Duplicate video and object detection by video keyframe uisng F-SIFT

Tanvi Gadgi
Department of Information Technology,
Dattakala Group of Institute of Technology,
Savitribai Phule Pune University, Pune, India
tanvigadgi.gadgi44@gmail.com

Prof. Amrit Priyadarshi
Department of Information Technology,
Dattakala Group of Institute of Technology,
Savitribai Phule Pune University, Pune, India
amritpriyadarshi@gmail.com

***Abstract*: To describe and detect various features in images scale-invariant feature transform can be used efficiently. Initially from a set of reference images SIFT key points of objects are extracted and stored in a database. An object in a new image can be recognized by individually balancing each feature from the new image to this database and find features for candidate matching. As an effective local SIFT can be employ as a key point descriptor for its invariance to, lighting, scale, and rotation changes in images. Since SIFT is not flip invariant, flip invariant SIFT is proposed. These F-SIFT is demonstrated to detect large scale copy videos, object detection as well as recognition. It requires to extract all the frames from query video and videos in dataset for similarity matching, time complexity of f-SIFT is more, So to eliminate such limitation we have proposed dual threshold technique.**

Our system will remove redundant video frames by applying auto dual threshold method. So there will be no need to perform extraction of features and matching of sequence with all video frames. Redundant frames are removed by making segments of video. Only the key frames are extracted for matching proposes. Here we are using two thresholds. One is for identifying immediate changes of visual information of extracted frames and other for detecting regular changes of visual information of extracted frames. Threshold values are decided according to the content of video. We are extracting three frames like first frame, last frame and key frame from video segment. By using average feature value of all the frames in the segment, key frames are decided. For matching propose key frame is used and remaining two frames are used to detect segment location.

Keywords - Threshold, key frames, SIFT, video segmentation

I. INTRODUCTION

With the quick growth in the multimedia technology and web, we are able to access and store huge numbers of video data quickly. That is huge numbers of video clips are transmitted, searched and stored on web. Some statistics of the YouTube shows that, there are about tons of user generated video clips are submitted to YouTube every minute. According to BBC motion gallery, it includes over 2.5 million hours of video contents. Among the huge numbers of video clips there exist a huge numbers of duplicated and near copied video clips. It is reported that about 27% video clips in videos search results obtained from yahoo, YouTube and Google video clips are copied or near copied duplicates of a popular version. For particular queries, the redundancy can be as high as 93%.A copy video clips can be divided into two types Duplicate Videos and Nearly Duplicated Videos. Duplicated Video will be extracted video duplicates that can be quickly detected. Near Duplicated video are transformed video clips and recognition of such duplicates is challenging. So we can define video clips copy as, it is a segment of video derived from another video clip usually through various transformations such as deletion, addition, cam coding and modification.

There is need to identify such duplicate videos for copyright propose. Scale invariant feature transform can be used to extract various features of videos. The beauty of SIFT is mainly because of its invariance to various picture transformations like: displacements, scaling, rotation and lighting changes of pixels in a local region. SIFT is normally calculated over a local silent region which is positioned by rotated and multi-scale detection to its leading orientation. The descriptor is invariant to both rotation as well as scale. In addition, due to rotation and spatial partitioning it is insensible to lighting, small pixel displacement and color. But the fact is that SIFT is not flip invariant.

Flipping video is one of the mostly used tricks to create duplicate videos. There are two types of flip operations vertical flipping and horizontal flipping. Vertical flipping is used mostly since it will not affect change into the content of video. Also the video of the same object taken from opposite direction can flip videos. To avoid this limitation F-SIFT is introduced. It improves the SIFT with flip invariant attribute. It also can be used for detection and recognition of similar objects from duplicate videos. F-SIFT require extracting all the frames of the query video and videos in dataset. So the time complexity for copy detection is much more. So we have introduced system which can help to reduce this time complexity.

Our system will remove redundant video frames by applying auto dual threshold method. So there will be no need to perform extraction of features and matching of sequence with all video frames. Redundant frames are removed by making segments of video. Only the key frames are extracted for matching proposes. Here we are using two thresholds. One is for identifying immediate changes of visual information of extracted frames and other for detecting regular changes of visual information of extracted frames. Threshold values are decided according to the content of video. We are extracting three frames like first frame, last frame and key frame from video segment. By using average feature value of all the frames in the segment, key frames are decided. For matching propose key frame is used and remaining tow frames are used to detect segment location.

II. RELATED **WORK**

Frame descriptors play very crucial in duplicate video detection performance. Here Mei-Chen Yeh [1] presents a compact, effective descriptor, which is theoretically efficient and simple. The descriptor is build by programming pair-wise correlations within a frame. Even if image property alter due to transformations, this descriptor uses the inside construction of a video frame, which makes it strong to attacks based on signals such as blurring , color changes as well as contrast enhancement to certain attacks such as scaling of frames.

Law-To et al. provided a relative research for video copy recognition and determined that, for small changes, temporal ordinal measurements are effective, while methods based on local features illustrate more appealing results in conditions of robustness [2]. However, Thomee et al. conducted a large-scale evaluation of picture duplicate recognition systems and achieved a somewhat different summary. Their selected technique that used interest factors conducted badly0 due to its lack of ability to find similar sets of factors between duplicates [3]. They determined that either a simple average technique or the retina technique works the best. To design a practical duplicate recognition program which satisfies the scalability specifications, a lightweight, frame-level descriptor that retains the most appropriate information, instead of just places of interest point descriptors, is suitable [4]. Furthermore, frame level descriptors are easily incorporated into fast detection frameworks such as the one provided in [5].

In actuality, SIFT [7] descriptors and HOG [6] descriptors are both well-designed gradient histograms used in object classification and detection tasks. In these tasks, to achieve robustness against objects' flipping and rotation is of great importance. Various extensions of SIFT descriptors have been proposed to address such transforms. RIFT [8] achieves flip and rotation invariance by dividing a region along the log polar direction instead of using 4 × 4 grids, which, however, is less distinctive than original SIFT. In contrast, FIND [9] and MIFT [10] preserve the distinctiveness while they are obtained by sorting original SIFT descriptors according to their relative magnitude, which is invariant under flip and rotation. Similarly, F-SIFT [11] infers the reference direction of SIFT based on the dominant curl associated with a local region, and performs selective flipping on the region before descriptor calculation. Contrary to those methods, which preserve original SIFT properties, MI-SIFT [12] applies direct flip-invariant transform to SIFT to produce flip-invariant descriptors. An overview of these flip and rotation invariant SIFT extensions is presented in [13].

MI-SIFT [14], instead, functions straight on SIFT while transforming it to a new descriptor which is flip invariant. This is achieved by clearly determining the categories of function components which are cluttered placed due to flip function. MI-SIFT brands 32 of such categories and symbolizes each group with four instants which are flip invariant. Nevertheless, the descriptor depending on time is not discriminative. As reported in [14], this outcomes in more than 10% of related performance degradation than SIFT when no-flip transformation happens.

In [15], the authors follow HSV to represent their key frames and further generate videos clip signature by cumulating all the key frames in it. This reflection accomplishes fast recovery speed as well as high precision in their dataset. However, a restriction is that global function based techniques generally become less efficient in handling video duplicates with layers of modifying cosmetics [16]. At the same time, global function centered techniques rely intensely on the selected function types.

The system that we have proposed will remove redundant video frames by applying auto dual threshold method. So there will be no need to perform extraction of features and matching of sequence with all video frames. Redundant frames are removed by making segments of video. Only the key frames are extracted for matching proposes. Here we are using two thresholds. One is for identifying immediate changes of visual information of extracted frames and other for detecting regular changes of visual information of extracted frames. Threshold values are decided according to the content of video. We are extracting three frames like first frame, last frame and key frame from video segment. By using average feature value of all the frames in the segment, key frames are decided. For matching propose key frame is used and remaining tow frames are used to detect segment location.

III. IMPLIMENTATION DETAILS

a. *System Architecture*

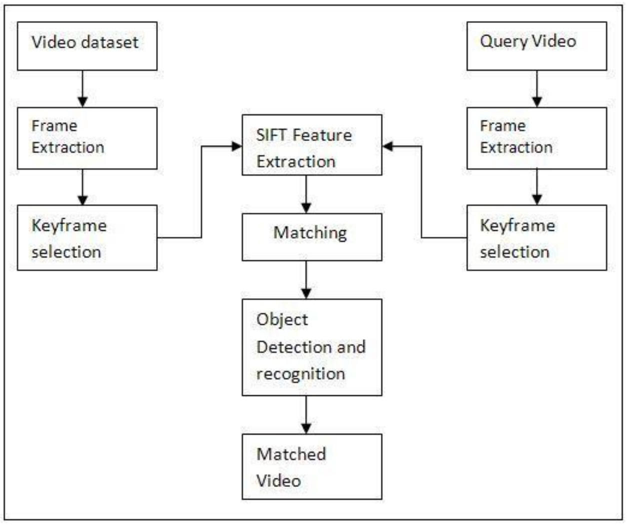

Fig 1: System Architecture

Above figure (Fig: 1) denotes system architecture of our proposed system.

This system works in two phases training phase and testing phase. IN training phase video dataset is taken as input. Different frames from all those videos are extracted sequentially. Similar frames for the set of frames are neglected and only the key frames are taken out.

Same process is done for input query video in testing phase. From those extracted key frames SIFT features are extracted for matching purpose. Similar objects are detected and recognized from matching frames. And finally similar matched videos are detected.

b. *Algorithm*

Step1: Query video is converted to number of frames.

Step2: Key frames are extracted from frames of query videos by using auto dual threshold method.

Step3: SIFT Features are extracted from each key frame.

Step4: Matching of query video key frames is done with the original video.

Step5: Detection and recognition of copied object is done.

Step6: Matched video is extracted from the database.

c. *Mathematical model for proposed work*

The lower and upper thresholds, T_L and T_U, are premeditated according to both *Rn k* and *Oa*(*n*) as following:

$$R_k^n = \frac{N_n}{N_n}, n = k+1, \dots, k+N$$

Let f_1… f_n… f_{Num} indicate the frames of the video and f_k be a key-frame. f_n indicate a frame among f_k and f_k+N,

$$T_L = \begin{cases} T_{ref} + \frac{R_k^n - T_{ref}}{4} & \overline{O_a} < 0.4 \\ T_{ref} + \frac{R_k^n - T_{ref} + 2\overline{O_a}}{4} + 0.1 & \overline{O_a} \geq 0.4 \end{cases} \quad \dots (1)$$

$$T_U = \begin{cases} T_{ref} + \frac{R_k^n max - T_{ref}}{2} & \overline{O_a} < 0.4 \\ T_{ref} + \frac{R_k^n max - T_{ref} + 2\overline{O_a}}{2} + 0.1 & \overline{O_a} \geq 0.4 \end{cases} \quad \dots (2)$$

$$\overline{O_a} = \frac{Num}{N+1} \sum_{n=k}^{k+N} \frac{dO_a(n)}{dn} / \sum_{n=1}^{Num} \frac{dO_a(n)}{dn} \quad \dots (3)$$

Where $dO_a(n)$ d_n is the derivative of the accumulative occlusion area, T_{ref} is an empirical parameter. $O_a(n)$. Rn k_{max} is the maximum R_n k for the key-frame f_k.

d. *Experimental Setup*

The system is built using Java framework (version jdk 6) on Windows platform. The Netbeans (version 6.9) is used as a development tool. The system doesn't require any specific hardware to run; any standard machine is capable of running the application.

IV. RESULTS AND DISCUSSION

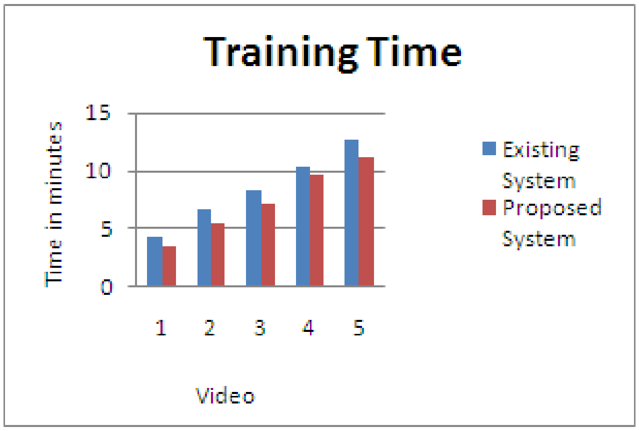

Fig 2: Time required for training videos

Above figure (Fig: 2) compares time required for training in existing system and our proposed system. We can see that time required for training in our proposed is less than that of existing.

Table 1: Time required for duplicate video and object recognition.

Video	Existing System	Proposed system
1	3.2	3
2	5.3	4.9
3	7.3	6.5
4	9.2	8.1
5	10.7	9.3

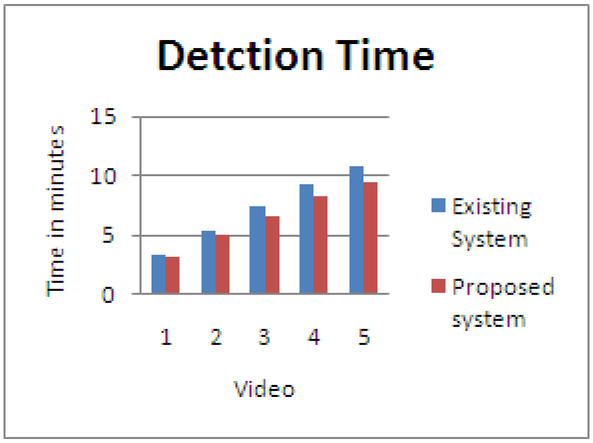

Fig 3: Time required for copy detection

Above figure (Fig: 3) denotes time required for duplicate video detection.

V. CONCLUSION

To find out duplicate videos from large video dataset numbers of systems are developed but those systems are infected by some limitations. As SIFT is not invariant to flip we have introduce F-SIFT which extracts key frames from query and videos in dataset to find duplicate videos and it finds and recognised similar objects in it for detection of duplicate video. Here we are using two thresholds. One is for identifying immediate changes of visual information of extracted frames and other for detecting regular changes of visual information of extracted frames. Threshold values are decided according to the content of video. Time complexity for duplicate video and object detection is less than other existing systems.

ACKNOWLEDGMENT

The authors would like to thank the researchers as well as publishers for making their resources available and teachers for their guidance. We also thank the college authority for providing the required infrastructure and support. Finally, we would like to extend a heartfelt gratitude to friends and family members

REFERENCES

[1] M.-C. Yeh and K.-T. Cheng, "A compact, effective descriptor for video copy detection," in Proc. Int. Conf. Multimedia, 2009, pp. 633–636.

[2] J. Law-To, L. Chen, A. Joly, I. Laptev, O. Buisson, V.Gouet-Brunet, N. Boujemaa, and F. Stentiford. "Video copy detection: a comparative study". In Proceedings of the ACM International Conference on Image and Video Retrieval, pages 371-378, 2007.

[3] B. Thomee, M. J. Huiskes, E. Bakker, and M. S. Lew. Large scale image copy detection evaluation. In Proceedings of the ACM International Conference on Multimedia InformationRetrieval, pages 59-66, 2008.

[4] S. Poullot, M. Crucianu, and O. Buisson. Scalable mining of large video databases using copy detection. In Proceedings of the ACM International Conference on Multimedia, pages 61-70, 2008.

[5] M. Yeh and K. -T. Cheng. Video copy detection by fast sequence matching. In Proceedings of the ACM International Conference on Image and Video Retrieval, 2009.

[6] Navneet Dalal and Bill Triggs, "Histograms of oriented gradients for human detection," in Proceed- ings of IEEE Computer Vision and Pattern Recognition (CVPR), 2005.

[7] David G Lowe, "Distinctive image features from scale invariant key points," International Journal of ComputerVision (IJCV), vol. 60, no. 2, pp. 91–110, 2004.

[8] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce,"A sparse texture representation using local affine regions,"IEEE

Transactions on Pattern Analysis and Machine Intelligence (PAMI), vol. 27, no. 8, pp. 1265– 1278, 2005.

[9] Xiaojie Guo and Xiaochun Cao, "FIND: A neat flip invariant descriptor," in Proceedings of IEEE International Conference on Pattern Recognition (ICPR), 2010.

[10] Xiaojie Guo and Xiaochun Cao, "MIFT: A framework for feature descriptors to be mirror reflection invariant," Image and Vision Computing, vol. 30, no. 8, pp. 546– 556, 2012.

[11] WL Zhao, CWNgo, et al., "Flip-invariant SIFT for copy and object detection," IEEE Transactions on Image Processing, vol. 22, no. 3, pp. 980–991, 2013.

[12] Rui Ma, Jian Chen, and Zhong Su, "MI-SIFT: mirror and inversion invariant generalization for sift descriptor," in Proceedings of ACM International Conference on Image and Video Retrieval (CIVR), 2010.

[13] Sreeraj M Asha S, "State-of-the-art: Transformation invariant descriptors," International Journal of Scientific and Engineering Research (IJSER), vol. 4, pp. 1994–1998, 2013.

[14] A. F. Smeaton, P. Over, and W. Kraaij, "Evaluation campaigns and TRECVid," in Proc. Int. Conf. Multimedia Inf Retr., 2006, pp. 321–330.

[15] X. Wu, A. G. Hauptmann, and C.-W. Ngo, "Practical elimination of near-duplicates from web video search," in Proc. ACM Multimedia, 2007, pp. 218–227.

[16] X.Wu, C.-W. Ngo, A.G. 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, 2009.