PERFORMANCE STUDY WITH THE MULTIPLE QUERY SPECIFIC DISTANCE MEASURES FOR VIDEO OBJECT RETRIEVAL

C R Bharathi[1], *Sanjay B Waykar[2]

[1]Associate Professor, Dept. of Electronics and Communications. Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Avadi, Chennai, India

*[2]Assistant Professor, Dept. of Computer Engineering, Sinhgad Institute of Technology, Lonavala, Maharashtra, India.

ABSTRACT

Nowadays, the occurrence of the recording capability of video has gained popularity in the mobile devices or surveillance systems. However, retrieval of video objects is still challenging. The main aim of this research is to design and develop a Deep Long Short-Term Memory (Deep-LSTM) systembased on the video object retrieval. Initially, the key frame gets extracted from the video and the objects are detected from the extracted key frames using the nearest neighborhood algorithm. Then, the trajectory of the objects detected is trackedbased on the Deep LSTM in such a way that the location of the detected objects will be tracked in the video.Once the video objects are tracked, the object retrieval mechanism is utilized by means of multiple query specific distance measures, such as Euclidean distance, Bhattacharyya distance,Canberra distance, Jaro–Winkler distance, tanimoto similarity and Hausdorff distance. However, the effectiveness of the proposed method is evaluated based on the performance metrics, such as precision, recall, Multiple Object Tracking Precision (MOTP) and F1-score. The results achieved will be compared with that of existing works for revealing the efficiency of the proposed method.

***Keywords*:** Video object retrieval, Object detection, Object tracking, Nearest neighborhood algorithm, Deep LSTM.

1. Introduction

In recent years, video retrieval has been considered as the fundamental research due to the emerging technical advances in the field of video retrieval, and it has increased the presence of numerous videos [5]. In addition, surveillance cameras present in various areas raised the total number of the videos in the video-based records and perform the security operations. An efficient mechanism should be developed for the efficient retrieval of the videos, which is mainly needed for handling the videos available in the archive. Therefore, the use of the video retrieval is rising day-by-day [6]. In general, there are two methods available for the retrieval of the videos, namely content based and text based. In the content-based retrieval approach, the input is given as the video clips or the image, and in the text-based retrieval approach, the input is in the form of text and various methodologies are adopted for retrieving the videos [12][5]. Video retrieval system based on the objects is considered as the dynamic research field, and it consists of three fundamental steps, namely video segmentation, feature extraction and grouping. Few procedures-based on the extraction of feature involves the Gabor filters or wavelet in order to extract the features from the frames [7] [4].

In addition, video retrieval technique is considered very significant and also considered as challenging issue due to the emerging rise of the social media applications, and the platform-based on sharing of videos. Based on the increasing amount of the videos published in several platforms, namely you-tube, it is much familiar to get the several videos of the same occurrence and also plane crash and terrorist attack are the examples, which may be the duplicate of some real videos, or may shows the same event from multiple perspective or at multiple times. The videos can be efficiently retrieved for various applications ranging from the copy detection for the protection of copyrights [8] in order to reconstruct the incidents [14][9], and also to verify the news [10] [3]. Several methods have been developed for the widespread applications of the video retrieval. In [5], a video retrieval approach was developed in order to perform the video retrieval for the required objects-based on the inputs of the appearance and the trajectory points of the objects, in such a way that the accuracy can be improved by developing various learning methods [4]. In [11], a content-based video retrieval and indexing was developed, and the frameworks for the video retrieval consists of feature extraction, video annotation, structure analysis, data mining, video browsing, queries and feedback [13][1].

The major objective of this research is to design and developaDeep LSTM-based video object retrieval system for the video objects retrieval, which offers a better efficiency in retrieval of objects present in the video. Multiple query specific distance measures are another contribution, which calculates the distance of the trajectory points among the detected video objects.

- **Proposed Deep LSTM:** Deep LSTM-based video object retrieval system is proposed for tracking the detected objects trajectory in order to track the location of the detected objects in the video. Moreover, multiple query specific distance measures are applied in such a way that the distance can be computed between the tracked object image, and the trajectory point of query image for the retrieval of video objects.

The organization of the paper is described as follows: Section 2 reviews the existing video objects retrieval methods, section 3 describes the proposed Deep LSTM-based VOR, section 4 elaborates the result and discussion of the proposed method and section 5 describes the conclusion of the paper.

2. Motivation

In this section, various existing video object retrieval methods are reviewed and their advantages and limitations are studied. The main motive of this research is to analyze the performance of the multiple query specific distance measures for video object retrieval.

2.1 Literature Survey

This section depicts a review of the literature on four existing video object retrieval techniques using multiple query specific distance measures. SihaoDing *et al.* [1]developed SurvSurf approach was introduced for the retrieval of humans on the surveillance of huge video data. In this method, motion information in the video was employed in order to partition the video data, and the partitioned data unit, called M-clip in such a way that the M-clips were capable of eliminating the redundancy in the video data contents and volumes. In addition, Map Reduce framework was utilized for detecting the humans by processing the M-clips and finally, V-Big Table was applied in order to formalize the information-based on the M-clips, thereby achieved efficient results in retrieval of data. However, this approach does not consider novel optimization algorithms to yield better results. Ning Zhang and Hwa-Young Jeong[2] introduced anew retrieval algorithm for the specific images related to face in the cloud computing framework. In this algorithm, a classifier, named face cascade classifier was employed for examining the face images in the airport surveillance of the multimedia video. In addition, a clock matching approach was utilized for tracking the face, and also can be capable of finding the missing face during the airport surveillance. This method suffered from computational complexity issues. Giorgos Kordopatis-Zilos *et al.*[3] designed an approach, called Fine-grained Incident Video Retrieval (FIVR) for the video retrieval. The main objective of this method is to retrieve the videos related to the video-based given queries with respect to the various duplicate videos from the same event in such a way that this FIVR framework consists of several retrieval tasks for the video objects. However, this method failed to consider the techniques-based on query expansion to solve the retrieval issues for better results. C. A. Ghuge *et al.* [4] developed ahybrid model-based on the integration of the Nearest Search Algorithm (NSA), and Non-Linear Autoregressive Exogenous (NARX) neural network. The major aim of this hybrid model is to retrieve the needed object with respect to the trajectory points of the objects. However, this method failed to utilize the deep learning approaches in order to enhance the retrieval performance.

2.2 Challenges

Some of the challenges faced by the existing video object retrieval techniques are explained below as follows,

- SurvSurf approach was introduced for the retrieval of humans on the surveillance of huge video data, but this method failed to consider the motion features, height, face and other appearance characteristics in order to achieve better results [1].
- FIVR was developed for the incident retrieval of videos, but the challenge lies in examining the matching practices-based on the frame-level with respect to the temporal alignment among the videos, thereby resulting in efficient retrieval results. [3].
- In [4], a hybrid method based on NSA and NARX was designed for the retrieval of video objects by considering the trajectories. However, the challenge lies in utilizing this method in the crowded scenes to perform efficient retrieval of video.

3. Deep LSTM-based VOR

This section describes the developed Deep LSTM method for the video objects retrieval using the trajectory points. Figure 1 represents the block diagram of the developed Deep LSTM-based VOR method. The series of steps carried out for the video object retrieval are elaborated in this section. It is processed using four phases, which involves key frame extraction, object detection, detected object tracking and object retrieval. Initially, the video input is fed into the key frame extraction phase so that the key frames from the video gets extracted, and then the extracted key frames are subjected to the object detection phase in order to detect the objects from the key frames, which is carried out using the Nearest Neighborhood algorithm. Then, the detected objects are given to the detected object tracking phase where the trajectory of the detected object is tracked using the Deep LSTM in such a way that the location gets tracked in the video. Once the video objects are tracked, object retrieval is performed by means of multiple query specific distance measures, such as

Euclidean, Bhattacharyya, Canberra, Jaro–Winkler, Tanimoto similarity and Hausdorff distance, thereby resulting in efficient video tracking and retrieval of video.

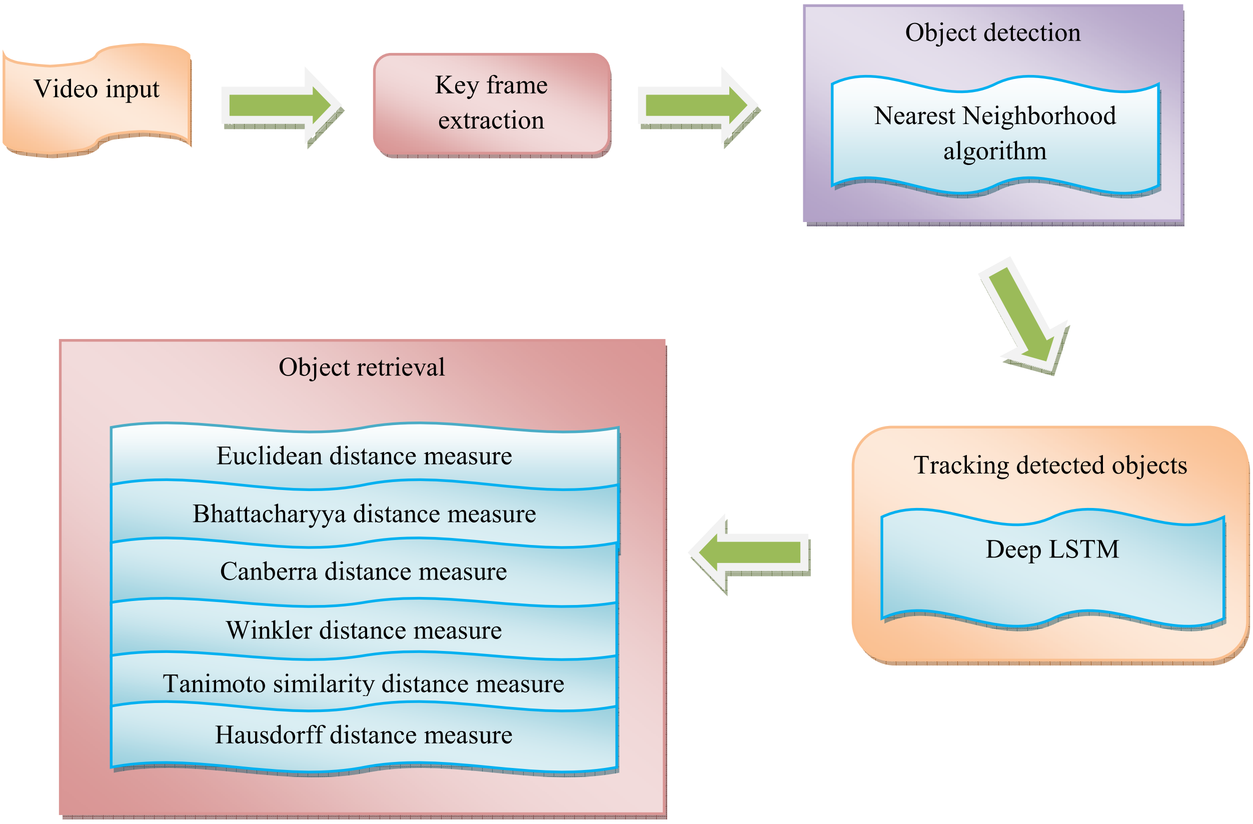

Video input
Key frame extraction
Object detection
Nearest Neighborhood algorithm
Tracking detected objects
Deep LSTM
Object retrieval
Euclidean distance measure
Bhattacharyya distance measure
Canberra distance measure
Winkler distance measure
Tanimoto similarity distance measure
Hausdorff distance measure

Figure1. System model for Deep LSTM-based VOR

4. Distance Measures Considered

The six distance measures, like Euclidean, Bhattacharyya, Canberra, Jaro-Winkler, Tanimoto and Hausdorff distance are considered for the video object retrieval.

4.1 Euclidean Distance

Euclidean distance measure is referred as the straight-line distance measure among the two points in the Euclidean space, and is expressed as,

$$d(r,s)=\sqrt{\sum_{i=1}^{n}(s_i-r_i)^2} \tag{1}$$

where, (r,s) denotes the points in the Euclidean space, s_i, r_i represents the Euclidean vectors and n represents the Euclidean n-space

4.2 Bhattacharyya Distance

Bhattacharyya distance measure is defined as the similarity-based on the two probability distributions and is represented as,

$$BC(f,g)=\sum_{i=1}^{n}\sqrt{f_i g_i} \tag{2}$$

where, (f,g) denotes the samples, f_i, g_i represents the members of samples in the i^{th} partition and n expresses the number of partitions.

4.3 Canberra Distance

The Canberra distance measure is the arithmetic distance measure between the pairs of vector points in the vector space and is denoted as,

$$d(y,z)=\sum_{i=1}^{n}\frac{|y_i-z_i|}{|y_i|+|z_i|} \tag{3}$$

where, y and z are represented as vectors.

4.4 Jaro-Winkler Distance

The Jaro-Winkler distance measure is defined as the measure of similarity among two strings and is given as,

$$d_w=\begin{cases}0 & p=0\\ \frac{1}{3}\left(\frac{p}{|q_1|}+\frac{p}{|q_2|}+\frac{p-n}{p}\right) & otherwise\end{cases} \tag{4}$$

where, $|q_i|$ denotes the string length, p represents the matching characters, and n indicates half the number of transposition characters.

.

4.5 Tanimoto Similarity

Tanimoto similarity distance is defined as the ratio of the intersecting set to the union set as a similarity measure, and is denoted as,

$$T(i,j)=\frac{F_k}{F_i+F_j+F_k} \tag{5}$$

where, the term F indicates the attributes of objects (i,j) and k expresses the intersection set.

4.6 Hausdorff Distance

Hausdorff distance measure is the distance between the two subsets of a metric space from each other and is represented as,

$$d_H(U,V) = \max\{d(U,V), d(V,U)\} \tag{6}$$

5. Results and Discussion

This section explains the results and discussion of the developed Deep LSTM method-based on the distance measures, such as Euclidean, Bhattacharyya, Canberra, Jaro-Winkler, Tanimoto and Hausdorff distance.

5.1 Experimental Setup

The experimental setup of the developed method is carried out in MATLAB tool with the OS 10, 2GB RAM, and the Intel i3 core processor.

5.2. Description of Dataset

The dataset used for the implementation of the proposed method is CAVIAR dataset [16]. The CAVIAR dataset comprises of various clippings of the video, like three persons shopping, one shop, walkby shop, and the walking. The lens-based on the wide-angle camera is used for collecting the video clips. Accordingly, the file size ranges from 6to 12 MB and some files are expanded to 21 MB.

5.3 Performance Metrics

The performance of the developed method is evaluated using four performance evaluation metrics, such as precision, recall, MOTP, and the F-measure.

a) Precision: It is the ratio of the relevant object instances from the retrieved object instances and is denoted as,

$$precision = \frac{|\{\mu \cap \tau\}|}{|\tau|} \tag{7}$$

where, μ indicates the relevant objects, τ denotes the retrieved objects.

b) Recall: Recall is referred as the overall instances in the relevant objects that are retrieved actually and is denoted as,

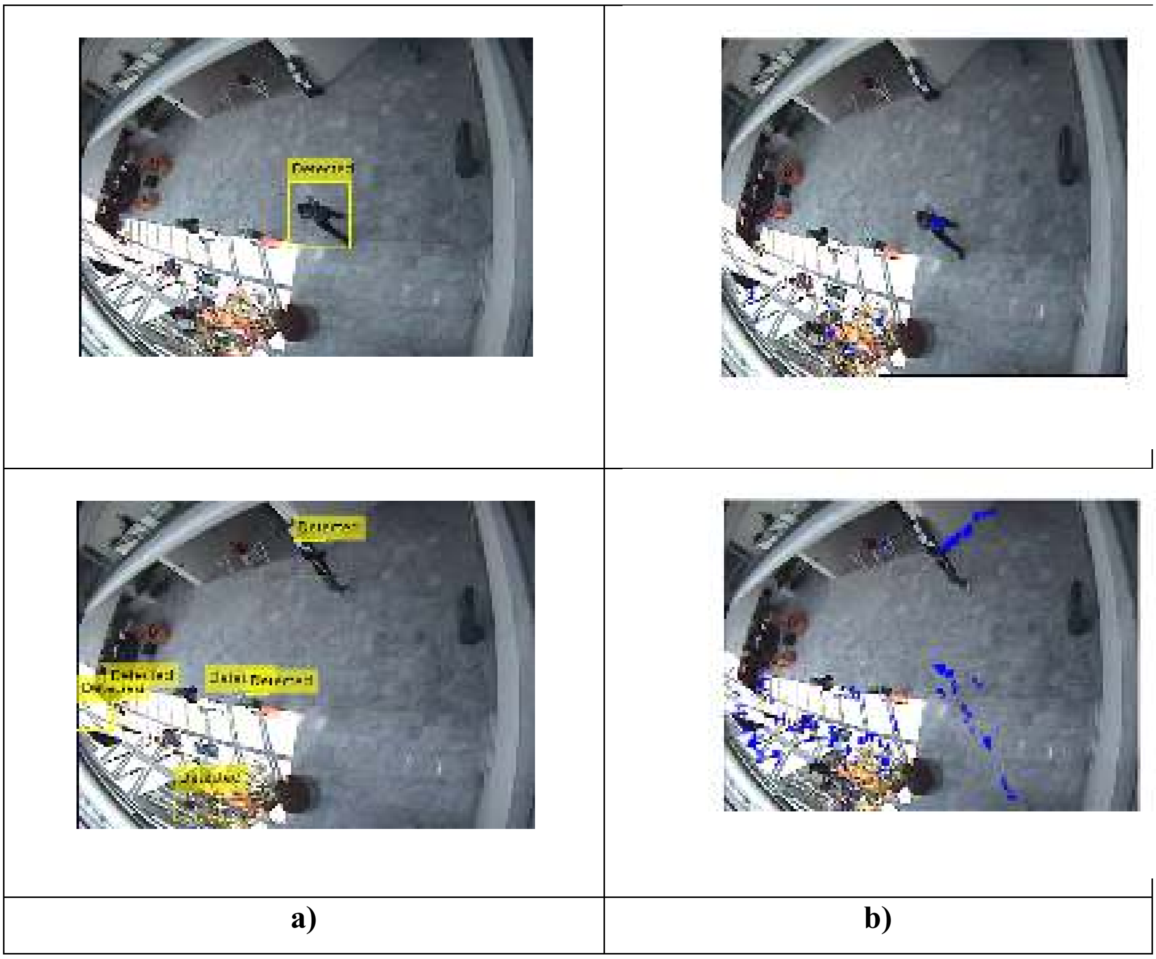
$$recall = \frac{|\{\mu \cap \tau\}|}{|\mu|} \tag{8}$$

c) F-measure: The harmonic mean of recall and precision is called as F-measure and is expressed as,

$$F - measure = 2 * \frac{precision * recall}{precision + recall} \tag{9}$$

5.4 Simulation results

The simulation results of the developed Deep LSTM based on video-1 and 2 are shown in figure 3.and figure 4. The detected objects and detected paths using video 1 is illustrated in figure 3a) and figure 3b) respectively .figure 4a) and figure 4b) shows the detected objects and detected paths based on video 2 respectively.

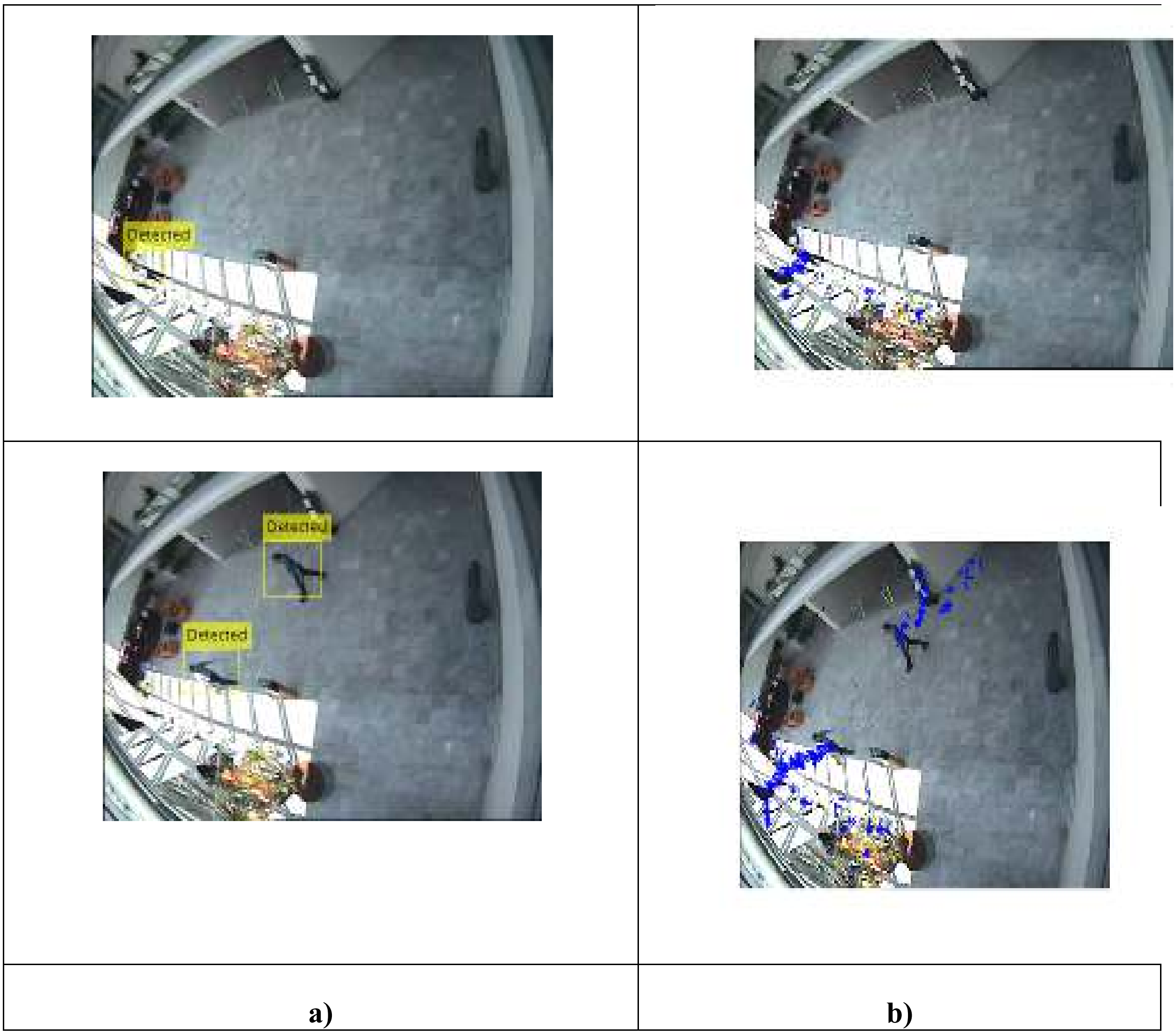


a) b)

Figure 3. Sample results based on video-1 a) Detected objects b) Detected paths

5.5 Comparative Analysis

This section elaborates the comparative analysis of video objects retrieval using Deep LSTM method by considering the CAVIAR dataset, which includes Euclidean, Bhattacharyya, Canberra, Jaro-Winkler, Tanimoto and Hausdorff distance.



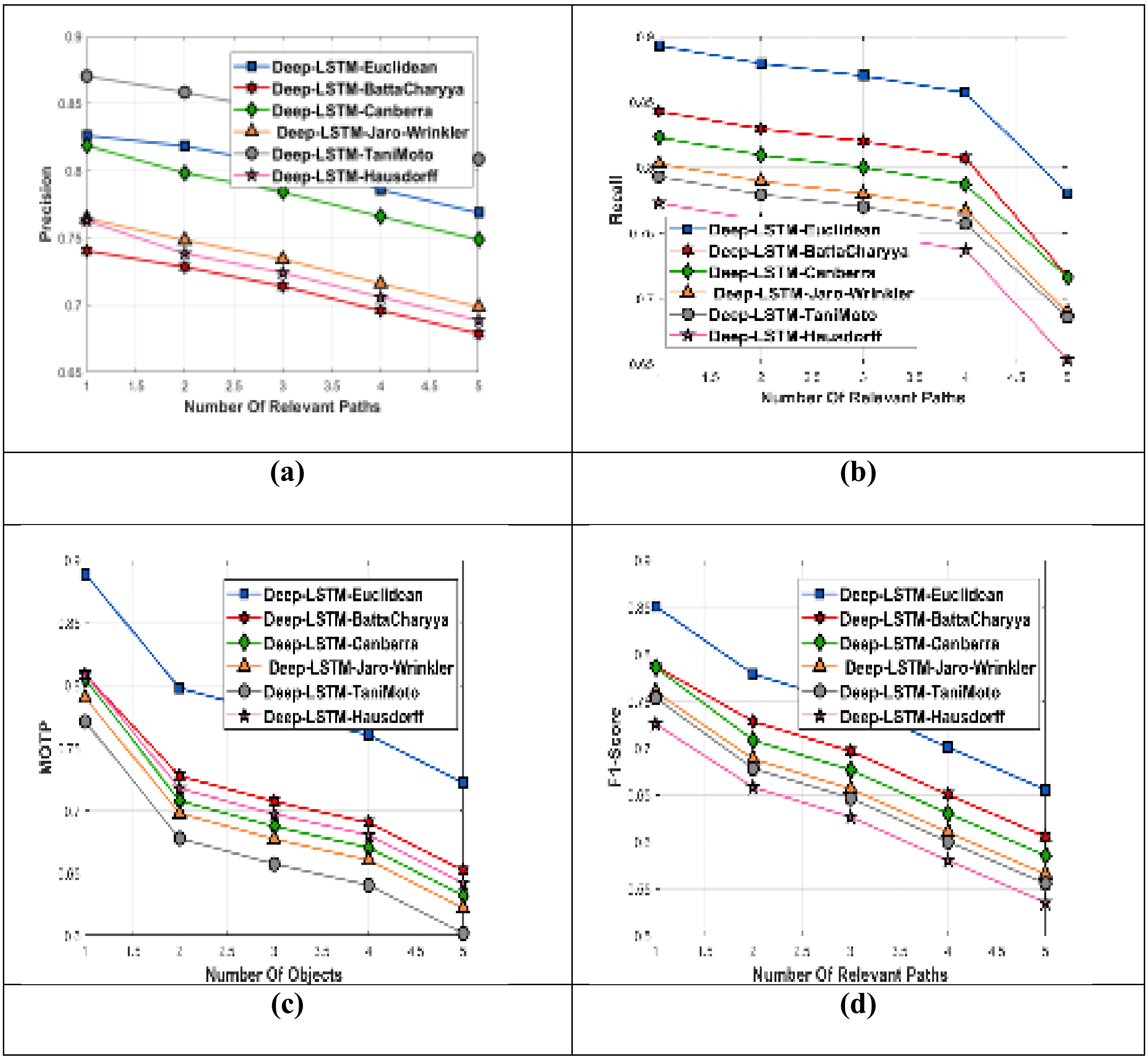
a) b)

Figure 4. Sample results based on video-2 a) Detected objects b) Detected paths

5.5.1. Analysis based on Video 1

Figure 5 shows the comparative analysis of Deep LSTM based on the performance metrics using video 1. The analysis based on precision metrics by varying the relevant paths is depicted in figure 5a). When the relevant path count is 1, the precision values computed by Euclidean, Bhattacharyya, Canberra, Jaro-Winkler, Tanimoto and Hausdorff distance measures are 0.826, 0.740, 0.818, 0.764, 0.870, 0.762. In figure 5b) the analysis with respect to recall metrics is potrayed. When the relevant path count is 1, the recall values computed by Euclidean, Bhattacharyya, Canberra, Jaro-Winkler, Tanimoto and Hausdorff distance measures are 0.892, 0.842,0.822, 0.802, 0.792, 0.772. Figure 5c) illustrates the analysis based on MOTP. When the count of object is 1, the MOTP metric computed by Euclidean, Bhattacharyya, Canberra, Jaro-Winkler, Tanimoto and Hausdorff distance measures are

0.889, 0.808, 0.805, 0.790, 0.771, 0.808. The analysis based on F1-score by varying the relevant paths is illustrated in figure 5d). When the number of path count is 1, the F1-score computed by Euclidean, Bhattacharyya, Canberra, Jaro-Winkler, Tanimoto and Hausdorff distance measures are 0.850, 0.787, 0.785, 0.759, 0.753, 0.725 respectively.



(a)

(b)

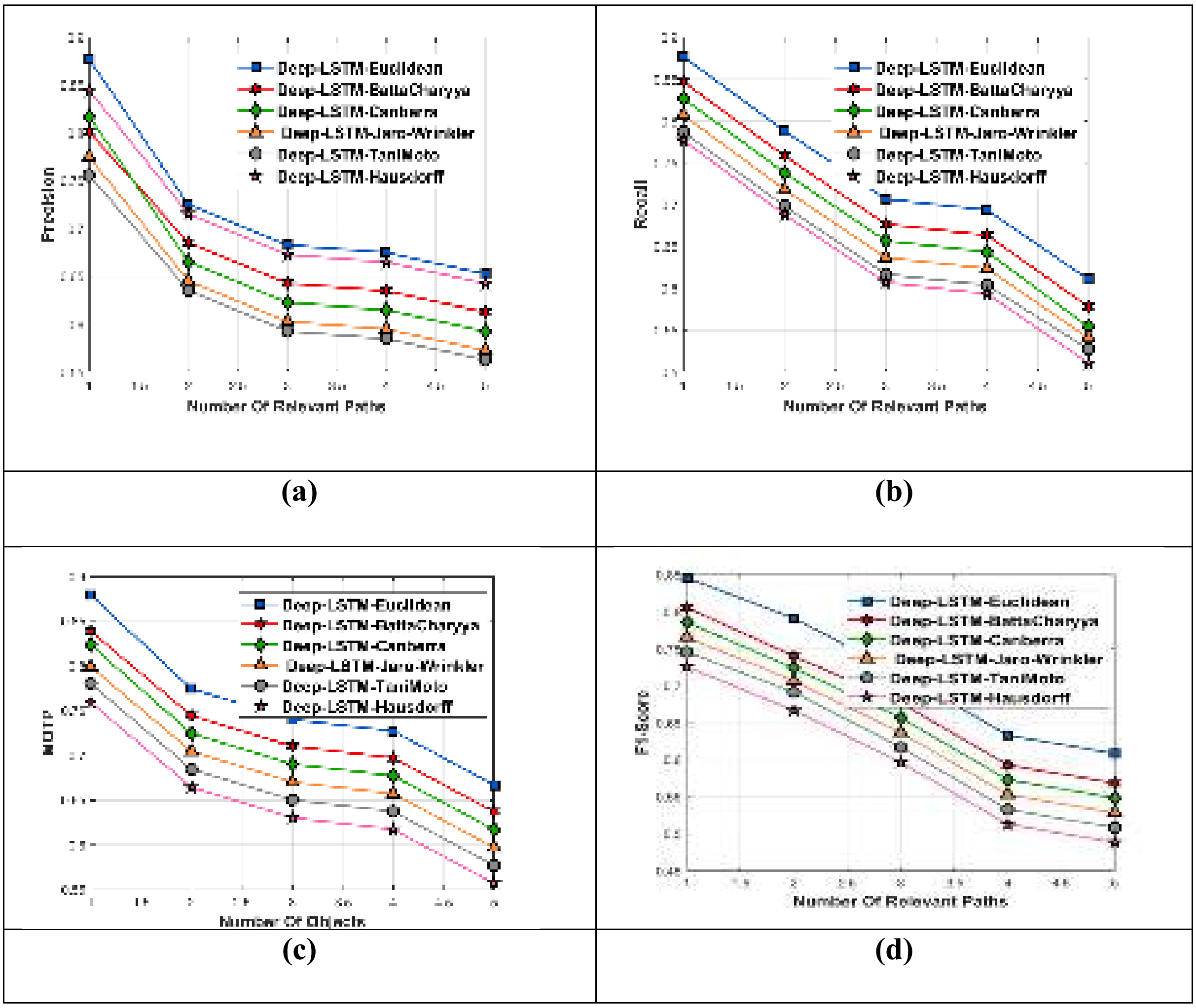
(c)

(d)

Figure 5. Comparative analysis with respect to video 1 for a) precision b) Recall c) MOTP d) F1-score

5.5.2. Analysis based on Video 2

Figure 6 shows the comparative analysis of Deep LSTM based on the performance metrics using video 2. The analysis using the precision metrics by varying the relevant paths is depicted in figure 6a). When the relevant path count is 1, the computed precision values of Euclidean, Bhattacharyya, Canberra, Jaro-Winkler, Tanimoto and Hausdorff distance measures are 0.877, 0.801, 0.817, 0.774, 0.755, 0.844. In figure 6b), the analysis with respect to recall metrics is potrayed. When the number of relevant path is 1, the recall values computed by Euclidean, Bhattacharyya, Canberra, Jaro-Winkler, Tanimoto and Hausdorff distance measures are 0.877, 0.847, 0.827, 0.807, 0.787, 0.777. Figure 6c) illustrates the analysis based on MOTP. When the count of object is 1, the MOTP metric computed by Euclidean, Bhattacharyya, Canberra, Jaro-Winkler, Tanimoto and Hausdorff distance measures are 0.8795, 0.839,0.823, 0.799, 0.780, 0.759. The analysis based on F1-score by varying the relevant paths is illustrated in figure 6d). When the number of path count is 1, the F1-score computed by Euclidean, Bhattacharyya, Canberra, Jaro-Winkler, Tanimoto and Hausdorff distance measures are 0.845, 0.805, 0.785, 0.765, 0.745, 0.725 respectively.



(a)

(b)

(c)

(d)

Figure 6. Comparative analysis with respect to video 2 for a) precision b) Recall c) MOTP d) F1-score

5.6 Comparative Discussion

The analysis of Deep LSTM method for the video objects retrieval with the Euclidean, Bhattacharyya, Canberra, Jaro-Winkler, Tanimoto and Hausdorff distance measure using video-1 and video-2 is illustrated below. The method is said to be effective if it achieves high values for the measures, like precision, recall, MOTP and F1-score. The maximum precision value measured by the proposed Deep LSTM method is 0.877 using video 2. The maximal recall value attained by proposed Deep LSTM considering video 1 is 0.892. Similarly, the

MOTP value obtained for the proposed Deep LSTM- based on video 1 is 0.889 respectively. The higher F1-score value measured by the proposed Deep LSTM-based on video 1 is 0.850 respectively. Table 1 portrays the comparative discussions of proposed method in terms of precision, recall, MOTP and F-measure.

Table 1. Comparative discussion

Number of videos	**Metrics**	**Euclidean**	**Bhattacharyya**	**Canberra**	**Jaro-Winkler**	**Tanimoto**	**Hausdorff**
Video 1	**Precision**	0.826	0.740	0.818	0.764	0.870	0.762
	Recall	0.892	0.842	0.822	0.802	0.792	0.772
	MOTP	0.889	0.808	0.805	0.790	0.771	0.808
	F1-score	0.850	0.787	0.785	0.759	0.753	0.725
Video 2	**Precision**	0.877	0.801	0.817	0.774	0.755	0.844
	Recall	0.877	0.847	0.827	0.807	0.787	0.777
	MOTP	0.879	0.839	0.823	0.799	0.780	0.759
	F1-score	0.845	0.805	0.785	0.765	0.745	0.725

6. Conclusion

In this research, Deep LSTM systembased on video object retrieval is developed for tracking the detected object location by tracking the trajectory of the detected objects in the video. Moreover, the multiple query specific distance measures are utilized in such a way that the distance can be computed for the retrieval of video objects.The dataset utilized for the retrieval of video objects is CAVIAR dataset. The developed Deep LSTM method with the distance measures, such as Euclidean, Bhattacharyya, Canberra, Jaro–Winkler, Tanimoto and Hausdorff are analyzed to evaluate the performance. However, the values achieved by the evaluation metrics like precision, recall, MOTP, and the F1-score are discussed and compared with the existing methods proven that the developed Deep LSTM method shows better and efficient results. Meanwhile, there are some limitations that should be considered for future enhancement. The developed Deep LSTM method combined with the new optimization algorithms may provide even more accurate results and improves the performance efficiently.

References

[1] Ding S, Li G, Li Y, Li X, Zhai Q, Champion AC, Zhu J, Xuan D, Zheng YF., "Survsurf: human retrieval on large surveillance video data", Multimedia Tools and Applications, vol.76, no.5, pp.6521-49, March 2017.

[2] Zhang N, Jeong HY., “A retrieval algorithm for specific face images in airport surveillance multimedia videos on cloud computing platform”, Multimedia Tools and Applications, vol.76, no.16, pp.17129-43, August 2017.

[3] Kordopatis-Zilos G, Papadopoulos S, Patras I, Kompatsiaris I., “FIVR: Fine-grained incident video retrieval”, IEEE Transactions on Multimedia, vol.21, no.10, pp.2638-52, March 2019.

[4] Ghuge CA, Chandra Prakash V, Ruikar SD., “Weighed query-specific distance and hybrid NARX neural network for video object retrieval”, The Computer Journal, vol.63, no.11, pp.1738-55, November 2020.

[5] Lai YH, Yang CK., “Video object retrieval by trajectory and appearance”, IEEE Transactions on Circuits and Systems for Video Technology, vol.25, no.6, pp.1026-37, September 2014.

[6] Guo H, Wang J, Lu H., “Multiple deep features learning for object retrieval in surveillance videos”, IET Computer Vision, vol.10, no.4, pp.268-72, February 2016.

[7] Li, Y., Wang, R., Cui, Z., Shan, S. and Chen, X., “Spatial pyramid covariance-based compact video code for robust face retrieval in TV-series”, IEEE Transactions on Image Processing, vol.25, no.12, pp.5905-5919, 2016.

[8] M. Douze, H. J´egou, and C. Schmid, “An image-based approach to videocopy detection with spatio-temporal post-filtering”, IEEE Transactions on Multimedia, vol.12, no.4, pp. 257–266, 2010.

[9] L. Gao, P. Wang, J. Song, Z. Huang, J. Shao, and H. T. Shen, “Eventvideo mashup: From hundreds of videos to minutes of skeleton”,In proceedings of 31stAAAI Conference on Artificial Intelligence, pp.1323–1330, 2017.

[10] Z. Jin, J. Cao, Y. Zhang, J. Zhou, and Q. Tian, “Novel visual and statisticalimage features for microblogs news verification”, IEEE transactions on multimedia, vol.19, no.3, pp.598–608, 2017.

[11] Hu, W., Xie, N., Li, L., Zeng, X. and Maybank, S., “A survey on visual content-based video indexing and retrieval”, IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol.41, no.6, pp.797-819, 2011.

[12] Puttaswamy M R, "Improved Deer Hunting Optimization Algorithm for video based salient object detection", Multimedia Research, Vol 3, No 3, 2020.

[13]V. Mallikalava, S. Yuvaraj, K. Vengatesan, A. Kumar, S. Punjabi and S. Samee, "Theft Vehicle Detection Using Image Processing integrated Digital Signature Based ECU," 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, 2020, pp. 913-918

[14] Amit Sarkar, "Optimization Assisted Convolutional Neural Network for Facial Emotion Recognition", Multimedia Research, Vol 3, No 2, 2020.

[15] Vasamsetti Srinivas,Santhirani Ch, "Hybrid Particle Swarm Optimization-Deep Neural Network Model for Speaker Recognition", Multimedia Research, Vol.3,No.1, pp.1-10,2020

[16].K.Srinivas, G.Madhukar rao, K.Vengatesan, P.Shivkumar Tanesh, A. kumar,and S. Yuvaraj,"An implementation of subsidy prediction system using machine learning logistical regression algorithm",Advances in Mathematics: Scientific Journal 9 (2020), no.6, 3407–3415.

[17].VengatesanK, E Saravana Kumar, S. Yuvaraj, Punjabi Shivkumar Tanesh, Abhishek Kumar. (2020). An Approach for Remove Missing Values in Numerical and Categorical Values Using Two Way Table Marginal Joint Probability. International Journal of Advanced Science and Technology, 29(05), 2745 - 2756