JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 753-759
https://publishoa.com
ISSN: 1309-3452

Real-Time Conceptual Video Interpretation for Surveillance Systems using Euclidean Norms

[1] LNC. Prakash K,

Associate professor, Department of C.S.E, CVR College of Engineering, Mangalpalli, Hyderabad.

[2] Chengamma Chitteti,

Assistant professor, Department of I.T, Sree Vidyanikethan College of Engineering, Tirupati.

[3] Dr. G. Rama Subba Reddy,

Professor, Department of C.S.E, Sai Rajeswari Institute of Technology, Proddatur.

[4] Dr. S. Saranya,

Assistant professor, Department of I.T, Hindustan Institute of Technology and Science, Chennai.

[1]klnc.prakash@gmail.com

[2]sailusrav@gmail.com

[3]subbareddy1227@gmail.com

[4]saran.aamec@gmail.com

Abstract

Information retrieval is intended to help people who are constantly looking for information. Partitioning the video into frames is the first stage in video information retrieval. Most video frames are brief and do not provide much information about the image content. On the other hand, scene border recognition or video fragmentation into scenes provides a better understanding of the video scene by clustering images based on similar image content. This paper is about video scene identification, specifically video formation mining for template matching with deep characteristics. The study proposed and created a workflow that included phases for frame extraction, finding similarities between consecutive frames, grouping frames, identifying key frames, and seeing detection by merging the relevant frames. Python's OpenCV generates the frames. The process is evaluated using scene identification metrics. The results show that scene detection and quality are significant, as measured by several criteria. In addition, we examined and studied current recognition and analysis criteria. Furthermore, our proposed methodologies have been thoroughly tested on various public scene video datasets, and they outperform some state-of-the-art approaches. This work's findings can be used to create real-time conceptual video interpretations.

Keywords: Frames, surveillance, clustering, Image recognition, Scene detection.

1. Introduction

A video scene, also known as a Logical Narrative Unit or just a tale unit, is a conceptually connected sequential series of optical frames that portrays and communicates an elevated notion like incident, theme, item, place, or movement and forms part of a video story. An incident, for instance, can be described as an event or scenario that happens in a specific location at a particular time, such as a home run in a baseball game, an actor's arrival

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 753-759
https://publishoa.com
ISSN: 1309-3452

on the platform, a vehicle accident on an expressway, and so on. According to these criteria, video scenes and activity identification locate all video frames from a given video that correlate to a specific occurrence. Because of the affordable and speedy internet, the number of video collections is experiencing exponential growth. The identification and searching of images have become more and more difficult. Because of technological advancement, people have high expectations. The primary video platforms, such as YouTube, Livestream, and Google, are spending heavily on efficient and intelligent classification and retrieval to keep respective portals appealing and addicted to viewers. The first step in processing videos for data retrieval is to divide them into pictures and retrieve sample images, referred frames, from every shot. These essential images are then utilized for searching, efficient categorization, scene development, and video classification. The main reason behind choosing keyframes is to decrease computation overhead. The video is a sequence of photographs recorded in a temporal sequence; for example, each video released on the internet is 30 frames per second or greater. The higher the frame rate, the greater the visual effect. Although very advanced technology, all structures can be analyzed in real-time scenarios, similar to action recognition as CCTV streams. After processing it for probable object recognition, it requires 0.5 to 1.5 seconds to identify items in a picture. The video is segmented into images, and similar images are joined to form action sequences in video scene fragmentation. Shots are continuous and unbroken series of video frames in which the content and camera do not vary [1]. There are two sorts of video sequences: sudden shootings and progressive shootings. A fast shift in the scene, such as an alteration of the presenter during a Media interview, seems to be an instantaneous shot border. At the same time, progressive shootings, like slides and dissolves, require many images to switch the picture.

Several scenes in recordings replicate in a relatively short period; when those photographs are merged, such groupings of pictures are referred to as scenes. When two performers are conversing, for instance, the camera constantly panning to each of them with a slight alteration in the backdrop. In a two-minute short video, there are occasionally 25-30 images. The study of merging similar or repeated pictures into one clip or splitting films into semantically or visually linked or identical clips is known as scene detection, also known as scene border recognition or video scene fragmentation. Traditional video segmentation for web pages and Discs is time-consuming and impractical while working with massive datasets. Automatic video fragmentation into frames and scenes has recently acquired popularity throughout business and scholars [2,5]. The task of segmenting a movie into meaningful portions is video scene detection. This is a critical first step toward successfully analyzing diverse video footage. In this paper, we provide a fresh description of this job as a generalized optimization method with a different standardized linear model to cluster successive pictures into scenes as optimally as possible. The suggested normalized stored procedure mathematical capabilities enable effective image recognition, even in complex real-world settings. The proposed approach segments images into sharp shot borders, which are then combined to generate the scenes based on similarities. Figure 1 depicts the suggested generalized chronological sequence diagram.

The following is how the article is structured. In Section 2, we address relevant research such as modern image and video categorization techniques and scene recognition. Subsection 3 relates to our suggested scene identification mechanism. We report experimental findings, and these findings are then examined in Portion 4, and conclusions are reached in Portion 5.

2. Review of literature

The collection and saving of organized relevant information, including tags, annotation, entities, and occurrences, is required for video comprehension and searching. Granular searching results are provided, assuming that movies are composed of thematically coherent pieces. As we indicated initially, investigators often approach semantic and structural mining challenges individually. We describe a video pattern mining technique in our study based on video semantics. This section will review previous studies aimed at picture and video annotations and scene identification technologies. Even though picture categorization and image identification tasks have been investigated for a lot longer, suitable universal techniques are still lacking. A few of the causes is that estimating classification performance is dependent on the quantity and quality of categories. The correctness and length of label sets are limited by human experience and language processing. As a result, the [4] deep network achieves only 50% efficiency for the Places dataset with 205 classes and

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 753-759
https://publishoa.com
ISSN: 1309-3452

95% for the Scene15 database [5]. The disparity in findings was explained in the study [4]. The authors demonstrated that increasing levels and the number of datasets improve classification performance. There is also an issue with label set size: Scene15, which is even less than Places205, produces improved outcomes due to fewer labels [6]. Public image categorization competitions give an excellent insight into cutting-edge algorithms.

The TRECVID [7] competitions are all on video analysis. Semantic indexing (SIN) is the essential component, where technologies recognize ideas on a per-frame basis. Extremely highly methods in the Large-Scale Visual Recognition Challenge (ILSVRC) [8], notably the identification and item positioning challenge ("Taster challenge"), employ region proposal networks (RPN) integrated with region-based convolutional networks (R-CNN). The solution presented by the Oxford Vision Geometry Group [9] is one instance. These approaches apply to both picture and semantic video identification. New databases created by laboratories and organizations drive the research and development of new methodologies: Places [4] and the current Places2 from MIT, Scene15 [5] again from Ponce group (the University of Illinois at Urbana-Champaign), ImageNet [10] from Stanford University, and many others. Organizations and companies use these resources to create exact deep networks for detecting, localization, and identification.

In the study [6], Torralba et al. developed a system for conceptual video annotation. Researchers used a prepared transition map to capture several inside and outdoor locales. The method produced per-frame annotation that included precise (office 400/628) and clustered (office) classes. Output reliability is dependent on prior transformation knowledge, and their technique works significantly worse than it. Del Fabro et al. [11] present a detailed survey of scene identification techniques. Designers use the same key terms that they use in their articles. A picture is a single picture from an image sequence. Shot—a continuing series of images or frames comparable in feature space and proximity metric. A scene (also known as a narrative unit, LSU) is a contiguous succession of shots reflecting a conceptually solid film section. According to a poll [11], the scene recognition job is viewed as a three-stage challenge. Images are divided into shots throughout the first phase (shot identification). The second stage is to choose frames to depict the shot. This is performed to lower the computational load. The third phase involves clustering images into action sequences based on a similarity metric and predictions about the film's design. The survey researchers say feature variation is often used for shot edge detection. The features are identified and altered with the issue region. RGB, HSV, or LUV color histograms [12] [13] [14], background comparison [15] [16], movement characteristics [17], edge ratio difference and SIFT [18], and spectral characteristics [19] are all often utilized features.

We may use a variety of strategies to choose critical frames. [13] describes essential frame selection procedures. The study discusses four approaches: two side edge frame identification, initial frame collection, intermediate frame choice, and chain collection. The initial frame of a photograph has consistently been selected in the chained approach. If we hit the threshold position from the preceding keyframe inside the shot [16], its next key frame is inserted. We like this strategy since it provides more unified images and allows us to cluster a more extensive range of photos. The research [20] describes an excellent scene-detection method for an anonymous news video. First, they partition the nameless news video into frames using shot recognition based on the hand-craft characteristic. Second, a convolutional neural network and host scene characteristics got integrated to determine the news channel. Lastly, they use the regularity of news footage to offer a novel scene recognition approach. Their team has put the proposed methodology to the test on recordings from CCTV4, CCTV13, Beijing News, Anhui News, and Shanghai News. In [21], shot sequencing is taken into account throughout sequence alignment. Researchers present an enhanced spectral clustering technique that calculates the number of clusters and applies the fast global k-means approach following the eigenvector generation of the similarity matrix in the clustering step to group the photos. The same spectral clustering approach is used to identify the critical frames of every shot, and empirical studies show that the technique they present there effectively summarizes the information for every shot. Experiments on TV shows and movies show that the suggested scene identification approach effectively identifies almost all the scene borders while maintaining an acceptable balance between recall and precision. In [22], researchers provide a unique configurability for animal patterns that improve the state-of-the-art. The whole

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 753-759
https://publishoa.com
ISSN: 1309-3452

technique is demonstrated using genuine video sequences of three various animals. They show which can follow and recognize the provided animal mechanically. The learnt architectures are used to distinguish animals using different datasets: photographs shot by industry professionals from the Corel repository and pictures from the Web offered by Google. On both data sets, they show decent efficiency. Designers demonstrate that, for the Google set, they can identify, locate, and restore part expressions from a collection that is difficult for object identification.

3. Proposed methodology

Object tracking technique linked to computer vision applications recognizes and characterizes things in digitized photos and videos like people, automobiles, and creatures. This procedure could classify single or several items in a digitized picture or video at an identical time. Object recognition will be about for an extended period, but it has become much more prevalent in various sectors than ever before. Object detection and tracking is the practice of finding moving objects in video streams utilizing the cameras across the period. Object recognition seeks to link objects of interest in successive video frames. Object recognition necessitates the position, structure, or characteristics of things in video frames. As a result, object recognition and categorization come before object tracking in a computer vision task. Object recognition is the initial stage in monitoring and is used to recognize or identify moving objects in the image. Tracking objects onto subsequent frames is a demanding or challenging operation in image analysis. Numerous issues might develop due to complicated object motion, asymmetrical object form, shadowing of the object to object and object to the scene, and actual processing needs. Object identification systems commonly employ derived characteristics and learning algorithms to characterize representations of an object or photos matching an object category. The proposed object identification is given in Figure 1.

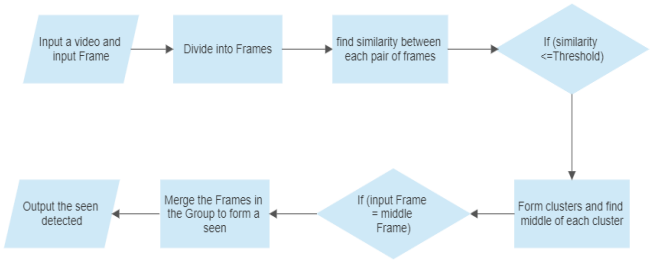

Figure 1: Proposed workflow for scene detection

Object class identification is concerned with categorizing objects into a given group or category. In contrast, object detection is involved localizing a single image of the object in digitized pictures or videos. Each item or object class has unique characteristics that distinguish it from the others, aiding in detecting the same or comparable things in the other photos or movies. One of the object detection applications is detecting a seen in the video that matches the given frame as an object. The corresponding methodology is presented in algorithm 1.

Algorithm 1:

Input: The video, the frame and the thresh hold.
Output: The video that matches the given input frame.

step 1. Input a video.
step 2. Divide the video into frames.
step 3. Find the similarity between each pair of consecutive frames.
step 4. Construct the groups(clusters) with respect to the threshold given.
step 5. Find the middle frame of each group.
step 6. Compare the input frame with each of the middle frames formed in step 5.
step 7. If match found merge the frames in the corresponding cluster.

Algorithm 2:

Input: an array of images.
Output: similarity between given pair of images.

1. Read the images in an array format as input data.
2. flatten each image thus that it must be a distinct 1-D array.
3. For every image, create a histogram using 1-D array.
4. determine the difference between two histograms, using Euclidean measure given in equation 1.
5. find the similarity of images using histogram difference.

In the above methodology, OpenCV is used to retrieve images from a given video. Now it will make an item of the type of Image Capture that will enable us to extract images from a video. Next, one must supply a phrase as input providing the location of the video in the system files and then use the read function on the VideoCapture

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 2, 2022, p. 753-759
https://publishoa.com
ISSN: 1309-3452

class to take a picture out from the video. Finally, the frames are stored in the stated place. We will utilize the technique described in algorithm 2 to determine the comparison among the pictures.

The algorithm two technique reads the picture data as an array; because the frames are coloured, there will be three streams for RGB values. Then we'll flatten frames so that each picture is a separate 1-D array. Once we have our picture data in an array, we will construct a histogram for every image by counting the occurrences of every pixel value in the image for every index 0 – 255. After finding the histograms, we use the L2-Norm or Euclidean Distance given in equation 1 to find the variance between the two histograms. Using the difference between the histogram of reference images, we can find the similarity between the images. This can be done by using the functions of OpenCV that are cv2.imread(), cv2.cvtColor(), cv2.calcHist(). The most similar frames (the distance between two consecutive frames is less than the thresh hold) are grouped to form the clusters.

$$E = \sqrt{\sum_{i=1}^{n}(Hist_1 - Hist_2)^2}$$
……………………………….. (1)

To find the centroid of all frames in a cluster, we pick the middle frame, which represents all frames in the group. Similarly, find all representatives of all clusters.

As said above, find the similarity between the given frame and each representative frame. If the most similar frame is found, merge all the corresponding cluster frames to form the required video. The programming code is developed to divide the video into images and identify suitable images for the given image with reasonable accuracy. It is proposed in future to increase the accuracy and need to develop the remaining parts of the above-said methodology as shown in the research [23, 24].

4. Experiments and Results

Scene boundary identification is made with cinematic and dramatic videos, and effectiveness analysis has been done with various movies and plays. For scene border recognition, F-score is employed as an evaluation metric. There isn't a standard dataset available. Two methodologies were being used to achieve ground truth: first-party ground truth and third-party ground truth. The researchers designed first-party and third-party verification, while third-party verification is gathered through specialists with adequate experience in shoots and scene boundaries. Figure 2(a), 2(b) and 2(c) depict the suggested system's performance. Our dataset is divided into two categories, each with its own set of films. One type of film is a cinematic film with a completely different setting and demanding effects with complicated scene movements. The second collection of data, on the other hand, comprises indoor play periodicals that are easier to segment than movie-making films due to their essential scenes and lack of complex effects, and that is why the value of the threshold is different for both datasets.

Figure 2(a): Input Frame Figure 2(b): Input Video Figure 2(c): Seen detection
Figure 2: Scene detection in indoor drama periodicals

5. Conclusion

Video segmentation is a crucial stage in the information retrieval of videos. The videos are divided into tiny parts using shot boundary identification. These little pieces don't provide enough information to understand the video's content or concept. Nevertheless, putting

comparable frames together gives a deeper understanding of the video, and this gathering can remain referred to as a video sequence. This study establishes a basis for scene boundary detection that employs cutting-edge exploring processes that are extensively utilized for image and video retrieving. We presented a method for conceptual video processing and developed a video framework analysis and markup workflow that includes steps for video frame extraction, shot grouping, and seen. For scene identification, video fragmentation is employed as the data generator. For scene identification, we would use the same data frames. Then we used metrics for both scene identification and description to examine the effectiveness of our workflow. We produced findings analogous to or improved than that stated in the original studies. As a result, we may conclude that the used methodology characteristics can not only be incorporated into traditional computer vision pipelines but also contribute new domain knowledge. We believe that our findings will be helpful in video exploration results and other videos semantic extracting applications. We also investigated different video categorization and annotation performance estimate criteria and suggested our approach. In the future, we intend to fine-tune our workflow by including characteristics such as other deep networks and data preprocessing techniques. We'll also have object and person recognition algorithms for our process to augment definitional knowledge.

REFERENCES

1. S. Lef'evre and N. Vincent, "Efcient and robust shot change detection," Journal of Real-Time Image Processing, vol. 2, no. 1, pp. 23–34, 2007.
2. J. Baber, N. Afzulpurkar, and S. Satoh, "A framework for video segmentation using global and local features," International Journal of Pattern Recognition and Artificial Intelligence, vol. 27, no. 5, Article ID 1355007, 2013.
3. J. Baber, N. Afzulpurkar, and M. Bakhtyar, "Video segmentation into scenes using entropy and SURF," in Proceedings of the 2011 7th International Conference on Emerging Technologies (ICET' 11), pp. 1–6, IEEE, 2011.
4. Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., Oliva, A.: Learning deep features for scene recognition using places database. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., Weinberger, K. (eds.) Advances in Neural Information Processing Systems, vol. 27, pp. 487–495. Curran Associates, Inc., Dutchess (2014)
5. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. 2, 2169– 2178 (2006). https://doi.org/10.1109/CVPR.2006.68
6. Torralba, A., Murphy, K.P., Freeman, W.T., Rubin, M.A.: Context-based vision system for place and object recognition. In: Proceedings of the Ninth IEEE International Conference on Computer Vision - Volume 2, ICCV '03, p. 273. IEEE Computer Society, Washington, DC, USA (2003). http://dl.acm.org/citation.cfm? id=946247.946665
7. Over, P., Awad, G., Michel, M., Fiscus, J., Kraaij, W., Smeaton, A.F., Queenot, G., Ordelman, R.: Trecvid 2015—an overview of the goals, tasks, data, evaluation mechanisms and metrics. In: Proceedings of TRECVID 2015. NIST, USA (2015)
8. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet large scale visual recognition challenge. Int. J. Comput. Vis. (IJCV) 115(3), 211–252 (2015). https://doi.org/10. 1007/s11263-015-0816-y.
9. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. CoRR abs/1409.1556 (2014). http://arxiv.org/abs/1409.1556
10. Deng, J., Li, K., Do, M., Su, H., Fei-Fei, L.: Construction and Analysis of a Large-Scale Image Ontology. Vision Sciences Society, Baltimore (2009).
11. Del Fabro, M., Böszörmenyi, L.: State-of-the-art and future challenges in video scene detection: a survey. Multimed. Syst. 19(5), 427–454 (2013). https://doi.org/10.1007/s00530-013-0306-4.
12. Hanjalic, A., Lagendijk, R.L., Biemond, J.: Automated high-level movie segmentation for advanced video-retrieval systems. IEEE Trans. Circuits Syst. Video Technol. 9(4), 580–588 (1999). https:// doi.org/10.1109/76.767124
13. Truong, B.T., Venkatesh, S., Dorai, C.: Scene extraction in motion pictures. IEEE Trans. Circuits Syst. Video Technol. 13(1), 5–15 (2003). https://doi.org/10.1109/TCSVT.2002.808084.

14. Odobez, J.M., Gatica-Perez, D., Guillemot, M.: Spectral Structuring of Home Videos, pp. 310–320. Springer, Berlin (2003). https:// doi.org/10.1007/3-540-45113-7_31.
15. Aner, A., Kender, J.R.: Video Summaries Through Mosaic-Based Shot and Scene Clustering, pp. 388–402. Springer, Berlin (2002). https://doi.org/10.1007/3-540-47979-1_26.
16. Yeung, M., Yeo, B.L., Liu, B.: Segmentation of video by clustering and graph analysis. Comput. Vis. Image Underst. 71(1), 94–109 (1998). https://doi.org/10.1006/cviu.1997.0628
17. Kwon, Y.M., Song, C.J., Kim, I.J.: A new approach for high-level video structuring. In: IEEE International Conference on Multimedia and Expo (2000).
18. Mitrovi´c, D., Hartlieb, S., Zeppelzauer, M., Zaharieva, M.: Scene Segmentation in Artistic Archive Documentaries, pp. 400– 410. Springer, Berlin (2010). https://doi.org/10.1007/978-3-642-16607-5_27.
19. Huayong, L., Hui, Z.: The Segmentation of News Video into Story Units, pp. 870–875. Springer, Berlin (2005). https://doi.org/10. 1007/11563952_95.
20. Y. Cui, Y. Cai, C. Qiu and X. Gao, "Scene detection of news video using CNN features," 2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), 2017, pp. 1-5, doi: 10.1109/CISP-BMEI.2017.8301916.
21. Vasileios T. Chasanis, Aristidis C. Likas, and Nikolaos P. Galatsanos," Scene Detection in Videos Using Shot Clustering and Sequence Alignment", IEEE Transactions on Multimedia, vol. 11, no. 1, January 2009.
22. D. Ramanan, D. A. Forsyth and K. Barnard, "Building models of animals from video," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 28, no. 8, pp. 1319-1334, Aug. 2006, doi: 10.1109/TPAMI.2006.155.
23. G Suryanarayana, LNC Prakash K, PC Mahesh, T Bhaskar, "Novel dynamic k-modes clustering of the categorical and non categorical dataset with optimized genetic algorithm based feature selection", Multimedia Tools and Applications, 1-20.
24. LNC.Prakash K, K.Anuradha, "clustering multivalued attribute data using transaction weights as utility-scale", Journal of Advanced Research in Dynamical and Control Systems 9 (18), 3132-3151.