the meaning or purpose of the objects or scenes depicted.

In this paper, this attempts to identify the unique research issues in image mining. This will be followed by a review of what are currently happening in the field of image mining, particularly, image mining frameworks, state-of-the-art techniques and systems. [5]

2. LITERATURE SURVEY

With the rapid growth of digital devices, internet infrastructures, and web technologies, video data nowadays can be easily captured, stored, uploaded, and shared over the Web. Although general search engines have been well developed for searching videos over the internet is still a tough task. Typically, most Web search engines index only the metadata of videos and search through a text-based approach. However, without the understanding of media content, general search engines have limited capacity of retrieving relevant video information effectively. Thus, there is much scope to improve the retrieval performance of traditional meta-data based search engines through exploiting media content. Content-based video retrieval (CBVR) is becoming a promising direction for implementing reliable video search engines in future. This is retrieved from [6], [7].

Most existing content based video retrieval (CBVR) systems select video shots [1], homogeneous video regions, or semantic video objects. The major problem for using the semantic video objects for video content representation and

feature extraction is that automatic semantic video object extraction in general is very hard, if not impossible [8]

The paper [9] addresses the limitations of the novel multimodal and multilevel ranking scheme approach. The main idea has mainly employed text and visual information’s in the ranking tasks. In addition, they have suggested that will include additional information from other modalities. For example, this can study high-level concept detection techniques [10] and investigate some concept models to improve the ranking performance [11], [12]. It is likely that the proposed scheme could be improved by engaging additional information. Second, in the implementation, the fusion parameters are simply set to default parameters. In future work, this could study more intelligent solutions to determine the optimal parameters for multimodal fusion.

Third, in the current ranking solution, this considers only the query free approach for automatic search tasks. For future work, the query-class dependent weighting methods [12] can also be extended to the solution for further improving the retrieval performance. How to develop an effective query-class dependent algorithm using the graph based ranking framework will be an interesting research issue in future works. Moreover, this will apply the solution to solving other problems of content-based video retrieval, such as interactive video retrieval [13] and image/video annotation [14]. To these problems, this will explore the proposed multimodal and multilevel framework together with active learning techniques [15]. The table1.1 summarized the advantages and disadvantages of some **duplicate image** deducting techniques.

Table -1 Duplicate image deducting Techniques

REF.NO	YEAR	ALGORITHMS	ADVANTAGES	IMPROVED PARAMETER
[11]	2010	SI, HD, GISI PT	Extract the robust & discrimination feature from video frames.	1.Finding 2.location,3.Time, 4.Binary feature
[12]	2010	BBF, SVO	By considered key frame it provide exact video	1. Detection Time. 2. Related video
[13]	2007	TIRI, DCT	Based on time it matches the video. And also it matches the percentage of the video.	1. Detection time. 2. Ability to catch the percentage.
[14]	2012	FPEA, FSA, TIRI	To find the video segment for finger print to extract spatial and temporal informative video segment.	1. Robustness. 2. Pairwise. 3. Database search efficiency.
[15]	1996.	TIRI, SVD	Video catching process.	1. Pixel evaluation.
[16]	2004	ACM, BCM	Automatic video retrieval.	1. Query extraction.2. Query delay 3. Location finding.
[17]	2000.	MA, STA	Retrieve video based on motion analysis	1. Location 2.Time 3. Storage.
[18]	2006.	IMR, VCR	Image annotation refinement	1. Finding location. 2.Security
[19]	2001	SVM, MA	Support vector machine for learning.	1. Location identity 2. Time
[20]	2004.	TA, VRM	Image location finding scheme.	1. Location Accuracy

3. DISCUSSION OF PROPOSED METHOD AND RESULTS

This section has elaborated with the method named hybrid method with duplicate elimination of image for detecting duplicate image by using image retrieval technique with the results to achieve the goal.

A. Duplication Elimination Process: Duplication Elimination is a Compression method. From this we compress the frames, which are not relevant to next frame, in order to provide the video as effective one, with avoiding the irrelevant frames. For this process gain we need to

implement. D-SIFT method. With the use of description method, we can get some described information, about the frames with the use of this we knew easily that, which is the duplicate (or) irrelevant one. With the use of duplication elimination process, we can provide as effective one.

B. Architecture of Duplicate elimination: Image matching techniques are the techniques used to find existence of a pattern within a source image. Matching methods can be classified in two categories i.e. Area based matching techniques and feature based matching techniques.

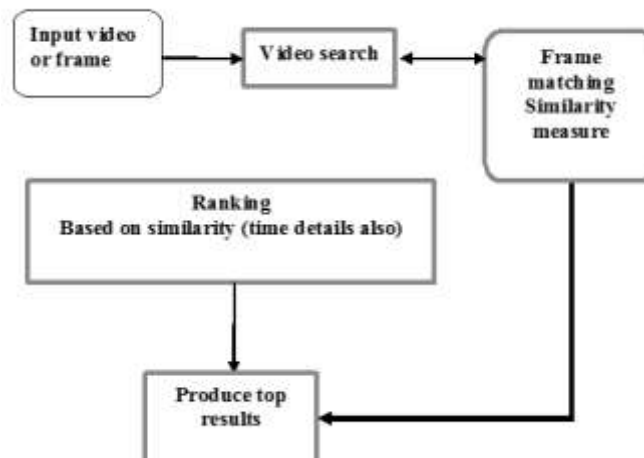


Figure-4 providing rank for each frames.

The query video gives as input, and each frame are matched with similarity measure. After the ranking process will provided the frames based on time. Based on these process, provide the Top result.

C. Proposed Algorithm and Flowchart

Step 1: Input the Video (or) Frames.

Step 2: In database it will search all the frames relevant to Query video.

Step 3: Matching the Frames.

Step 4: With the use of D-SIFT, Provide Ranking For each frames.

Step 5: Produced the Top Result.

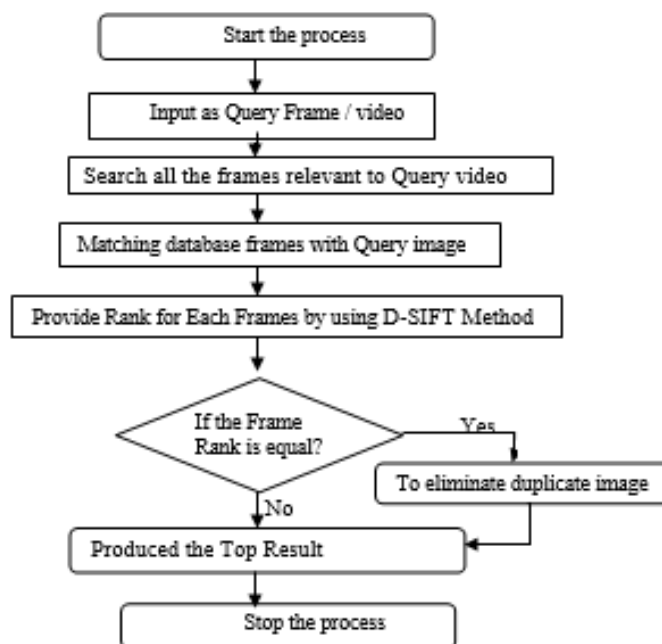
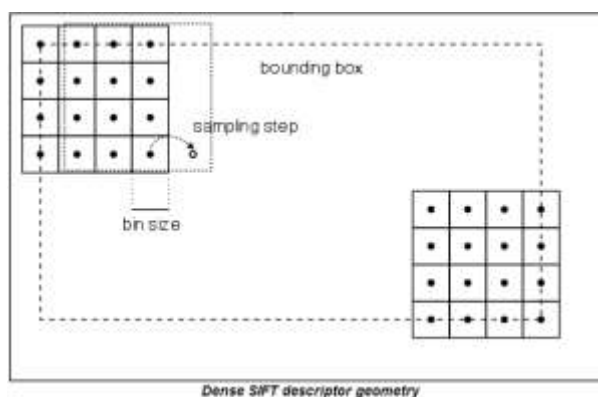


Figure-5 Flow chart for duplicate elimination.

Depending on query frames matching process is done. Next with the use of providing rank method we can provide a rank for each frames. Based on rank we can align the frames.



D. Result of Duplicate eliminate: In [16] similarity between two images was calculated with the help of enhanced Canberra distance. In [17] image similarity computed using L1 norm. Aksoy and Haralick [18] calculated similarity between images with the help of distances between feature vectors. According to their experiments, probabilistic

similarity measures performed better in terms of precision and recall compared to the cases where the geometric measures were employed. It was better to consider the feature distributions while deriving similarity measures. Image similarity computed using Hamming distance within the binary codes in [19] and [20].



Figure-6 Duplicate Omitting Process

4. CONCLUSION AND RESEARCH WORK

In this paper, the study addressed the problem of extracting the effective image related to the query image or video. The frame descriptors are found using a strong descriptor such as Dense SIFT learnt from a learning algorithm. Then, the similarity of video clips is calculated by the spatio-temporal pattern matching which includes spatio temporal dimension into the matching schema using the local and global descriptors. the experimental evaluation using traffic related videos and other synthetic dataset shows that the spatio temporal dimension is an effective feature to match video clips improves the accuracy and performance. In future the system may implement strong data mining technique to deal the storage problems. The current study does not deal with the storage based issues. The system may improve with the above constraint in future. In future the system may implement strong data mining technique to deal the security problems. Because the current study does not deal with the

Security based issues. If the user has stored some frames and video in particular repository path or database, there is no secure process to predict the frames and video in that particular database the system may improve with the above constraint in future.

REFERENCES

1. K. Jayamalini and M. Ponnavaikko, "Research on web data mining concepts, techniques and applications," 2017 International Conference on Algorithms, Methodology, Models and Applications in Emerging Technologies (ICAMMAET), Chennai, India, 2017, pp. 1-5, doi: 10.1109/ICAMMAET.2017.8186676.
2. Rolysent K. Paredes, Arnel C. Fajardo, Ariel M. Sison, "A Fuzzy-Based Dynamic Bandwidth Allocation Approach for Campus Area Networks", Engineering Technologies and

- Applied Sciences (ICETAS) 2018 IEEE 5th International Conference on, pp. 1-6, 2018
3. H. K. Azad and K. Abhishek, "Entropy measurement and algorithm for Semantic-Synaptic web mining," 2014 International Conference on Data Mining and Intelligent Computing (ICDMIC), Delhi, India, 2014, pp. 1-5, doi: 10.1109/ICDMIC.2014.6954238.
 4. J. M. Gago, C. Guerrero, C. Juiz and R. Puigjaner, "Web Mining Service (WMS), a Public and Free Service for Web Data Mining," 2009 Fourth International Conference on Internet and Web Applications and Services, Venice/Mestre, Italy, 2009, pp. 351-356, doi: 10.1109/ICIW.2009.58.
 5. C. V. Mahamuni and N. B. Wagh, "Study of CBIR methods for retrieval of digital images based on colour and texture extraction," 2017 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 2017, pp. 1-7, doi: 10.1109/ICCCI.2017.8117784.
 6. H. D. Wactlar, T. Kanade, M. A. Smith, and S. M. Stevens, "Intelligent access to digital video: Informedia project," *IEEE Computer*, vol. 29, no. 5, pp. 46–53, 1996.
 7. X. Wu, C.-W. Ngo, A. Hauptmann, and H.-K. Tan, "Real-Time Near-Duplicate Elimination for Web Video Search with Content and Context," *IEEE Trans. Multimedia*, vol. 11, no. 2, pp. 196-207, Feb. 2009.
 8. J. Yuan, J. Li, and B. Zhang, "Learning concepts from large scale imbalanced data sets using support process machines," in *Proc. 14th ACM Int. Conf. Multimedia*, Santa Barbara, CA, 2006, pp. 441–450.
 9. Hoi, Steven CH, and Michael R. Lyu. "A Multimodal and Multilevel Ranking Scheme for Large-Scale Video Retrieval." *IEEE TRANSACTIONS ON MULTIMEDIA* 10.4 (2008): 607.
 10. A. Amir, G. Iyengar, J. Argillander, M. Campbell, A. Haubold, S. Ebadollahi, F. Kang, M. R. Naphade, A. Natsev, J. R. Smith, J. Tesic, and T. Volkmer, "IBM research trecvid-2005 video retrieval system," in *Proc. TRECVID Workshop*, Washington, DC, 2005.
 11. R. Yan, J. Yang, and A. G. Hauptmann, "Learning query-class dependent weights for automatic video retrieval," in *Proc. ACM Multimedia Conf. (MM 2004)*, 2004.
 12. S. Dagtas, W. Al-Khatib, A. Ghafoor, and R. Kashyap, "Models for motion-based video indexing and retrieval," *IEEE Trans. Image Process.*, vol. 9, no. 1, pp. 88–101, Jan. 2000.
 13. M. Christel and R. Yan, "Merging storyboard strategies and automatic retrieval for improving interactive video search," in *Proc. Int. Conf. Image and Video Retrieval (CIVR)*, Amsterdam, The Netherlands, 2007.
 14. C. Wang, F. Jing, L. Zhang, and H.-J. Zhang, "Image annotation refinement using random walk with restarts," in *Proc. 14th ACM Int. Conf. Multimedia*, Santa Barbara, CA, 2006, pp. 647–650.
 15. S. Tong and E. Y. Chang, "Support vector machine active learning for image retrieval," in *Proc. Ninth ACM Int. Conf. Multimedia (MM'01)*, 2001, pp. 107–118.
 16. G.-H. Liu and J.-Y. Yang, "Content-based image retrieval using color difference histogram," *Pattern Recognit.*, vol. 46, no. 1, pp. 188–198, 2013.
 17. C. Kim, "Content-based image copy detection," *Signal Process., Image Commun.*, vol. 18, no. 3, pp. 169–184, 2003.
 18. S. Aksoy and R. M. Haralick, "Probabilistic vs. geometric similarity measures for image retrieval," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2000, pp. 357–362.
 19. 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.
 20. F. Zou et al., "Nonnegative sparse coding induced hashing for image copy detection," *Neurocomputing*, vol. 105, no. 1, pp. 81–89, 2013
 21. Dr. Anna Saro Vijendran ,2012. S.Thavamani "ANALYSIS STUDY ON CACHING AND REPLICA PLACEMENT ALGORITHM FOR CONTENT DISTRIBUTION IN DISTRIBUTED COMPUTING NETWORKS", *International Journal of Peer to Peer Networks (IJP2P)* Vol.3, No 6, November 2012
 22. U.Sinthuja, Dr.S.Thavamani "Evaluating systems and tools for vulnerability study on multi-broker MQTT instances", *GEDRAG & ORGANISATIE REVIEW - ISSN:0921-5077, VOLUME 33 : ISSUE 04 – 2020.*