Control and Cybernetics

vol. **49** (2020) No. 4

Systematic analysis and review of video object retrieval techniques*

by

C. A. Ghuge[1], V. Chandra Prakash[1] and Sachin D. Ruikar[2]

[1] Department of Computer Science and Engineering, KoneruLakshmaiah Education Foundation, Vaddeswaram, Guntur district, AP, India
caghugeklu@gmail.com
[2]Department of Electronics Engineering, Walchand College of Engineering, Sangli, Maharashtra, India

Abstract: Video object retrieval is a promising research direction, developing in the recent years, and the current video object retrieval strategies are used for visualizing, digitizing, modeling, and retrieving the objects especially in graphics and in architectural design. The research performed led to the design of proficient video object retrieval techniques. Yet, although, a number of algorithms had been devised for tracking objects, the problems persist in enhancing the performance, for instance – with regard to non-rigid objects. In this review article we provide a detailed survey of 50 research papers presenting the suggested video object retrieval methodologies, based on approaches such as deep learning techniques, graph-based techniques, query-based techniques, feature-based techniques, fuzzy-based techniques, machine learning-based techniques, distance metric learning-based technique, and also other ones. Moreover, analysis and discussion are presented concerning the year of publication, employed methodology, evaluation metrics, accuracy range, adopted framework, datasets utilized, and the implementation tool. Finally, the research gaps and issues related to various proposed video object retrieval schemes are presented for guiding the researchers towards improved contributions to the video object retrieval methods.

Keywords: video object retrieval, computer vision, deep learning, fuzzy-based techniques, machine learning, query-based techniques, graph-based techniques

1. Introduction

The video analysis techniques and computer vision are, in particular, considered an important part of the video surveillance systems (Ahmed et al., 2019; Birkas, Birkas and Popa, 2016; Lai and Yang, 2014). The video based systems

*Submitted: January 2020; Accepted: January 2021.

Abbreviations used throughout the paper in the sequence of their appearance

Abbreviations/Acronyms	**Description**
NIP	Nested Invariance Pooling
CDVA	Compact Descriptor for Video Analysis
CNN	Convolution Neural Network
LSH	Locality-Sensitive Hash
MAC	Maximum activations of convolutions
3D	3 Dimensional
AVWs	Auxiliary Visual Words
2D	2 Dimensional
QEVR	Query by Example Video Retrieval
CS	Compressed Sensing
DCT	Discrete Cosine Transform
3DOR	3D Object Retrieval
COBRA	COntentBasedRetrievAl
HMMs	Hidden Markov Models
IMU	Inertial Measurement Unit
PCA	Principal Component Analysis
RGB	Red, Green, Blue
HSL	Hue, Saturation, Lightness
NDVR	Near-duplicate Video Retrieval
SIFT	Scale Invariant Feature Transform
DGTC	Differential Geometric Trajectory Cloud
BoF	Bag-of-Features
QSD	Query-Specific Distance
SQL	Structured Query Language
MAP	Mean Average Precision
MOTA	Multi-object Tracking Accuracy
MOTP	Multi-object Tracking Precision
ACM	Association for Computing Machinery
IEEE	Institute of Electrical and Electronics Engineers

can capture huge variety of the desired information and are in principle of low cost due to easy installation of relatively inexpensive surveillance equipment (Kim and Hwang, 2000; Hu et al., 2004; Ghuge, Prakash and Ruikar, 2018a), its functioning and maintenance (Cheng and Hwang, 2011). The complexity of retrieval task rises to a higher level while processing different frames in digital videos (Chuang et al., 2014) as videos are usually filmed under different lighting conditions in an unimpeded manner (Yang et al., 2013). Due to large deployment of video camera installations, there arises an imperative requirement for automatic video understanding methods, which can effectively replace human operators (Ji and Liu, 2009; Sivic and Zisserman, 2003) for monitoring the areas under surveillance. In the video-based systems, the adequately robust systems discover and keep the record of moving objects. The recognition model is constructed after obtaining the tracking results and is used for extracting the features to determine the objects (Cheng and Hwang, 2011). Multiple modalities for video browsing and retrieval, including text search through the spoken dialogue, image matching against shot key frames (Ardizzone and La Cascia, 1997; Padmakala, Anandha Mala and Shalini, 2011; Yeo and Yeung, 1997), and object matching against segmented video objects (Shan, 2010; Farag and Abdel-Wahab, 2003) are utilized.

In modern video surveillance systems, video object retrieval is a major problem and an obstacle (Broilo et al., 2010; Valdés and Martínez, 2011), due to similarity of appearances, different poses, multiple object interactions, and object confusions. Owing to various particular visual manifestations, like occlusions, position changes, and diverse illumination, the object detection process becomes difficult for many of the techniques oriented at retrieval of the video objects. Besides, the retrieval of an object using video (see Chou, Chen and Lee, 2015; Dimitrova and Golshani, 1995) is influenced by the motion of the camera and object movement. In effect of the aforementioned problems, the size and shape of the objects are easily changed. Moreover, the matching methods are devised based on the measurement of similarities and features related to shape related to the resemblance of objects, which is in itself a complex task, taken up by many researchers (Aslandogan and Yu, 1999; Priyaa and Karthikeyan, 2013; Mezaris, Kompatsiaris and Strintzis, 2004), aiming to provide complete coverage of video retrieval (Smeaton and Browne, 2006; Ghuge, Prakash and Ruikar, 2018c). Recognition of objects in video can offer significant benefits to video retrieval, including automatic annotation and responding to content based queries, based on the object characteristics (Visser, Sebe and Bakker, 2002; Yan and Hauptmann, 2007; Eidenberger, Breiteneder and Hitz, 2002). Face recognition (Sivic, Schaffalitzky and Zisserman, 2006; Sivic and Zisserman, 2008), otherwise a very popular and important subject, has also been considered in the content-based video retrieval setting.

The primary intention here is to provide a detailed survey of the various video object retrieval techniques for the effective retrieval of objects using videos

(Anjulan and Canagarajah, 2007; Zhang et al., 2017). This review considers the existing video object retrieval methods adapted in the available techniques. The present survey is done by considering the year of publication, the employed methodology, the evaluation metrics, the datasets utilized, and the implementation tool. The scrutinized methods are classified into distinct approaches, and then, the survey is carried out with a view on the potential problems arising. Thus, the survey may also be considered to be an inspiration for the future extension of the effective video object retrieval methods.

The article is structured as follows: Section 1 provides the introduction to the video object retrieval, and Section 2 presents the survey of the currently available video object retrieval approaches. Then, Section 3 describes the research gaps and issues, related to the existing approaches, while Section 4 presents a shorthand evaluation of the existing methods. Finally, the conclusions from the survey are given in Section 5.

2. Literature survey on different video object retrieval techniques

In this section, the review of different video object retrieval techniques is provided. Figure 1 shows the here adopted classification of distinctive video object retrieval techniques, according to which the review has been performed. The techniques based on video object retrieval techniques are broadly categorized into eight groups of techniques, namely: deep learning, graph-based, query-based, feature-based, fuzzybased, machine learning-based, distance metric learning-based, and, finally, the remaining, other techniques.

2.1. Classification schemes for video object retrieval techniques

This section illustrates the specific research works that employ different video object retrieval classification schemes. The consecutive subsections are devoted to the studies, categorised according to the previously presented scheme.

2.2. Studies involving the deep learning techniques

Deep learning is a domain, belonging to the machine learning field, which has become recently immensely popular, and that is why we treat it separately and as the first group of techniques. Deep learning finds applications in a wide variety of fields, including object detection,video retrieval (Yang et al., 2018; Jin and Kim, 2017), and so on. Regarding the use of deep learning techniques in video retrieval, the following research reports and papers have been considered:

Lou et al. (2017) developed a method named NIP, based on deep-learning features,which incorporated also the CDVA framework for analyzing the videos. The method was used to obtain the compacted CNN descriptors. The CNN

descriptors are generated by applying three different pooling operations for generating the feature maps of CNN for scale invariant feature representation. The integration of CNNs and handcrafted descriptors can be used for enhancedconsideration of the complementary effects of the deeply learnt and handcrafted features. The method reduced the size of feature vector in CNN and can be used to improve the geometric invariance of the deep descriptors.

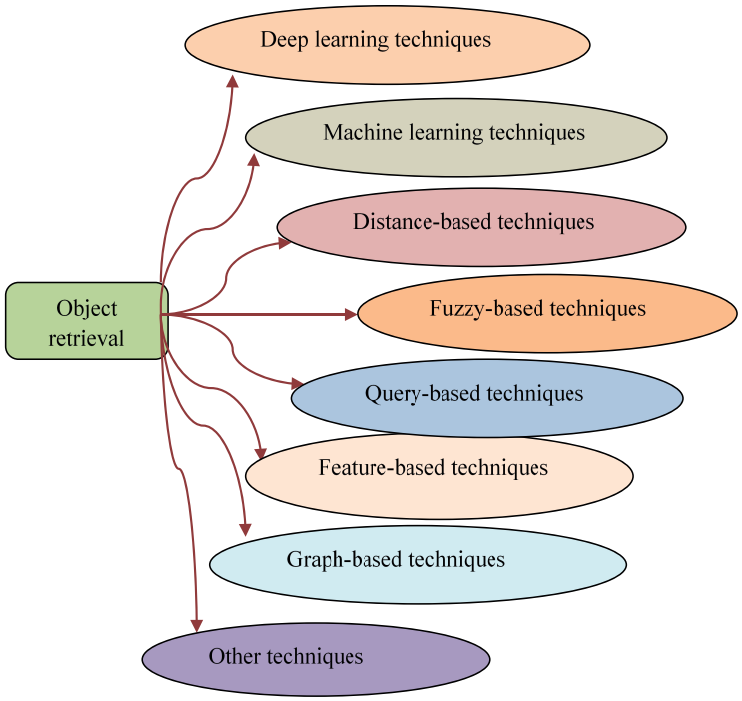

Figure 1. Categorization of distinct video object retrieval techniques used in the survey

Guo et al. (2015) developed a classifier, named discriminative CNN for learning the deep structures and colour features to form an effective multi-view object representation. Here, a CNN was trained using ImageNet and was used to abstract the structural information. The method used eleven colour names for extracting the deep colour features. As compared to conventional colour descriptors, the deep colour features captured more colour properties with alteration in the illumination conditions. The multi-view deep features (Zhao, F. et al., 2015; Li, Zaïane and Tauber, 1999) were encoded with binary codes using LSH and were fused for retrieving the objects.

Guo, Wang and Lu (2016) designed multiple deep features learning approach for retrieving the objects using the surveillance videos (see also Van Den Hengel et al., 2007; Hu et al., 2011). The discriminative CNN employed different deep features for describing the visual objects. Here, the CNN model was pre-trained using the ImageNet and fine-tuned for abstracting the structural information. Furthermore, the CNN model was trained using supervised colour names for delivering the colour information. For enhancing the performance of retrieval, the deep features were encoded into small binary codes and then fused to retrieve the interesting objects. Tolias, Sicre and Jegou (2015) used two retrieval stages, namely the initial search and the re-ranking strategies, which use the primitive information obtained from the CNN. Here, the compact feature vectors were built for encoding different image regions without subjecting the inputs to the network. The resulting bounding box is utilized for ranking the images. This method enhanced the performance by tuning the classifier based on maximum activations of convolutions (MAC) similarity. Region segmentation is designed in this approach so that the pre-image of the region corresponds to the same surface region. Videos of people have the definite benefit of containing multiple exemplars of each person in a form that can easily be associated automatically using straightforward visual tracking (Sivic, Everingham and Zisserman, 2004).

2.3. Studies involving the graph-based techniques

In this section, we present the studies, in which different graph-based techniques were used by the respective authors. Thus, Zhang et al. (2016a) developed a 3D video object retrieval technique based on features and the bipartite graph matching mechanisms. A view-based retrieval framework with multi-feature collaboration and bipartite graph matching is used to extract complement descriptors from both the contour and the interior region of 3D objects effectively. Here, the greedy bipartite graph matching algorithm and feature concatenation are used to enhance the performance of the 3D object retrieval tasks. The method extracted three descriptors and integrated them into the bipartite graph. The user feedback information was explored for attaining improved performance. Kuo et al. (2011) designed auxiliary visual words (AVWs) based approach for augmenting the image for purposes of searching for the required targets. The AVWs were identified automatically by propagating the features and selecting the textual and visual graphs in an unsupervised manner. The method employed different optimization methods to secure the efficiency and scalability for the very large scale image data. Furthermore, the method improved over the conventional bag of words approach.

Cao et al. (2016) devised an object-level method for extracting the foreground objects from the videos. The method generated object-like regions for segmenting the candidates. Based on the corresponding map, the next set of frames was used for addressing the video segmentation problem and was used to acquire the most significant object region from each frame. Also, the shortest

path algorithm was used for obtaining the global optimum solution amongst the graphs. Kuo et al. (2012) developed a framework for controlling the content of the image and the related textual information in the videos. The method was used to augment the image using pertinent semantic features, employing the graphs linking the images. The framework was used to discover the pertinent semantic features by selecting the textual and visual image graphs in an unsupervised manner (see Zhu et al., 2016). The method used different optimization methods for enhancing its efficiency and.it helped to retain the representative tags.

Zhao, S. et al. (2015) developed a feature fusion method based on multimodal graph learning for the view-based 3D object retrieval. At first, different visual features were used that include 2D Fourier descriptor, 2D Zernike moments, and 2D Krawtchouk moments for describing the 3D objects. The method was used for measuring the similarity between the two 3D objects in the presence of multiple views. Lastly, multiple graphs were constructed using different features and the optimized weights of each graph were learned for fusing the features.

2.4. Studies involving the query-based techniques

This section addresses the different query-based techniques employed for object retrieval. Thus, Hou, Zhou and Siddique (2014) developed a compressed sensing retrieval approach for QEVR. The method used the CS and the conventional DCT for yielding improved efficiency in compression, and it used the CS for measuring the similarity between the keyframes. Thus, the similarity measure between the videos and shots was calculated in order to perform retrieval (see Ren et al., 2009; Arman et al., 1994) using video databases. The method demonstrated improved performance. Gao et al. (2012) developed a 3D object retrieval algorithm by addressing the limitations of camera array restrictions and certain camera constraints. The method proceeded using different sets of views, without camera constraint, based onown camera array settings for capturing views of 3-D objects. At first, the clustering of query views was done to produce the view clusters, which were then utilized for building the query models. For the precise 3D object comparison, a positive matching model and a negative matching model were identified and then were used for training using the positive and negative matched samples. However, the selection of best queryviews for users was challenging.

Qin et al. (2011) developed a fuzzy logic based data association algorithm for tracking multiple objects. Here, a fuzzy inference system based on knowledge was developed using fuzzy rules. These rules were utilized to describe the motion, shape and appearance models. However, the method still needed an improvement concerning the performance for tracking multiple objects. Pang et al. (2019) developed an event-driven visualization technique for processing

data from a huge video surveillance system. The method helped to improve the functioning of the security personnel. Gong et al. (2013) developed a 3D OR technique for defining the objects using object similarities for known classes, but the knowledge mining methods were not used, which affects the performance of the systemin extracting the significant features.

Yang et al. (2011) designed an object retrieval method using the information obtained from visual context and made use of this information for processing the feature-based query object representation. Qin, Wengert and Van Gool (2013) designed a probabilistic framework for modelling the similarities of features. Here, adaptive query distance was devised for evaluating the global similarities, but the parameter tuning was difficult in this methodin terms of the potential attainment ofoptimal performance. Petkovic and Jonker (2001) developed the COBRA video data model, which provided the framework for utilizing different knowledge-based methods to interpret the raw data in terms of semantic content. The model used video processing and pattern recognition techniques to provide flexibility to the system. This model utilized object and event grammars to formalize the descriptions of advanced concepts and to facilitate the extraction on the basis of spatio-temporal logic. The model supported stochastic modelling meant to identify the objects. In this method, initiating feature selection and training tasks were difficult.

Chen et al. (2014) developed a method for computing the similarities between contents for image measurements. The method was supposed to function on the basis of the available ranking information. The method proved to be efficient and allowed for performing quick calculations. Castanon et al. (2016) developed a content-based retrieval method using several videos. The aim was to recover video segments by determining objects. The method possessed the ability to retrieve the abnormal and recurrent activities using the videos, but involved high computational complexity while processing large videos. Bency et al. (2017) proposed a method for annotating the video libraries for tracking the objects using unseen video sequences. For each of the video sequences, a document representing the information on motion was generated. The method outperformed other methods using other surveillance datasets. Czúni and Rashad (2017) proposed an object retrieval approach in which the IMU sensors and camera were utilized for object retrieval. The method employed deep learning recognition and retrieval solutions for processing within an autonomous systems integrated with memory and computing power (see also Zhang, Huang and Liu, 2018; DeMenthon and Doermann, 2003). The method was fast and robust for different image descriptors and camera orientations. This method used the Hough transformation paradigm for evaluating the query results using different video frames. However, this method had problem in analyzing the probability of candidates.

2.5. Studies involving the feature-based techniques

We shall now present the results of the query with respect to the feature-based techniques. Zhang and Jeong (2017) designed a retrieval algorithm for different face images using the cloud computing platform. The method used the Harr face cascade classifier for determining the face images in the multimedia video. The method used the principal component analysis (PCA) for extracting the specific facial features and for extracting the feature information in the particular face. The method produced improved retrieval accuracy and minimized the matching time with improved retrieval efficiency and precision. Yang et al. (2013) designed a framework for retrieving the objects, based on object extraction and layer information integration. The object was detected using unconstrained videos with the help of the multimodal cues method.Furthermore, this method used object extraction algorithm based on GrabCut, which was used for separating the foreground object from the background. The method used a multi-layer object-level information integration strategy for improving the object retrieval and classification (see also Hong et al., 2016; Song et al., 2011), but the complexity of the method was high.

Hong et al. (2011) developed a method for retrieving objects using the videos. Initially, the feature descriptors were compressed for tracking the features between the consecutive frames and encode the tracked feature points. In this case, the algorithm was used for locating the objects by searching the position of related feature points in the frames, but lost some feature related information in the processing.Priyaa and Karthikeyan (2013) proposed an algorithm for implementing the tests on colour histograms and a retrieval algorithm for different video frames. The method determined the objects using Intuitionistic Fuzzy-based Hausdorff similarity measure for colour and shape by considering the membership and non-membership values of each frame. This method was used for improved identification and retrieval. Gong and Caldas (2011) developed a video interpretation method for providing guidance in selecting the computer vision algorithm. The method used an automated video interpretation technique for computing the productivity information, like cycle times, delays, and processes and thereby reduced the manual efforts.

Tang et al. (2011) used a word quantization technique for retrieving the objects using the spatial information and the contextual synonyms. The method used queries for retrieving the objects from huge datasets. This method had problems in recognizing the objects. Kanagamalliga and Vasuki (2018) developed a Movement Estimation and object tracking algorithm using videos. The method utilized in its functioning of the features of optical flow distribution. The method was not applicable for non-rigid objects. Fernandez-Beltran and Pla (2015) devised a generalized re-ranking algorithm for video object retrieval using the videos. The method used similarity propagation for reconstructing the vectors of queries. The algorithm was memory-wise and computationally effi-

cient, but the query processing was difficult in this method using the standard database. Chuang et al. (2014) designed spatial-temporal sampling for extracting the video objects and for measuring the spatial-temporal boundaries, based on the devised object model. This method used such computational approaches as dynamic programming and Hough voting to learn the optimal key-object codebook sequence from different video clips in dynamic programming. The detection of multiple video objects from a large database was difficult with this method.

Mitrea et al. (2014) developed a classification-based automated surveillance system for retrieving different objects with the aim of tracking the persons. The method used suitable motion detectors, feature extraction and classification methods for attaining improved classification accuracy. However, the recall significantly decreased for smaller samples. Cai et al. (2015) developed a video object tracking technique based on the Adaptive Hybrid Difference approach. The method extracted the features and used those features for producing the retrieval results. Gómez-Romero et al. (2011) developed a computer vision framework that aimed at constructing a symbolic model for tracking the objects and the related information. The method provided scene interpretation and achieved improvements in tracking for effective video retrieval. Li, Larson and Hanjalic (2015) developed a strategy using the pair-wise geometric relations and incorporated these relations in the procedure for minimizing the computational cost that made the technique suitable for huge-scale object retrieval (see also Sasithradevi, Roomi and Maragatham, 2017; Song et al., 2011). Czúni and Rashad (2018) developed a lightweight sensory and processing technique for retrieving the objects using the sensors with several modalities. In this study, the Hough framework was used to fuse optical and orientation information of the different views of different objects. The method used spatial-temporal perception technique for analyzing the initial measurement and for determining the hit rate. The method needed relatively little memory space and computations for the video object retrieval.

Arandjeloviç and Zisserman (2011) developed a 3D smooth object retrieval technique for retrieving the objects from huge video datasets. The method was used for determining the appearance, shape and viewpoint changes. Further, it used descriptors and quantization methods for retrieving the objects. This method had problemswith segmenting the classes. Zhao, Precioso and Cord (2011) designed a video object retrieval system on the basis of spatiotemporal data representation and a statistical learning tool for retrieving the video objects. Here, the video objects were tracked using real-time movie video shots. The method extracted the coherent features for the retrieval process. Also, an active learning strategy was introduced into the framework from Zhang et al. (2016b) for dealing with the unsupervised data. However, the method failed to incorporate the generative models into the video retrieval system. Hao et al. (2017) developed a stochastic multi-view hashing algorithm for construct-

ing the huge-scale NDVR system. The method used reliable mapping functions for converting different types of keyframe features using auxiliary information, like video-key frame association and ground truth. The method was used to estimate the retrieval scores.

Ma et al. (2010) developed a colour-feature model for determining the objects present in the video by converting the RGB pixels to a colour circle hue. The framework was devised on the basis of a colour feature model. The model showed improved accuracy and good performance using the devised colour feature model. The method failed to produce a feature vector for video object retrieval. Normally, models with bigger numbers of features are used to enhance the performance of the system,while in this method, a model with a lower number of features was used. Lai and Yang (2014) developed a method for video object retrieval using the trajectories of the desired objects. The method used a 3D graphical user interface to produce satisfactory results. The method, however, did not consider the machine learning approaches for tuning the system. Hu, Tang and Zhang (2008) developed the SIFT algorithm for addressing the problems of visual tracking, in which the appearance of tracked objects and scene background changes during tracking. The SIFT algorithm used the Euclidean distance between the object features and frames. The method showed improved performance in tracking video objects and in their retrieval.

2.6. Studies involving the fuzzy-based techniques

This section considers the application of the fuzzy-based techniques employed for video object retrieval purposes. In this context, we would like to mention the work of Liang-qun et al. (2018), who developed a videosurveillance methodology that contributed to tracking the trajectories of people by discovering questionable actions in a certain context. Initially, the technique applied image segmentation for locating the foreground objects using the scenes. Then, the detected blobs were used for determining the human target. Finally, the identified human trajectories were used for analyzing the behaviour of humans and detecting the mistrustful behaviour.

2.7. Studies based on machine learning techniques

Of the techniques based on machine learning we would like to mention the following ones: Gomez-Conde and Olivieri (2015) developed a DGTC method for capturing the fine and large-scale structures of the spatio-temporal optical flow field. In this methodology, a distance metric was used for transforming the query trajectory into the training set. The method developed improved the discovery process using full and short videos. Lin and Brandt (2010) developed a global BoF framework for supporting local matching. In this work, an algorithm was presented, which was used for performing object localization and local histogram matching in a simultaneous manner. Then, a localization scheme was used for imposing a weak spatial consistency for retrieving the objects.

2.8. Studies based on distance-metric learning technique

We would like to attract attention to the work of Joy and Peter (2018), who developed a colour-independent tracking approach for tracking visual objects in intricate, and hence hard to analyse, videos. Initially, the method was devised for determining the level of illumination using a fast discrete curvelet transform. Here, a metric was used for quantifying the distances between the successive frames for regulating the object templates to handle dynamic occlusions. This method had problems regarding its adaptability.

2.9. Studies based on other techniques

The remaining surveyed techniques, employed for the video object retrieval, referred to a variety of methodologies. Thus, Arroyo et al. (2015) designed a complete expert system, based on real-time detection of mistrustful behaviours in shopping malls. The method contributed several innovative proposals, which altogether formed a robust application that was able to track the trajectories and identify questionable actions. The method used a segmentation algorithm for locating the foreground objects from the scene. Fan, Wang and Huang (2017) developed a method for the video-based video object retrieval. The method was based on a k-reciprocal nearest neighbour structure using the image space. During the query time, the information obtained from the process was used for treating different distance measures, which led to the re-ranking of different retrieved images. The method outperformed many other methods, but its memory overhead was high. Cheng and Hwang (2011) developed an incremental topic model for effective video object retrieval. The method addressed the efficiency issues of the topic extraction methods for obtaining the targeted objects. However, the quantization techniques for extracting the objects were not used so that the performance was reduced. Araujo and Girod (2018) developed a technique, referred to asymmetric, based on Fisher vectors,which explored the query for determining the objects. The method used video descriptors and Fisher vectors for producing video segments. The deep learning techniques were not used in this method, which affected the performance results.

Ding et al. (2017) designed Survsurf, a system for retrieving images of humansusing very large scale video data, which used data characteristics and processing tools for video object retrieval. The method used motion information for video segmentation. The data unit obtained after the segmentation process is known as M-clip. The method accelerated the processing of huge volumes of data by processing significant motion vectors.

3. Research gaps and issues

Despite the development of quite many distinct video object retrieval methodologies, there exist certain limitations, which must be addressed for attaining the truly effective retrieval of objects. This section considers the research gaps

and issues based on the characteristics of different methods employed for the retrieval of objects, appearing in the literature surveyed.

The retrieval of videos over a big database needs knowledge of the videos and textual information (see Inbavalli, Sathya and Manjula, 2017). The textual queries were effective for extracting the precise knowledge by searching and indexing the structured databases, but the conventional techniques were unable to work well with an unstructured database while finding the video clips without the video information. The goal of query-based techniques was to retrieve the video from the huge set of videos. However, obtaining similar videos from the database is a major problem in the video object retrieval process. The compressed sensing (CS) based algorithms were used for Query by Example Video Retrieval (QEVR) (see Hou, Zhou and Siddique, 2014; Gomez-Conde and Olivieri, 2015).

The feature-based techniques are flexible and scalable, but the object descriptions become complex, posing problems for retrieving the objects. These techniques face the high computational cost due to the storage needs, related to the feature descriptors (Hong et al., 2011). The feature-based techniques turned out to be inappropriate for complex image analysis. Hence, the provided combined images were not appropriately recognized. In the same context, the automation process for the rule generation was hard coded in complex image analysis approach, but the computational cost was high (see also Lin and Brandt, 2010). The accuracy range achieved by the devised feature-based methods was not equal to the expected range (Yang et al., 2013). Likewise, the Near Duplicate Video technique did not apply for the complex video object retrieval systems (Phalke and Jahirbadkar, 2018).

The distinct challenge, faced by the fuzzy classifier was related to the requirement of a large training sample and this method did not adequately apply to the recognition of the objects (Liang-qun et al., 2018). Namely, the fuzzy technique was not capable of recognizing the objects. The video object retrieval performance, achieved by the feature-based methods was not equal to the expected performance quality (Priyaa and Karthikeyan, 2013). Quite similarly, the query-based systems developed did not attain the anticipated rate of retrieval (Gao et al., 2012). This method had problems in achieving adequate performance for the crowded scenes (Ghuge, Prakash and Ruikar, 2018b).

The video object retrieval experienced, in general, the difficulties from non-uniform illumination, noise, complicated background, variations in font size and style. These issues lead to complications during the process of video object retrieval (Haseyama, Ogawa and Yagi, 2013). Finally, the assessment of a video object retrieval system is relatively subjective and lacks standards that would make easier the introduction of improvements in the object retrieval task (Yang et al., 2013).

4. Analysis and discussion

This section presents the analysis based on different techniques for the video object retrieval regarding the publication year, adapted methods, datasets used, evaluation metrics, software tools, and performance evaluation values.

4.1. Statistics based on publication year

The distribution of the papers and reports surveyed in terms of publication year is shown here in Fig. 2.

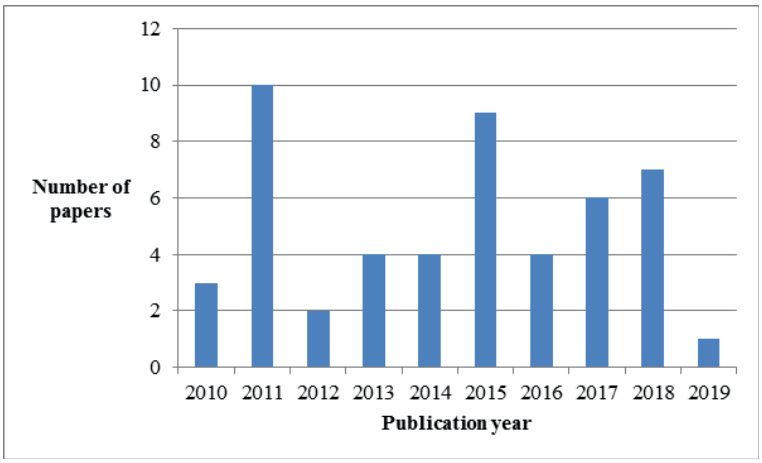

Figure 2. Statistics based on publication year

4.2. Statistics based on techniques

We now provide the illustration for the shares of the particular groups of techniques in the publications surveyed. This illustration is given in Fig. 3. Based on this figure, it can be noted that 40% of the research used the feature-based techniques, 24% of these references used other techniques, 10% of the references used the graph-based techniques, similarly, 10% of the references considered based on query techniques, 8% of the research papers used deep learning-based techniques, 4% of the studies based on machine learning techniques, and 2% of the research work reviewed used the distance metric learning-based techniques, just likethe fuzzy-based techniques. Thus, obviously, the feature-based techniques are most commonly employed for video object retrieval. (Note that we consider here just the sole or leading methodologies, and that is why the shares add up to 100%.)

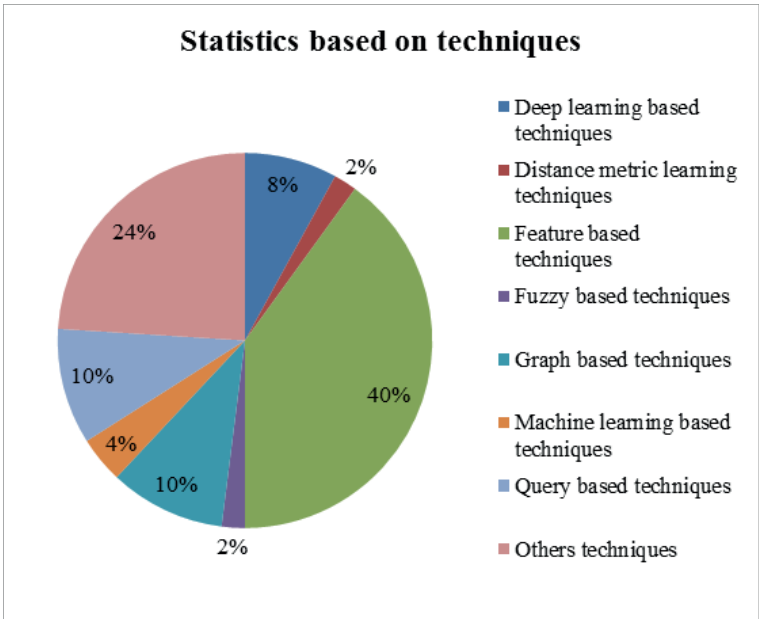

Figure 3. Shares of the categorized video object retrieval techniques in the surveyed set of papers

4.3. Distribution of work based on the software tools

This section concerns the software toolsused in the studies, reported in the research papers. Table 1shows the software tools adapted for performing effective video object retrieval. The major software tools employed in the investigations, reported in the research papers are Java, OpenCV, C++, Visual C++, SQL server, and MATLAB. From table 1, it can be evaluated that MATLAB is a frequently used software tool for effective video object retrieval.

4.4. Statistics based on datasets used

This subsection concerns the distribution of the work reviewed, based on the datasets used in the research studies. Various datasets employed for the effective video object retrievals are presented in Fig. 4. The frequently used datasets employed for the video object retrieval are Oxford dataset, PETS database, personal video databases, and CAVIAR Dataset. Other datasets considered are eBay dataset, ETH dataset, Flickr dataset, ImageNet dataset, SHREC's dataset, UCF dataset, VOC challenge dataset, and Wang databases.

Table 1. Analysis based on the software tool

Implementation tools	**Research papers**
C++	Gong and Caldas (2011); Liang-qun et al. (2018); Qin et al. (2011); Joy and Peter (2018)
Java	Li, Larson and Hanjalic (2015)
MATLAB	Priyaa and Karthikeyan (2013); Pang et al. (2019); Tolias, Sicre and Jegou (2015); Castanon et al. (2016); Bency et al. (2017); Czúni and Rashad (2017, 2018); Hao et al. (2017); Kuo et al. (2011); Kuo et al. (2012)
OpenCV	Arroyo et al. (2015); Chuang et al. (2014); Cai et al. (2015); Zhao, Precioso and Cord (2011); Ding et al. (2017); Ma et al. (2010)
SQL server	Hou, Zhou and Siddique (2014)
Visual C++	Zhang and Jeong (2017); Hu, Tang and Zhang (2008)

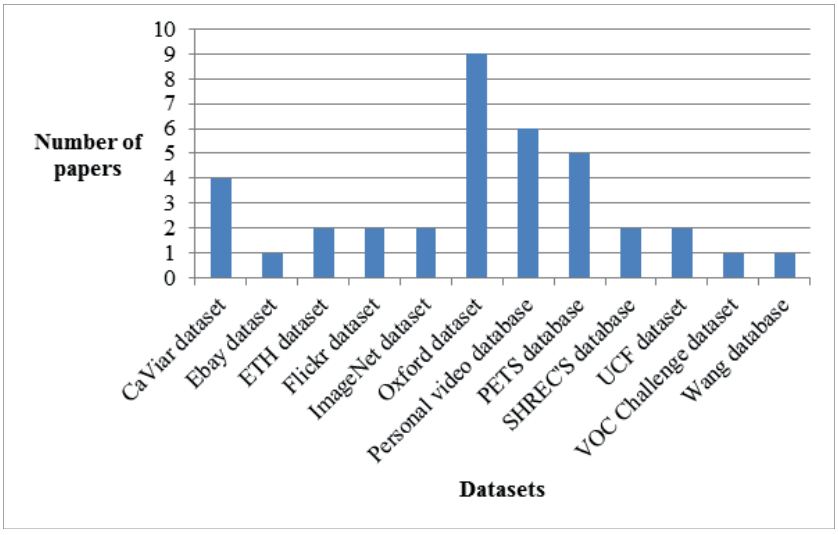

Figure 4. Distributon of the datasets employed

4.5. Analysis of the performance metrics employed

The present section is devoted to the use of various performance metrics for assessing the quality of video object retrieval. The distribution of different metrics among the studies here surveyed is provided in Table 2.

It can be gathered from this table, without any surprise, indeed, that the commonly used performance metrics includemean average precision (MAP), accuracy, recall, precision, and F-measure. Other metrics, some of them also having been used in several cases by the respective authors, include time per query, false alarm rate, Jaccard similarity, MOTA, MOTP, success rate, and gain in %. Attention is attracted by the reference of Kanagamalliga and Vasuki (2018), who used, as the only ones, both the false alarm rate and the Jaccard similarity.

4.6. Studies classified according to the values of the performance metrics

In this section we present the analysis based onthe values, attained by the performance metrics in the particular studies. The analysis is shown in terms of accuracy, MAP, accuracy, recall, and precision.

The values of recall reported

The publications reviewed are here presented with regard to the ranges of recall, reported in them. Table 3 shows the respective publications according to the recall ranges of 0.0-0.5, 0.5-1.0, and 1.0-1.5, respectively.

The values of accuracy reported

Now, in turn, we present the results, concerning the achieved values of the accuracy measure. Table 4 shows these values for the respective studies, referred to here, using three value ranges, 70%-80%, 80%- 90%, and 90%-100%. It can be concluded from this table that the methodologies proposed performed very differently, indeed, in terms of this measure.

The values of MAP reported

The statistics from the survey regarding MAP (mean average precision) is provided in Table 5. The values of the MAP parameter are divided up into six ranges, namely 0.0-0.5, 0.5-0.6, 0.6-0.7, 0.7-0.8, 0.8-0.9, and 0.9-1. According to this table, the highest values of MAP (between 0.9 and 1.0) were reported in the papers by Hao et al. (2017) and Li, Larson and Hanjalic (2015).

Table 2. References according to the performance metrics used

Performance metrics	**References**
Accuracy	Gomez-Conde & Olivieri (2015); Priyaa & Karthikeyan (2013); Gong & Caldas (2011); Fan, Wang & Huang (2017); Pang et al. (2019); Guo et al. (2015); Guo, Wang & Lu (2016); Chen et al. (2014); Ma et al. (2010); Hu, Tang & Zhang (2008)
False alarm rate	Kanagamalliga & Vasuki (2018)
F-measure	Hou, Zhou & Siddique (2014); Kanagamalliga & Vasuki (2018); Fernandez-Beltran & Pla (2015)
Gain %	Gong et al. (2013); Yang et al. (2011); Arandjeloviç & Zisserman (2011); Zhao et al. (2015)
Jaccard similarity	Kanagamalliga & Vasuki (2018)
MAP	Lin & Brandt (2010); Yang et al. (2013); Lou et al. (2017); Tang et al. (2011); Qin et al. (2011); Araujo & Girod (2018); Guo et al. (2015); Yang et al. (2011); Li, Larson & Hanjalic (2015)
MOTA	Liang-qun et al. (2018); Bency et al. (2017)
MOTP	Shan (2010); Li, Zaïane & Tauber (1999)
Precision	Kuo et al. (2011); Hou, Zhou & Siddque (2014); Zhang & Jeong (2017); Priyaa & Karthikeyan (2013); Lou et al. (2017); Gao et al. (2012); Fernandez-Beltran & Pla (2015); Chuang et al. (2014); Ding et al. (2017); Cao et al. (2016)
Recall	Zhang et al. (2016a); Hou, Zhou & Siddique (2014); Mitrea et al. (2014); Zhao, Precioso & Cord (2011)
Success rates	Joy & Peter (2018)
Time per query	Lin & Brandt (2010); Hong et al. (2011); Castanon et al. (2016); Hao et al. (2017)

Table 3. Analysis based on recall

Recallrange	**References**
<0.5	Kuo et al. (2012)
0.5 to <1	Hou, Zhou & Siddique (2014); Priyaa & Karthikeyan (2013); Mitrea et al. (2014); Zhao, Precioso & Cord (2011); Ding et al. (2017)
>= 1	Zhang et al. (2016a)

Table 4. Analysis based on accuracy

Accuracy range	**References**
70% to 80%	Hu, Tang & Zhang (2008); Priyaa & Karthikeyan (2013)
80 to 90%	Guo et al. (2015); Guo, Wang & Lu (2016); Chen et al. (2014); Ma et al. (2010)
90 to 100%	Gong & Caldas (2011); Gomez-Conde & Olivieri (2015); Fan, Wang & Huang (2017)

Table 5. Analysis based on MAP

MAP range	**References**
<0.5	Zhu et al. (2016); Qin et al. (2011)
>= 0.5 to <0.6	Cao et al. (2016)
>= 0.6 to <0.7	Lin & Brandt (2010); Guo et al. (2015); Tolias, Sicre & Jegou (2015)
>= 0.7 to <0.8	Yang et al. (2013); Lou et al. (2017); Tang et al. (2011); Araujo & Girod (2018)
>= 0.8 to <0.9	Yang et al. (2011); Qin et al. (2013)
>= 0.9 to <1	Hao et al. (2017)

The values of precision reported

The statistics, based on the results from the here reported survey, regarding the achieved values of the precision measure, are provided in Table 6. The results in Table 6 are shown with respect to the precision parameter values divided up into four ranges, i.e. above 90%, 80-90%, 70-80%, and below 70%. Like in some of the preceding cases, the values of this measure are highly differentiated among the studies surveyed. Of special interest is the group of studies, which reported high values of this criterion, i.e. between 90% and 100%.

Table 6. Analysis based on precision values

Precision range	**Research papers**
<70	Kuo et al. (2011); Mitrea et al. (2014)
>= 70 to <80	Priyaa & Karthikeyan (2013); Lou et al. (2017)
>= 80 to <90	Hou, Zhou & Siddique (2014); Zhang & Jeong (2017)
>= 90 to <100	Gao et al. (2012); Chuang et al. (2014); Ding et al. (2017); Cao et al. (2016)

4.7. Statistics according to journals

We shall now present the distribution of the papers reviewed among the journal publishers. The distribution among the publishers is shown in Fig. 5 in the form of the pie chart of shares for each of the publishers accounted for. It can easily be seen that the biggest share of the papers reviewed were published by IEEE (44%), followed by Springer and Elsevier (18% and 17%, respectively), and then ACM (12%), the remaining 9% of the papers surveyed having been published by other publishing houses.

5. Conclusion

In this survey, we address the variety of techniques that are employed for achieving effective retrieval of objects from the video material. The techniques have been categorized for the purposes of this survey into relatively broad classes. The survey classifies also the investigated techniques in terms of publication years, methodologies adopted, datasets used, performance metrics used and implementation tools, as well as values of performance metrics on the basis of 50 research papers, appearing to be representative for the field.

The techniques utilized for the video object retrieval have been classified into eight groups, namely deep learning techniques, graph-based techniques, query-based techniques, feature-based techniques, fuzzy techniques, machine learning-based techniques, distance metric learning-based technique, and other,

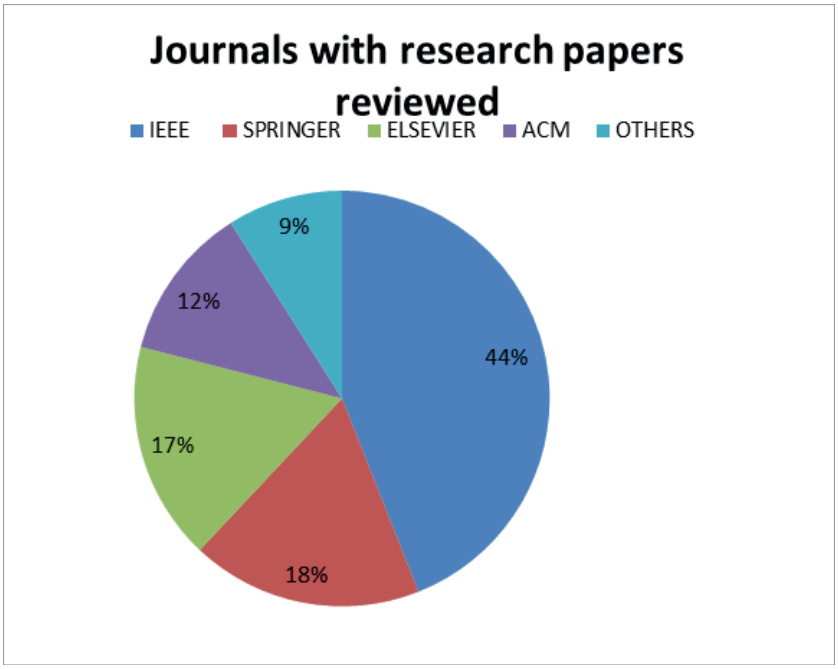

Figure 5. Statistics based on journals

remaining techniques. Besides, the research gaps and the issues in the video object retrieval techniques are presented in this paper in order to suggest effective future scope of research. The most commonly used techniques for attaining effective video object retrieval are feature-based. The majorchallenge faced by the video object retrieval systems is that most of the existing techniques cannot effectively detect objects in the crowded videos due to occlusions. This appears to be the major limitation that needs to be addressed in the future by adopting more advanced video object retrieval techniques.

References

Ahmed, S.A., Dogra, D.P., Kar, S. and Roy, P.P. (2019) Trajectory-based surveillance analysis: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*, **29**(7): 1985-1997.

Anjulan, A. and Canagarajah, N. (2007) Object-based video retrieval with local region tracking. *Signal Processing, Image Communication*, **22**(7-8): 607-621.

Arandjelović, R. and Zisserman, A. (2011) Smooth object retrieval using a bag of boundaries. In: *Proceedings of International Conference on Computer Vision*, IEEE, 375-382.

Araujo, A. and Girod, B. (2018) Large-scale video retrieval using image

queries. *IEEE Transactions on Circuits and Systems for Video Technology*, **28**(6): 1406-1420.

Ardizzone, E. and La Cascia, M. (1997) Automatic video database indexing and retrieval. *Multimedia Tools and Applications,* 4: 29-56.

Arman, F., Depommier, R., Hsu, A. and Chiu, M.Y. (1994) Content-based browsing of video sequences. In: *Proceedings of the Second ACM International Conference on Multimedia*, 97-103.

Arroyo, R., Yebes, J.J., Bergasa, L.M., Daza, I.G. and Almazán, J. (2015) Expert video-surveillance system for real-time detection of suspicious behaviors in shopping malls. *Expert Systems with Applications*, **42**(21): 7991-8005.

Aslandogan, Y.A. and Yu, C.T. (1999) Techniques and systems for image and video retrieval. *IEEE Transactions on Knowledge and Data Engineering*, **11**(1): 56-63.

Bency, A.J., Karthikeyan, S., De Leo, C., Sunderrajan, S. and Manjunath, B.S. (2017) Search tracker: human-derived object tracking in the wild through large-scale search and retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, **27**(8): 1803-1814.

Birkas, D., Birkas, K. and Popa, T. (2016) A mobile system for scene monitoring and object retrieval. In: *Proceedings of the 29th International Conference on Computer Animation and Social Agents,* ACM, 83-88.

Broilo, M., Piotto, N., Boato, G., Conci, N. and De Natale, F.G. (2010) Object trajectory analysis in video indexing and retrieval application. In: *Video Search and Mining*, Springer, Berlin, Heidelberg, 3-32.

Cai, Z., Liang, Y., Hu, H. and Luo, W. (2015) Offline video object retrieval method based on color features. In: *Proceedings of International Symposium on Computational Intelligence and Intelligent Systems*, Springer, 495-505.

Cao, X., Wang, F., Zhang, B., Fu, H. and Li, C. (2016) Unsupervised pixel-level video foreground object segmentation via shortest path algorithm. *Neurocomputing,* 172, 235-243.

Castanon, G., Elgharib, M., Saligrama, V. and Jodoin, P.M. (2016) Retrieval in long-surveillance videos using user-described motion and object attributes. *IEEE Transactions on Circuits and Systems for Video Technology,* **26**(12): 2313-2327.

Chen, Y., Li, X., Dick, A. and Hill R. (2014) Ranking consistency for image matching and object retrieval. *Pattern Recognition,* **47**(3): 1349-1360.

Cheng, H.Y. and Hwang, J.N. (2011) Integrated video object tracking with applications in trajectory-based event detection. *Journal of Visual Communication and Image Representation,* **22**(7): 673- 685.

Chou, C.L., Chen, H.T. and Lee, S.Y. (2015) Pattern-based nearduplicate video retrieval and localization on web-scale videos. *IEEE Transactions on Multimedia*, **17**(3): 382-395.

Chuang, C.H., Cheng, S.C., Chang, C.C. and Chen, P.P. (2014) Model

based approach to spatial-temporal sampling of video clips for video object detection by classification. *Journal of Visual Communication and Image Representation,* **25**(5): 1018-1030.

Czúni, L. and Rashad, M. (2017) The use of IMUs for video object retrieval in lightweight devices. *Journal of Visual Communication and Image Representation*, 48, 30-42.

Czúni, L. and Rashad, M. (2018) Lightweight Active Object Retrieval with Weak Classifiers. *Sensors,* **18**(3): 801.

DeMenthon, D. and Doermann, D. (2003) Video retrieval using spatio-temporal descriptors. In: *Proceedings of the Eleventh ACM International Conference on Multimedia*, 508-517.

Dimitrova, N. and Golshani, F. (1995) Motion recovery for video content classification. *ACM Transactions on Information Systems* (TOIS), **13**(4): 408-439.

Ding, S., Li, G., Li, Y., Li, X., Zhai, Q., Champion, A.C., Zhu, J., Xuan, D. and Zheng, Y.F. (2017) Survsurf: human retrieval on large surveillance video data. *Multimedia Tools and Applications,* **76**(5): 6521-6549.

Eidenberger, H., Breiteneder, C. and Hitz, M. (2002) A framework for visual information retrieval. In: *International Conference on Advances in Visual Information Systems,* 105-116, Springer.

Fan, C.T., Wang, Y.K. and Huang, C.R. (2017) Heterogeneous information fusion and visualization for a large-scale intelligent video surveillance system. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, **47** (4), 1-12.

Farag, W.E. and Abdel-Wahab, H. (2003) A human-based technique for measuring video data similarity. In: *Proceedings of the Eighth IEEE Symposium on Computers and Communications (ISCC)*, 769-774.

Fernandez-Beltran, R. and Pla, F. (2015) Incremental probabilistic latent semantic analysis for video retrieval. *Image and Vision Computing,* 38, 1-12.

Gao, Y., Tang, J., Hong, R., Yan, S., Dai, Q., Zhang, N. and Chua, T.S. (2012) Camera constraint-free view-based 3-D object retrieval. *IEEE Transactions on Image Processing,* **21**(4): 2269-2281.

Ghuge, C.A., Prakash, V.C. and Ruikar, S.D. (2018a) Weighed query specific distance and hybrid NARX neural network for video object retrieval. *The Computer Journal,* **63**(11): 1738-1755.

Ghuge, C.A., Prakash, V. C. and Ruikar, S.D. (2018b) Query-specific distance and hybrid tracking model for video object retrieval. *Journal of Intelligent Systems*, **27**(2): 195-212.

Ghuge, C.A., Prakash, V. C. and Ruikar, S.D. (2018c) Support vector regression and extended nearest neighbor for video object retrieval. *Evolutionary Intelligence.* DOI https://doi.org/10.1007/s12065-018-0176-y

Gomez-Conde, I. and Olivieri, D.N. (2015) A KPCA spatio-temporal differential geometric trajectory cloud classifier for recognizing human actions in a CBVR system. *Expert Systems with Applications,* **42**(13): 5472-5490.

Gomez-Romero, J., Patricio, M.A., Garcia, J. and Molina, J.M. (2011) Ontology-based context representation and reasoning for object tracking and scene interpretation in video. *Expert Systems with Applications,* **38**(6): 7494-7510.

Gong, B., Liu, J., Wang, X. and Tang, X. (2013) Learning semantic signatures for 3D object retrieval. *IEEE Transactions on Multimedia,* **15**(2): 369-377.

Gong, J. and Caldas, C.H. (2011) An object recognition, tracking, and contextual reasoning-based video interpretation method for rapid productivity analysis of construction operations. *Automation in Construction*, **20**(8): 1211-1226.

Guo, H., Wang, J. and Lu, H. (2016) Multiple deep features learning for object retrieval in surveillance videos. *IET Computer Vision,* **10**(4): 268-272.

Guo, H., Wang, J., Xu, M., Zha, Z.J. and Lu, H. (2015) Learning multi-view deep features for small object retrieval in surveillance scenarios. In: *Proceedings of the 23rd ACM International Conference on Multimedia*, 859-862.

Hao, Y., Mu, T., Hong, R., Wang, M., An, N. and Goulermas, J.Y. (2017) Stochastic multiview hashing for large-scale near-duplicate video retrieval. *IEEE Transactions on Multimedia,* **19**(1): 1-14.

Haseyama, M., Ogawa, T. and Yagi, N. (2013) A Review of Video Retrieval Based on Image and Video Semantic Understanding. *ITE Transactions on Media Technology and Applications* (MTA), **1**(1).

Hong, C., Li, N., Song, M., Bu, J. and Chen, C. (2011) An efficient approach to content-based object retrieval in videos. *Neurocomputing,* **74**(17): 3565-3575.

Hong, R., Hu, Z., Wang, R., Wang, M. and Tao, D. (2016) Multi-view object retrieval via multi-scale topic models. *IEEE Transactions on Image Processing,* **25**(12): 5814-5827.

Hou, S., Zhou, S. and Siddique, M.A. (2014) A compressed sensing approach for query by example video retrieval. *Multimedia Tools and Applications*, **72**(3): 3031-3044.

Hu, W., Tan., Wang, L. and Maybank, S. (2004) A survey on visual surveillance of object motion and behaviours. *IEEE Transaction on Systems, Man, and Cybernetics*, *Part C* (*Applications and Reviews*), **34**(3):334-352.

Hu, W., Xie, N., Li, L., Zeng, X. and Maybank, S. (2011) A survey on visual content-based video indexing and retrieval. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, **41**(6): 797-819.

Hu, X., Tang, Y. and Zhang, Z. (2008) Video object matching based on SIFT algorithm. In: *Proceedings of International Conference on Neural Networks and Signal Processing,* IEEE, 412-415.

Inbavalli, M., Sathya, G. and Manjula, R. (2017) A Study of Multimedia Visuals and Information Retrieval and Techniques. *International Journal of Scientific & Engineering Research*, **8**(4): 118-121.

Ji, X. and Liu, H. (2009) Advances in view-invariant human motion analysis: a review. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 50, 456-465.

Jin, R. and Kim, J (2017) Tracking feature extraction techniques with improved SIFT for video identification. *Multimedia Tools Appl.*, **76**(4):5927-5936.

Joy, E. and Peter, J.D. (2018) Visual tracking with conditionally adaptive multiple template update scheme for intricate videos. *Multimedia Systems,* **24**(2): 175-194.

Kanagamalliga, S. and Vasuki, S. (2018) Contour-based object tracking in video scenes through optical flow and Gabor features. *Optik,* 157, 787-797.

Kim, C. and Hwang, J.N. (2000) An integrated scheme for object based video abstraction. In: *Proceedings of the Eighth ACM International Conference on Multimedia,* 303-311.

Kuo, Y.H., Cheng, W.H., Lin, H.T. and Hsu, W.H. (2012) Unsupervised semantic feature discovery for image object retrieval and tag refinement. *IEEE Transactions on Multimedia,* **14**(4): 1079-1090.

Kuo, Y.H., Lin, H.T., Cheng, W.H., Yang, Y.H. and Hsu, W.H. (2011) Unsupervised auxiliary visual words discovery for large-scale image object retrieval. In: *Computer Vision and Pattern Recognition (CVPR 2011)*, IEEE, 905-912.

Lai, Y.H. and Yang, C.K. (2014) Video object retrieval by trajectory and appearance. *IEEE Transactions on Circuits and Systems for Video Technology,* **25**(6): 1026-1037.

Li, X., Larson, M. and Hanjalic, A. (2015) Pairwise geometric matching for large-scale object retrieval. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5153-5161.

Li, Z.N., Zaïane, O.R. and Tauber, Z. (1999) Illumination invariance and object model in content-based image and video retrieval. *Journal of Visual Communication and Image Representation*, **10**(3): 219-244.

Liang-qun, L., Xi-yang, Z., Zong-Xiang, L. and Wei-Xin, X. (2018) Fuzzy logic approach to visual multi-object tracking. *Neurocomputing,* 281, 139-151.

Lin, Z. and Brandt, J. (2010) A local bag-of-features model for large-scale object retrieval. In: *Proceedings of European Conference on Computer vision*, Springer, 294-308.

Lou, Y., Bai, Y., Lin, J., Wang, S., Chen, J., Chandrasekhar, V., Duan, L.Y., Huang, T., Kot, A.C. and Gao, W. (2017) Compact

deep invariant descriptors for video retrieval. In: *Proceedings of Data Compression Conference*, IEEE, 420-429.

Ma, X., Huang, H., Cai, Z., Wang, C. and Zou, Y. (2010) Video Object Retrieval Based on Color Feature Modeling. In: *Proceedings of International Conference on Machine Vision and Human-machine Interface,* IEEE, 101-104.

Mezaris, V., Kompatsiaris, I. and Strintzis, M.G. (2004) Region-based image retrieval using an object ontology and relevance feedback. *EURASIP Journal on Advances in Signal Processing*, 6, 231-946.

Mitrea, C.A., MironicÂ, I., Ionescu, B. and Dogaru, R. (2014) Multiple instance-based object retrieval in video surveillance Dataset and evaluation. In: *Proceedings of IEEE 10th International Conference on Intelligent Computer Communication and Processing,* 171-178.

Padmakala, S., AnandhaMala, G.S. and Shalini, M. (2011) An effective content based video retrieval utilizing texture, color and optimal key frame features. In: *2011 IEEE International Conference on Image Information Processing,* Shimla, India, 1-6.

Pang, S., Ma, J., Zhu, J., Xue, J. and Tian, Q. (2019) Improving object retrieval quality by integration of similarity propagation and query expansion. *IEEE Transactions on Multimedia*, **21**(3): 760-770.

Petkovic, M. and Jonker, W. (2001) Content-based video retrieval by integrating spatiotemporal and stochastic recognition of events. In: *Proceedings of IEEE Workshop on Detection and Recognition of Events in Video*, IEEE, 75-82.

Phalke, D.A. and Jahirbadkar, S. (2018) Systematic Review of Near Duplicate Video Retrieval Techniques. *International Journal of Pure and Applied Mathematics*, **118**(24): 1-11.

Priyaa, D.S. and Karthikeyan, S. (2013) An Innovative Approach for Video Object Retrieval based on Color and Shape Using Intuitionistic Fuzzy Hausdorff Distance. *International Journal of Computer Technology and Applications*, **4**(4): 710.

Qin, D., Gammeter, S., Bossard, L., Quack, T. and Van Gool, L. (2011) Hello neighbor: Accurate object retrieval with k-reciprocal nearest neighbors. In: *Proceedings of IEEE CVPR*, 777-784.

Qin, D., Wengert, C. and Van Gool, L. (2013) Query adaptive similarity for large scale object retrieval. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1610-1617.

Ren, W., Singh, S., Singh, M. and Zhu, Y.S. (2009) State-of-the art on spatio-temporal information-based video retrieval. *Pattern Recognition*, **42**(2): 267-282.

Sasithradevi, A., Roomi, S.M.M. and Maragatham, G. (2017) Content based video retrieval via object based approach. In: *TENCON 2017- 2017 IEEE Region 10 Conference*, 781-787.

Shan, C. (2010) Face recognition and retrieval in video. In: *Video Search and Mining*, 287, 235-260.

Sivic, J. and Zisserman, A. (2003) Video Google: A text retrieval approach to object matching in videos. *Proceedings of Ninth IEEE International Conference on Computer Vision*, IEEE, 1470.

Sivic, J. and Zisserman, A. (2004) Efficient visual content retrieval and mining in videos. In: *Pacific-Rim Conference on Multimedia*, Springer, 471-478.

Sivic, J. and Zisserman, A. (2008) Efficient visual search for objects in videos. In: *Proceedings of the IEEE*, **96**(4): 548-566.

Sivic, J., Everingham, M. and Zisserman, A. (2005) Person spotting: video shot retrieval for face sets. In: *International Conference on Image and Video Retrieval*, Springer, 226-236.

Sivic, J., Schaffalitzky, F. and Zisserman, A. (2004) Object level grouping for video shots. In: *European Conference on Computer Vision*, Springer, 85-98.

Sivic, J., Schaffalitzky, F., and Zisserman, A. (2006) Object level grouping for video shots. *International Journal of Computer Vision*, **67**(2): 189-210.

Smeaton, A.F. and Browne, P. (2006) A usage study of retrieval modalities for video shot retrieval. *Information Processing & Management*, **42**(5): 1330-1344.

Song, J., Yang, Y., Huang, Z., Shen, H.T. and Hong, R. (2011) Multiple feature hashing for real-time large scale near-duplicate video retrieval. In: *Proceedings of the 19th ACM International Conference on Multimedia*, 423-432.

Tang, W., Cai, R., Li, Z. and Zhang, L. (2011) Contextual synonym dictionary for visual object retrieval. In: *Proceedings of the 19th ACM International Conference on Multimedia*, ACM, 503-512.

Tolias, G., Sicre, R. and Jégou, H. (2015) Particular object retrieval with integral max-pooling of CNN activations. *arXiv* preprint arXiv:1511.05879

Valdés, V. and Martínez, J.M. (2011) Efficient video summarization and retrieval tools. In: *9th International Workshop on Content-Based Multimedia Indexing (CBMI)*, IEEE, 43-48.

Van Den Hengel, A., Dick, A., Thormaehlen, A., Ward, T.B. and Torr, P.H. (2007) VideoTrace: rapid interactive scene modelling from video. *ACM Transactions on Graphics (ToG),* **26**(3).

Visser, R., Sebe, N. and Bakker, E. (2002) Object recognition for video retrieval. In: *International Conference on Image and Video Retrieval*, Springer, 262-270.

Yan, R. and Hauptmann, A.G. (2007) A review of text and image retrieval approaches for broadcast news video. *Information Retrieval,* **10**: 4-5, 445-484.

Yang, H., Qu, S., Zhu, F. and Zheng, Z (2018) Robust objectness tracking with weighted multiple instance learning algorithm. *Neurocomputing*, 288: 43-53.

Yang, L., Geng, B., Cai, Y., Hanjalic, A. and Hua, X.S. (2011) Object

retrieval using visual query context. *IEEE Transactions on Multimedia*, **13**(6): 1295-1307.

Yang, Y., Fleites, F.C., Wang, H. and Chen, S.C. (2013) An automatic object retrieval framework for complex background. In: *Proceedings of IEEE International Symposium on Multimedia*, 374-377.

Yeo, B.L. and Yeung, M.M. (1997) Retrieving and visualizing video. *Communications of the ACM*, **40**(12): 43-53.

Zhang, Y., Jiang, F., Rho, S., Liu, S., Zhao, D. and Ji, R. (2016a) 3D object retrieval with multi-feature collaboration and bipartite graph-matching. *Neurocomputing*, 195, 40-49.

Zhang, H., Cao, X., Ho, J.K.L and Chow, T.W.S. (2016b) Object Level Video Advertising: An Optimization Framework. *IEEE Transactions on industrial informatics*, **13**(6): 1295-1307.

Zhang, D., Han, J., Jiang, L., Ye, S. and Chang, X. (2017) Revealing event saliency in unconstrained video collection. *IEEE Transactions on Image Processing*, **26**(4):1746-1758.

Zhang, H., Ji, Y., Huang, W. and Liu, L. (2018) Sitcom-star-based clothing retrieval for video advertising: a deep learning framework. *Neural Computing and Applications*, **13**(6): 1295-1307.

Zhang, N. and Jeong, H.Y. (2017) A retrieval algorithm for specific face images in airport surveillance multimedia videos on cloud computing platform. *Multimedia Tools and Applications*, **76**(16): 17129-17143.

Zhao, F., Huang, Y., Wang, L. and Tan, T. (2015) Deep semantic ranking based hashing for multi-label image retrieval. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1556-1564.

Zhao, S., Precioso, F. and Cord, M. (2011) Spatio-Temporal Tube data representation and Kernel design for SVM-based video object retrieval system. *Multimedia Tools and Applications*, **55**(1): 105-125.

Zhao, S., Yao, H., Zhang, Y., Wang, Y. and Liu, S. (2015) View based 3D objects retrieval via multi-modal graph learning. *Signal Processing*, 112, 110-118.

Zhu, L., Shen, J., Xie, L. and Cheng, Z. (2016) Unsupervised visual hashing with semantic assistant for content-based image retrieval. *IEEE Transactions on Knowledge and Data Engineering*, 29(2): 472-486.