A Comprehensive Survey On Various Semantic Based Video/Image Retrieval Techniques

Tamil Priya D, Divya Udayan J

Abstract: ***Semantic based video/image retrieval is a most critical issue in the multimedia search enginerelated research. Multimedia content annotation improves the accuracy of semantic based content retrieval system where multimedia content annotation can be done in two ways: One is content based annotation and another is context based annotation. The content based annotation of images and videos consist of both low level and high level features which would be derived from the detailed pixel intensity information of images whereas context information would provide the semantic details about the images/videos. In this analysis work, discussion about the various techniques used to retrieve the multimedia content from the web engines with the semantic knowledge awareness is provided. The various research methods that intend to perform the annotation on multimedia contents is discussed in detailed. And also, this analysis work provides the discussion about the techniques in terms of their merits and demerits. Finally this work also provides the numerical evaluation of eachtechnique based on their accuracy in the videoretrieval outcome. The overall evaluation of the surveyed methodologies is performed and provides discussion about the outcome obtained from each technique.***

Index Terms: Semantic based video retrieval system, context information, content information, low level feature, high level features, and image annotation.

I. INTRODUCTION

The dramatic growth of multimedia data for few decades on web has paved the way for content and context based image annotation frameworks in organizing of data retrieval and indexing techniques [1]. This attracts the focus of various researchers to concentrate on Content based image retrieval (CBIR) techniques. As of now, CBIR has turned out to be a proficient strategy for observing, perusing, and retrieval of advanced images from large collection of databases [2]. CBIR utilizes features of images instead of meta-data for addressing, indexing and retrieving the data. This method of filtering images gives well enhanced indexing and query result [3]. The CBIR system is the appropriate methodology for representing any images in multimedia system. This method is used for addressing the elements of image to establish a CBIR system based on fundamental structures. In practical, the content based image retrieval systems (CBIR) depends on the extracted feature set from images [4]. Visual concept detection (VCD) is the fundamental task in content based multimedia information retrieval (CBMIR), in which vocabularies is used in the images that is characterized by set of concepts which fuse scenes,objects and definitelabelled substances[5]. Large volume of multimedia information increases on account of video sharing websites, some electronic telecom, video recording locations and news affiliation. This approach incorporates close-by descriptors count, vector quantization through clustering [6], composed of scene or object representation by methods via localized vector code histograms, closeness measure to construct kernel and the classifier learning. Similarity search concepts is considered asmost considerable complex task in multimedia retrieval systems and has discovered broad utilization in both the business and logical applications in various situations, duplicate and near-duplicate detection of images and videos particularly with the end goal of content-based image retrieval (CBIR) [7]. For fulfilling the framework prerequisites and furthermore the client needs as for versatile media retrieval, improvement of content-based similarity models have been completed. One of the basic difficulties is that the inherent properties of the information objects are gathered by methods of feature representations that are used for contrasting the individual information objects in view of their segments. In this survey, analysis about the visual concept detection strategies [8] and the semantic based interactive media content retrieval procedures [9] has been examined in detail. In this survey, typical research strategies are examined with their working procedure. Here the methods are analyzed as far as their benefits and adverseresults and it is contrasted with other methods in terms of variation and furthermore degradation than other research strategies. The overall organization of the survey is given as follows: In the above section, discussion about the introduction of content based video/image retrieval and their necessity is discussed. In section 2, discussion about the different techniques that are used to perform the content based video retrieval and the annotation has been provided. In section 3, comparison evaluation of the video retrieval techniques in terms of their merits and demerits has been

Revised Manuscript Received on April 07, 2019.

Tamil Priya D received the M.E degree in Computer Science and Engineering from Anna University, India. She is Assistant Professor in the Department of Information Technology, VIT University, India.

Divya Udayan J received the PhD degree in Internet and Multimedia Engineering from Konkuk University, South Korea. She is Associate Professor in the Department of Information Technology, VIT University, India.

given. In the section 4, numerical evaluation of various proposed techniques based on accurate retrieval outcome is provided for the future research work. Finally, in the section 5, overall comprehensive conclusionabout the research work is provided based on numerical outcome obtained.

II. SEMANTIC BASED MEDIA RETRIEVAL SYSTEM

The main goal of video retrieval process is to retrieve the specific video shots according to client needs. Video retrieval is an effective working process as a result of the huge inventive change that contributes to direct video catching and sharing. Insignificantly, it is more complex to measure the efficiency of the videos found on web or on the storage media for video retrieval. Indexing of low level feature videos cannot be able to see the most majority of client's requirement and needs. Majority of client's queries can be either investigated by illustrations or by a content analysis. It is necessary to overcome the needs of the clients based on the queries articulated by the client. In order to reduce the semantic gap between client perceptions and the prerequisites, features of low level video can be considered. Likewise, new semantic approaches have been developed in video retrieval to link the client needs usingthe provided low level features in the video. These types of semantic approaches exist, inconcern of eliminating the high level features from video. Semantic concepts are concentrated by various research methods to address the issues found in content based retrieval. These techniques need to satisfy the client queries and their perspectives. In this way, there is an unquestionable requirement to develop extensive semantic concept discovery. Yet, to make non-specific huge scale concept detectors, there are a few issues, for example,

1. There are an unlimited eminence amount of high level concepts which exist in client perceptions, and there is no any unpretentious method to create concept detector for incredible amount of high level concepts.

2. To build up the concept detector, an excessive approach is required which would consume enormous time, and requires various intends to take after, for instance: (i) Data pre-processing, (ii) extraction of low level features, (iii)classification of machine learning techniques.

In the remaining subsection, description about the varying techniques whose principle objective is to perform the image/video annotation and retrieval of videos/images based on semantic knowledge is given.

A. Image/Video Annotation Techniques

In this section the various image/video annotation techniques has been discussed. Tian and Shen [10] discussed the issues in using noisy datasets that hinder the annotation process. Additionally, the author introduced the approach called semantic neighborhood learning model on noisy data, that re-established the missing labels, and semantic adjusted neighborhood is created. This method allows the coordination of various labelled data based learning and neighborhood nonnegative deficient coding. It assembled semantic meaning of reliable neighborhood for every case, thus relating neighbors have higher global closeness, fractional correlation, and sensible resemblance with nearby semantic change. Meanwhile, an iterative denoising methodology is moreover proposed.Jing et al [11] proposed Multi-label learning approach, viz. Multi-Label Dictionary Learning (MLDL), by constancy regularization and fractional label insertion in indefinite manner, and at the same time it conducts multi- label lexicon learning and halfway vague label installation. It associates the dictionary learning techniques as multi-label learning in the information feature space, and sketches the regularization of label consistency period to well-study the description of features. In the output label space, it outlines the halfway vague label embedding, that tests with the exactly same set of label that can group together, and tests with fractional vague label sets can cooperatively communicate with each other. Yao et al [12] proposed a frame work, which has synthetic aperture radar (SAR) image description for a semi-automated hierarchical clustering and classification context. This execution performance of the system permits the order and description of image information extracted from the immense satellite information documents. This system encompasses of three phases: In the first phase, each one of the image is cut into numberof patches and each and every patch is changed into a vector consisting texture feature,, in second phase, the comparative feature vectors are grouped together, whererepeated clustering is used to quantify the cluster group in order to advance their gaussianity and in the third phase, for each and every image patch, the fitting classes is assigned and this is accomplishedfinally by semi-supervised learning. Ding et al [13] proposed a novel approach to synchronize the contexts of featureand labelinto a common structure called Context-aware MIML (CMIML) model. Especially, the context of feature isestablished by various graphs, whereas the label context is established over a straight blend the conceptions of low level features and high level semantic labels.Feng et al [14] suggested a deep architecture designed for customizing image annotation by using the affluence of information in client's labeling history. The proposed framework includes three sections: two sections is used for learning features of the image substance and client's history names and the other one for joining the two insightful features to foresee the labels. It also researched two distinctive approaches to show client's history names:

1) simply average the embedding client's history names and 2) demonstrate client's history labels with a gathering model using deep neural network framework. Analysis of this deep architecture on a generous scale and sensible informational index, containing ~22.8 million open images exchanged by ~4.69 million clients. Zhang et al [15] proposed three intellectual annotation structures: First is **multi-data extraction:** to improve the multi-helper data, the label co-event, and client premium vector are added; other than different visual highlights; Second, **introductory labelling**: in sight of the conventional term recurrence—reverse record recurrence model is proposed and third, is the **label refinement**: it proposed the multi-data, all labels display for label refinement by considering data from multiple multimedia contents including multi-visual substance, label co-event, and client conspiracy comparability. The label refinement process is dignified as a streamlining issue by fluctuating confidence score set by the primary naming model. Ramisa et al [16] have untaken an adaptive CNN model; those proposals have the common structure for different errands that contains: (1) source identification, (2) article outline and (3) article geo-location. Deep Canonical Correlation Analysis is done using object representation, and Great Circle Distance is proposed using geo-location. Moreover, it grants the Breaking News, that contains dataset of 100K news articles including images, content and subtitles.Hu et al [17] proposed an innovative concept, known as robust multi-view semi-supervised learning (RMSL) which ensures the performance of image annotation assignment. Also, the model accomplishes to reveal the essential information using both the labeled images and unlabeled images. The work utilizes the correlated and complemented data which obtained various features of image information, systematically to represent an individual data. In order to accomplish the annotation consistency, the method ensures the robust pairwise requirement by consequences of various perceptions. Moreover, it incorporated a robust classifier learning segment which can give feasible noise distinguishing proof influence between the learning processes. Li et al. [18] proposed a technique to implement automatic image annotation, so-called a ranking-preserving low-rank factorization (rPLrf) technique. The training set for each test image is ranked and this leads to low-rank matrix factorization on the model coefficient lattice. Likewise, the prediction model is refined through label positioning standardized by test similarities and label relationships to reduce the equivocalness presented by missing labels and combined into the factorization scheme This technique anticipates the requirement for relaxing paired choices on unreliable data and it is more accurate to find missing labels, beside all the earlier stated methods together.

B. Feature Extraction Methods

Dimensionality reduction is a concept used in the field of machine learning and pattern recognition based applications. Broadly, in this survey, certain feature selection and feature extraction techniques has been analyzed with the purpose of how these techniques can be used effectively for learning algorithms to achieve high performance that eventually increases predictive accuracy of classifier.

In order to encrypt the local topological structure of face images, the author Penev et al. [30] proposed the techniques known as Local Feature Analysis (LFA), to detect the local features of the facialparts; LFA is considered as the most appropriate method using a set of kernels. Timofte et al. [3] proposed a Local Binary Pattern (LBP), to investigate the area of texture analysis forthe facial representation. By using this technique, apply the LBP operator, and thensubsequently resultant LBP imageis divided into small regions, where histogram features are extracted. The component-based method is used to divide the face image, and then the face image is divided into smaller blocks. Now, these image blocks are used as inputs of classifiers or for supplementary feature extraction process (e.g., PCA, FLD). Fan et al. [31] deliberated a novel face recognition approach which combinesseveraldiscriminating Gabor features model in multi-scales and multi-orientations. In order to build a collection of feature vectors, a series of appropriate Gabor filters is applied for specified Normalized facial images. To improve the face recognition techniques more effectively, the different channels are tuned by combining the different texture information and different consistent representations. Lee et al. [32] deliberate a local and global feature extraction for face recognition. This technique depends on Local Feature Analysis (LFA) and it comprises of three phases, for example, construction, selection, and combination of local structures. Subsequent to building kernels utilizing LFA, computed their fisher scores where the score esteems are shown on the area of the comparing kernels. It demonstrates that kernels having a place with the important territories for recognition, for example, eyebrow, nose, cheekbone, and jaw-line, got higher scores than the rest. This confirms the convenience of Fisher Score for kernel determination. The framework builds a subset of kernels, which is productive for recognition and it consolidates the nearby structures to address them in a smaller shape. Jiang et al. [33] proposed another image feature extraction and recognition technique 2D Linear Discriminant Analysis (2DLDA). LDA's compression rate is better than Image Matrix based LDA (IMLDA). 2DLDA outperforms IMLDA twice both in horizontal level and vertical level for the transpose operation matrix. It is more productive in recognition rate and low computational cost which is better than IMLDA.
In spite of the fact that 2DLDA have low

computational cost when contrasted with IMLDA which appears complex due to performing segregation more than once. In the work of Zuo et al. [34],feature extraction, is done by combiningbidirectional PCA (BDPCA) and LDA (BDPCA + LDA), that implements and out performs the LDA in BDPCA subspace. To reduce the image dimension, two-dimensional PCA is used and it is applied to original image for covariance matrix and its transposed version is applied to generate a correct set of feature representation. Both BDPCA and LDA technique are used for the extraction of facial features that connects on a low-dimensional BDPCA subspace. However, computational cost is an issueof these techniques. Song et al. [35] composed of another feature extraction technique named as Parameterized Direct LDA (PD-LDA) for small size issues. The main advantages of PD-LDA are: that it can be directly applied to high-dimensional input spaces, and implemented with prodigious efficiency. Subsequently, working procedure of PD-LDA is (i) to perform eigen value decomposition on scatter matrix to obtain discriminant matrix, (ii) to map each sample vector to obtain its intermediate representation, (iii) to perform eigen value decomposition on the within-class matrix of projected samples, (iv) to calculate regulating matrix and discriminant matrix then mapping each sample. Sahoolizadeh et al. [36] used a combination model of Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA) and Neural Networks. The context of the model is divided into four units, like pre-processing, dimensionality reduction, feature extraction and classification. Pre-processing removes the background information which is not necessary for recognition, to reduce the computational complexity most significant Eigen-vector and eigen-faces are computed using PCA. In LDA eigen-vectors with greater Eigen values are selected to increase the discriminating power and finally three layersMLP neural network trains and classify the database using simple back propagation algorithm.

Kinage et al. [37] proposed 2D Principal Component Analysis (2DPCA) on Wavelet sub band for dynamic face recognition. By utilizing different wavelet changes like Haar, Daubechies, Coiflet and so on., image features of facial images are separated and with sub-band disintegration running from 1 to 8. 2D-DWT is performed using progressive low pass and high pass filtering of the image and image is divided into four sub groups LL, LH, HL and HH. 2DPCA develops an image covariance matrix specifically utilizing the original image matrices and its Eigen vectors are inferred for feature extraction of image. Higher accuracy is accomplished for face recognition; however irregular wavelets cannot control2DPCA. The author Yang et al. [38] proposed a framework for face recognition, the local pattern descriptor, called Local Derivative Pattern (LDP), which is used to encode the directional pattern features based on local derivative variations. However, in order to encode the $(n-1)^{th}$ order local derivative direction variations, the n^{th} order LDP is used to capture more detailed information than the first-order local pattern used in LBP. The proposed model provides high-order local information by encoding various distinctive spatial interactions limited to given local regions, which has different relationships among central point and its neighbors.

Yang et al. [39] proposed a computing algorithm to handle occluded face images, a Gabor-feature based SRC (GSRC) system with Gabor occlusion dictionary. To create the occlusion dictionary compressible, image Gabor-features of image and of Gabor kernels are used for SRC. To extract the local directional features of image, a Gabor filters are used broadly at multiple scales. When the proposed system uses Gabor occlusion dictionary computing algorithm, it reduces the computational cost and increases SRC accuracy. Gabor features might not to be a best approach for training data, which ensures that it is not encompassed of particular face information learned from face training data. Tang et al. [40] aimed at developing a novel method for face feature extraction techniques recognition method using contourlet transform and CSA, which has properties like, energy aggregation, multi-resolution and directional image expansion. To infer the dimension, Coupled Subspace Analysis (CSA) employs optimal bi-directional projection based matrix that allows it to attain enhanced result for recognition and to inferior computational complexity. In initial stage, each face is decomposed in same scale both by contourlet transform and frequency coefficients and numerous directions are merged into a sub band. At last, by improved CSA, face discriminant features are extracted in merged sub bands.

Zhang et al. [41] developed a Monogenic Binary Coding (MBC) effective local feature extraction approach for face recognition. The Monogenic gesture illustration crumbles an original gesture into 3 different components namely, amplitude, alignment, and segment. Respectively, the monogenic variation of local region and monogenic feature of pixel is encrypted by using this approach, and then extracts the local features with computed statistical(features) values (e.g. histogram) which merges the monogenic component for effective facial recognition. The experimental result shows that the proposed Monogenic Binary Coding (MBC) system has lower time and space complexity when compared to Gabor-transformation based local feature methods.Yu et al. [42], for facial expression recognition, the author combined both the descriptors and geometric features of the images. In order to characterize the facial expression, the covariance descriptor is used which composes of

different textural feature. Then the geometric features are captured from the facial expressions. Finally, the combined expression and geometric features are form a vector representation for the facial expression recognition.

Ren et al. [43] aimed to develop the enhanced Local Gradient Order (LGO) features for face recognition.Hence, by considering the local features explicitly, structure of facial images is determined with the neighboring pixel points. Spontaneously, the most of the discriminant subspace are merged by the multimodal features. Adaptive interactionfunction is used in order to overcomethe outliers of each dimension of the pixel for the similarity measurement and discriminant analysis robustly. Two kinds of similarity functions are used such as kernel and entropy, in order measure the closeness of each pair of features.

C. Semantic Based Video/Image Retrieval Techniques

In this section, various research methods have been introduced for the purpose of retrieving the video contents based on semantic similarity that has been discussed.

Yang and Meinel [19] proposed a method for searching a video in a large number of video files, the method called automated video indexing. In order to suggest visual recommendations for the video content navigation, automatic video segmentation and key-frame detection are proposed. By applying Video Optical Character Recognition (OCR) technology and Automatic Speech Recognition (ASR), techniques, it extracts textual metadatafrom audio tracks. Also, OCR and ASR transcriptare used to detectslide text line types using video and segment-level keywords which were adopted for keyword extraction for content-based video browsing and searching.

Semi supervised kernel hyper-plane learning (SKHL)approach for semantic image retrieval is proposed by Kan et al [20] for hashing.Each hashing function is represented as a non-linear kernel hyper-plane which obtained from an unlabeled dataset. Furthermore, in order to study the optimal kernel hyper-plane and hashing functions,a Fisher-like benchmark model is anticipated byusing the weakly labeled training samples with adjacent data and further to integrate different types of features.Additionally,it also integrates multiple kernel learning (MKL) approach as SKHL approachwhich is proposed to promote the different kinds of integration, that leads to better hashing functions. Zhu and Shyu [21] introduced a model that emphasises on integrating the visual content and with its related meta-data called sparse linear integration (SLI) model, with consistently as content and context modalities for semantic concept retrieval. To estimate a feature set by means of asparse linear combination of other features and to reduce the difference between them, an optimization problem is framed. For features, predictionscore value measures of concept can be remodelled by the positive features of that concept.

Guo et al [22] proposed an effective architecture termed as Semantic-based Heterogeneous Multimedia Retrieval (SHMR), which retrieves semantic information from heterogeneous multimedia data and it is low cost to store the data. The proposed framework addresses the following activities for heterogeneous multimedia retrieval as follows: (i) address the accuracy of heterogeneous multimedia retrieval in large data environments, (ii) proposed an approach to extract and represent semantic information for heterogeneous multimedia documents, (iii) an approach called NoSQL, which is provided to process the multimedia data parallel in distributed nodes to semantic storage.(iv) a user opinion system is introduced to achieve high accuracy precision and to design noble user experiences and finally MapReduce-based retrieval is proposed. Fernandez-Beltran and Pla [23] to survive with incremental circumstances in definite and efficient way, a new idea is proposed, incremental topic model (IpLSA), which provides different retrieval systems and two reference benchmarking video databases.

In order to incorporate the semantic and visual features,Allani et al [24] proposed an image retrieval system. The knowledge behind the system is to spontaneously construct anintegrated ontology for the semantic information and to establish the visual features in apopular graph-based model. For the retrieval purpose, the two components are associated together as "pattern" in the image retrieval system.For videos, to produce the annotation as key frames based on ontology andin order to allow the users to access video data from the large volume of database, techniquecalled the video annotation technique is proposed byTulasi et al [25],which is used for retrieving the videos from largedatabase. However, the videos are scrutinized more rapidly and efficiently by video annotation system ontology. During analysis phase, key frame ingests a smaller amount of time for entire video analysis when compared with others. Key frame extraction process which increases the accuracy using entropy besides the prewitt operator and by the nearest- neighbor methods.

To overcome the semantic gap between low and high level feature of video, Fernandez-Beltran and Pla [26] proposed an innovative approach called Content-Based Video Retrieval (CBVD) approach. First, the proposed supervised learning model which enhances the traditional retrieval approach as a class discovery problem.Consequently, a novel probabilistic ranking function is recognized, to experiments the challenge ofsemantic gap among the low-level and high-level feature.Finally, sample

queries are initiated from the database and ashort-range relevance feedback is strengthened.

Supervised distributed hashing (SupDisH) method is proposed by Zhai et al [27]to studydiscriminative hash functions by the influence of semantic label information in a distributed manner. Toperform the issues of hashing function as the context of classification in distributed manner, the learned binary codes areexpectedfor semantic retrieval. Distributed model is separated asdecentralized sub-problem set by introducing supplementary variables with consistent limitations, which could be resolved in parallel manner by every vertex of distributed network. For web images and fast Fisher Vector products (VRFP), Han et al [28] developedthe video retrieval techniques based on the input queries as text input, once the queries is known itaccomplishes weakly labeled trainingimages from the web. The web images that are retrieved from the database is the representative of query and each video are treated as unordered image groups in the database, and each groups is represented as asingle fisher vector which built on Convolution Neural Network features. To speed up the inner product computation among the high-dimensional fisher vectors, lossless algorithm is proposed and the outcome shows that it decreases quadratically with the sparsity of Fisher Vectors. Jyothi et al (2018) [29] suggested an approach for video retrieval system with flowers dataset for training and learning features using Deep learning technique. By using this technique, network is trained in three ways; one by key frames, other by segmented flowers and another by flowers gradient asthe input. By using Multiclass Support Vector Machine (MSVM), the related videos are retrieved by the system based on the query given by the videos from the large database. The experimental result shows that the complexity could be reduced to a higher range by training the DCNN using the flowers gradient without conceding the performance of the system.

III. COMPARISION EVALUATION ON SEMANTIC BASED VIDEO RETRIEVAL TECHNIQUES

In this section, comparison evaluation of the different research techniques based on merits and demerits has been given. From this section, we can find the issues and also the benefits of each and every technique, so that novel technique can be introduced which can resolve the issues found in every technique. In the following sub section, comparison evaluation of these techniques in terms of their working functionality issues and performance benefits is provided for both techniques namely multimedia content annotation and multimedia content retrieval techniques.

Comparison of Image/video annotation techniques for various methods based on merits and demerits is shown in Table I. In Table II, various approaches of feature extraction techniques with its merits and demerits are given. Comparison of Semantic based Image/Video retrieval techniques based on merits and demerits is shown in the Table III.

A. Discussion of Image/Video Annotation Techniques

Table I
Image/Video annotation techniques comparison

S.No.	Author, Year	Method	Merits	Demerits
1	Tian and Shen (2015) [10]	Semantic neighborhood learning model(SNLM)	It is more effective in small scale databases Reduced time complexity	It might reduce in its performed with large scale data presence
2	Jing et al (2016) [11]	MLDL calledMulti-Label Dictionary Learning by using partial identical labelsand label consistency regularization.	Consumes less training time MLDL can annotate on images with more right labels, as well as can clarify diverse images utilizing a similar label with moderately appropriate weights	It requires full labeling information to perform accurate annotation
3	Yao et al (2016) [12]	Semi-automated hierarchical clustering and classification framework (SAHCCF)	Improved classification performance with the consideration of varying distance metrics	It required human interaction to obtain the increased classification accuracy rate
4	Ding et al (2016) [13]	Context-aware multi-feature multi-label learning (CMIML)	Increased precision rate Obtains satisfied performance with increased	Semantic accuracy is not guaranteed

		model	number of labels	
5	Feng et al (2017) [14]	Deep architecture (DA)	It is compelling on customized image annotation assignment with help of client's history labels data and the model consolidate LSTM TagNet and Image ConvNet perform best among a few models	It requires more training information to retrieve the accurate annotation outcome
6	Zhang et al (2017) [15]	AIAF-MAImethod (Automatic image annotation framework based on multi-auxiliary information)	higher exactness and scope than conventional techniques More power than numerous other well-known refinement techniques	Present of more irrelevant features would reduce the annotation outcome accuracy
7	Ramisa et al (2016) [16]	Adaptive CNN architecture (ACNN)	Increased performance ratio with the consideration of geo location attributes	More sensitive to inaccurate text and images data.
8	Hu et al (2017) [17]	Robust Multi-view Semi supervised Learning (RMSL)	It ensures the promising annotation outcome	Experiences the weight of more computational assets and over-fitting issue particularly in the event of little training set.
9	Li X et al (2018) [18]	Ranking-preserving low-rank factorization method (RPLRFM)	It achieves significant improvement over traditional methods	It still needs improvement in terms of semantic similarity which can be concentrated in future

B. Discussion about Feature Extraction Techniques

TableII
Various approaches in feature extraction methods

S.No.	Reference	Method	Merits	Demerits	Results
1	Penev et al. [30]	Local Feature Analysis (LFA)	Better error stability	Time complexity	It statistically derived local features and their position.
2	Fan et al. [31]	Null Space-based LDA (NLDA)	High retrieval rate Improved dimensionality reduction	Its normal sub examining plan may lose critical discriminant data in the face locale	It increases the final retrieval rate
3	Lee et al. [32]	Local Feature Analysis (LFA)	In order to improve the retrieval rate extract local structure	Time complexity	Better retrieval performance
4	Jiang et al. [33]	2DLDA	Better Compression rate	High Complexity	Efficient retrieval rate and low computational cost
5	Zuo et al. [34]	Bidirectional PCA (BDPCA) plus LDA (BDPCA + LDA)	Higher retrieval accuracy	Need to improve performance	Less computational and memory requirements
6	Song et al. [35]	PD-LDA	More effective and robust	Fine tuning of parameter has not been take place	High face recognition rate

7	Sahoolizadeh et al. [36]	PCA, LDA and Neural Networks	High Recognition rate, reduced computational complexity	New phase algorithms for security purposes	Improvement in feature extraction and selection
8	Kinage et al. [37]	2DPCA on Wavelet sub-band	Improved accuracy	Not suitable for handling curve discontinuities	7th order symlet gives highest accuracy
9	Yang et al. [38]	GSRC, Gabor occlusion computing algorithm	Compressed occlusion dictionary	Does not contain face specific information	Low computational cost and improved SRC accuracy
10	Tang et al. [39]	Contourlet transform	Low computational complexity	It does set the weights adaptively	It is efficient and can get solid right recognition rate.
11	Zhang et al. [40]	Local Binary Pattern (LBP)	It achieves Robust similarity measurement	Required to check domain adaptation ability	High face recognition rate
12	Yu et al. [41]	Appearance descriptors and geometric features	Great execution in the recognition of six essential outward appearances	Poor performance while dynamic changes of facial expression	Better recognition rate
13	Lei Z et al. [42]	Enhanced-Local Gradient Order so-called IGO features	It achieves Robust similarity measurement	Required to check domain adaptation ability	High face recognition rate

C. Discussion about Semantic Based Video Retrieval Techniques

TableIII
Semantic based Image/Video retrieval techniques comparison

S.No	Author, Year	Method	Merits	Demerits
1	Yang and Meinel (2014) [19]	Optical Character Recognition (OCR) technology Automatic Speech Recognition (ASR)	Accurate retrieval of videos is guaranteed More pertinent videos can be retrieved by considering both sound and video highlights	More overhead by considering both sound and video features for the retrieval of comparative videos
2	Kan et al (2014) [20]	Semi supervised kernel hyper plane learning (SKHL)	Accurate retrieval of videos on small scale video set Increased performance by matching more similar videos	In this work, each hashing function is refreshed autonomously in every emphasis which may not be reasonable for the expansive scale video set
3	Zhu and Shyu (2015) [21]	Sparse linear integration (SLI) model	min combination strategy demonstrates genuinely great outcomes and is superior to the normal and max combination strategies SLI indicates promising outcomes by outflanking the early combination and late combination approaches in view of logistic regression	Retrieving positive occasions from exceedingly imbalanced dataset is an exceptionally difficult issue

4	Guo et al (2015) [22]	Semantic-based Heterogeneous Multimedia Retrieval (SHMR)	Capacity and I/O cost in the design is altogether decreased	It cannot perform well in case of arrival of large volume of data entering into the system
5	Fernandez-Beltran and Pla (2015) [23]	Incremental topic model (IpLSA)	Efficient handling of incremental database Support scalability with increased accuracy of retrieval	Over fitting problem Cannot support multi model data
6	Allani et al (2016) [24]	Knowledge-based Image Retrieval System	Fundamentally increment the significance of retrieval comes about, by upgrading the ranking of images	Still it might reduce in its accuracy with image database of missing information
7	Tulasi et al (2016) [25]	Video annotation technique	Increased accuracy rate Faster computation time	Required more training samples to ensure the accuracy Video with missing labeling information would reduce the classification accuracy
8	Fernandez-Beltran and Pla (2016) [26]	Content-Based Video Retrieval approach (CBVR)	LTR positioning function can outflank whatever is left of the tried capacities predominantly in light of the fact that, approach has the similar probabilistic nature than the point models approach	More topic extraction cost More patterns diversity
9	Zhai et al (2017) [27]	Supervised distributed hashing (SupDisH)	SupDisH calculation works intensely against the best in class techniques in both scale and precision	It cannot support scalability Reduced in performance with increased arrival of videos
10	Han et al (2017) [28]	Video retrieval and fast Fisher Vector products (VRFP), using web images	It does not require discriminative training and ensures the accurate video retrieval In web images robust noise present.	Increased time overhead
11	Jyothi et al (2018) [29]	FVRS-DLA(Flower video retrieval system using Deep learning approach)	Reduced complexity Increased accuracy	It required more training sample to ensure the guaranteed accuracy

IV. NUMERICAL EVALUATION

In this section, comparison evaluation of the various research techniques which are used to perform the multimedia content annotation and retrieval techniques has been given. This comparison evaluation is made under matlab simulation environment based on which improvement of different research method in various aspects has been given. The comparison evaluation is done in terms of metrics namely accuracy, precision and recall. In the following subsection discussion about the varying research methods has been given.

A. Comparison of Image/video annotation techniques

In this section, comparison evaluation of the different image/video annotation techniques has been given. Here the comparison is made between the four research methods from different years to ensure the performance improvement of each technique over the year. The methodologies that are considered asthe comparison for Semantic neighborhood learning model (SNLM), Context-aware multi-feature multi-label learning (CMIML) model, Automatic image annotation framework based on multi-auxiliary information (AIAF-MAI), and Ranking-preservinglow-rank factorization method (RPLRFM).The comparison is made in terms of accurate retrieval of videos based on how well videos are annotated. The metrics that are considered for the comparison evaluation are Accuracy, precision and recall.

The evaluation of the methods SNLM, CMIML, AIAF-MAI, RPLRFM based on these performance metrics is done and it is shown in table IV. The numerical evaluation of the research methods is conducted by comparing it with each other

technique which is shown in Fig.1.

Methods / Performance Metrics	SNLM	CMIML	AIAF-MAI	RPLRFM
Accuracy	86	89	91	96
Precision	89	91	97	99
Recall	89	93	97.5	99.3

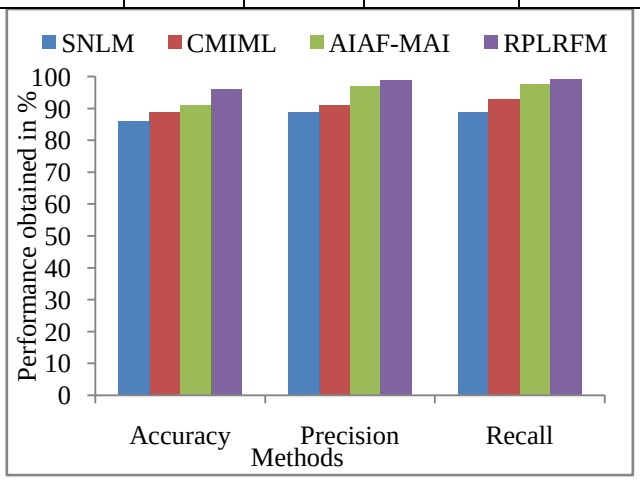

Fig.1. Numerical comparison outcome

From the Fig.1, it can be concluded that RPLRFM framework leads to provide improved performance than the existing research method by accurately retrieving the similar videos from the training database.

B.Comparison Evaluation of video/image Retrieval System

In this section, comparison evaluation of the different image/video retrieval techniques has been given. Here the comparison is made between the five research methods from different years to ensure the performance improvement of each technique over the years. The methodologies that are considered for the comparison evaluation is, Semi supervised kernel hyper plane learning (SKHL), Semantic-based Heterogeneous Multimedia Retrieval (SHMR), Content-Based Video Retrieval approach (CBVR), Video retrieval using Web images, fast Fisher Vector products (VRFP), and flower video retrieval system using Deep learning approach (FVRS-DLA). The comparison is made in terms of accurate retrieval of videos based on the accuracy of videos annotated.The metrics that are considered for the comparison evaluation are Accuracy, precision, recall and F-Measure.

The performance comparison measure for various methods is shown in the table V.The evaluation of the methods SKHL, SHMR, CBVR, VRFP, FVRS-DLA based on these performance metrics is done. The numerical evaluation of the research methods is conducted by comparing it with each other technique which is shown in the following Fig.2.

Table V
Performance Comparison Measures

Methods / Performance Metrics	SKHL	SHMR	CBVR	VRFP	FVRS-DLA
Accuracy	88	91	93.5	94.6	98.3
Precision	91	92.7	94.6	97	98.9
Recall	92.3	94.7	96.1	98.8	99.4
F-Measure	91.65	93.69	95.34	97.89	99.15

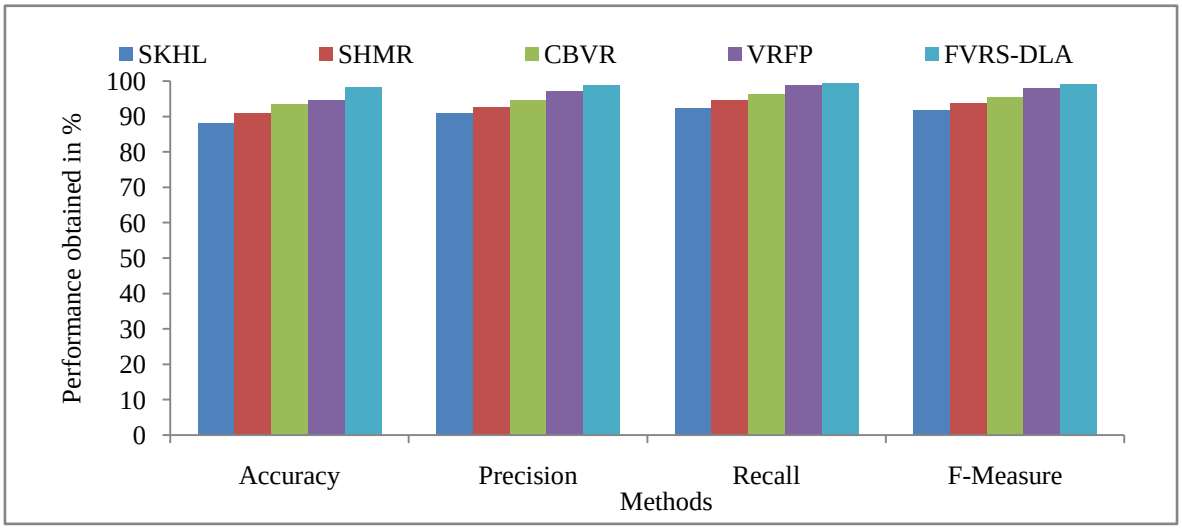

Fig.2. Numerical comparison outcome

From the Fig.2, it can be concluded that FVRS-DLA framework leads to provide improved performance than the existing research method by accurately retrieving similar videos from the training database.

V. CONCLUSION

In this survey, discussion about the various techniquesused to retrieve the multimedia content from the web engines with the semantic knowledge consciousness is provided. The various research methods that intend to perform the annotation on multimedia contents is discussed in detailed. And also, this survey provides the discussion about the techniques in terms of their merits and demerits. Finally this survey also provides the numerical evaluation of each technique based on their accuracy in the video/image retrieval outcome. Finally this survey also provides the numerical evaluation of each technique based on their accuracy in the video/image retrieval outcome. The overall evaluation of the research method is performed and discussion about the outcome obtained from each technique is provided in detail.

Acknowledgement

The authors thank VIT University for providing "VIT SEED GRANT" for carrying out this research work.

REFERENCES

1. Qazi, A., & Goudar, R. H. (2016). Emerging trends in reducing semantic gap towards multimedia access: A comprehensive survey. Indian Journal of Science and Technology, 9(30).
2. Liu, G. H., Yang, J. Y., & Li, Z. (2015). Content-based image retrieval using computational visual attention model. pattern recognition, 48(8), 2554-2566.
3. Xia, Z., Wang, X., Zhang, L., Qin, Z., Sun, X., & Ren, K. (2016). A privacy-preserving and copy-deterrence content-based image retrieval scheme in cloud computing. IEEE Transactions on Information Forensics and Security, 11(11), 2594-2608.
4. Ahmad, J., Sajjad, M., Rho, S., & Baik, S. W. (2016). Multi-scale local structure patterns histogram for describing visual contents in social image retrieval systems. Multimedia Tools and Applications, 75(20), 12669-12692.
5. Mei, T., Hua, X. S., Li, S., &Gu, Z. (2014). U.S. Patent No. 8,804,005. Washington, DC: U.S. Patent and Trademark Office.
6. Kolesnikov, A., Trichina, E., & Kauranne, T. (2015). Estimating the number of clusters in a numerical data set via quantization error modeling. Pattern Recognition, 48(3), 941-952.
7. Zhang, X., Liu, W., Dundar, M., Badve, S., & Zhang, S. (2015). Towards large-scale histopathological image analysis: Hashing-based image retrieval. IEEE Transactions on Medical Imaging, 34(2), 496-506.
8. Koniusz, P., Yan, F., Gosselin, P. H., & Mikolajczyk, K. (2017). Higher-order occurrence pooling for bags-of-words: Visual concept detection. IEEE transactions on pattern analysis and machine intelligence, 39(2), 313-326.
9. Alzu'bi, A., Amira, A., & Ramzan, N. (2015). Semantic content-based image retrieval: A comprehensive study. Journal of Visual Communication and Image Representation, 32, 20-54.
10. Tian, F., & Shen, X. (2015). Learning semantic concepts from noisy media collection for automatic image annotation. Chinese Journal of Electronics, 24(4), 790-794.
11. Jing, X. Y., Wu, F., Li, Z., Hu, R., & Zhang, D. (2016). Multi-label dictionary learning for image annotation. IEEE Transactions on Image Processing, 25(6), 2712-2725.
12. Yao, W., Dumitru, C. O., Loffeld, O., & Datcu, M. (2016). Semi-supervised hierarchical clustering for semantic SAR image annotation. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 9(5), 1993-2008.
13. Ding, X., Li, B., Xiong, W., Guo, W., Hu, W., & Wang, B. (2016). Multi-Instance Multi-Label Learning Combining Hierarchical Context and its Application to Image Annotation. IEEE Trans. Multimedia, 18(8), 1616-1627.
14. Feng, F., Liu, R., Wang, X., Li, X., & Bi, S. (2017). Personalized Image Annotation Using Deep Architecture. IEEE Access, 5, 23078-23085.
15. Zhang, P., Wei, Z., Li, Y., & Zhao, C. (2017). Automatic Image Annotation Based on Multi-Auxiliary Information. IEEE Access, 5, 18402-18411.
16. Ramisa, A., Yan, F., Moreno-Noguer, F., & Mikolajczyk, K. (2018). Breakingnews: Article annotation by image and text processing. IEEE transactions on pattern analysis and machine intelligence, 40(5), 1072-1085.
17. Hu, M., Yang, Y., Shen, F., Zhang, L., Shen, H. T., & Li, X. (2017). Robust web image annotation via exploring multi-facet and structural knowledge. IEEE Transactions on Image Processing, 26(10), 4871-4884.
18. Li, X., Shen, B., Liu, B. D., & Zhang, Y. J. (2018). Ranking-Preserving Low-Rank Factorization for Image

Annotation with Missing Labels. IEEE Transactions on Multimedia, 20(5), 1169-1178.
19. Yang, H., & Meinel, C. (2014). Content based lecture video retrieval using speech and video text information. IEEE Transactions on Learning Technologies, (2), 142-154.
20. Kan, M., Xu, D., Shan, S., & Chen, X. (2014). Semisupervised hashing via kernel hyperplane learning for scalable image search. IEEE Transactions on Circuits and Systems for Video Technology, 24(4), 704-713.
21. Zhu, Q., & Shyu, M. L. (2015). Sparse linear integration of content and context modalities for semantic concept retrieval. IEEE Transactions on Emerging Topics in Computing, 3(2), 152-160.
22. Guo, K., Pan, W., Lu, M., Zhou, X., & Ma, J. (2015). An effective and economical architecture for semantic-based heterogeneous multimedia big data retrieval. *Journal of Systems and Software*, *102*, 207-216.
23. Fernandez-Beltran, R., & Pla, F. (2015). Incremental probabilistic Latent Semantic Analysis for video retrieval. *Image and Vision Computing*, *38*, 1-12.
24. Allani, O., Zghal, H. B., Mellouli, N., & Akdag, H. (2016). A knowledge-based image retrieval system integrating semantic and visual features. *Procedia Computer Science*, *96*, 1428-1436.
25. Tulasi, R. L., Rao, M. S., Usha, K., & Goudar, R. H. (2016). Ontology-based annotation for semantic multimedia retrieval. *Procedia Computer Science*, *92*, 148-154.
26. Fernandez-Beltran, R., & Pla, F. (2016). Latent topics-based relevance feedback for video retrieval. *Pattern Recognition*, *51*, 72-84.
27. Zhai, D., Liu, X., Ji, X., Zhao, D., Satoh, S. I., &Gao, W. (2018). Supervised Distributed Hashing for Large-Scale.
28. Han, X., Singh, B., Morariu, V. I., & Davis, L. S. (2017). VRFP: On-the-fly video retrieval using web images and fast fisher vector products. *IEEE Transactions on Multimedia*, *19*(7), 1583-1595.
29. Jyothi, V. K., Guru, D. S., Sharath Kumar, Y. H. (2018). Deep Learning for Retrieval of Natural Flower Videos. *Procedia Computer Scienec*, 132, 1533-1542.
30. PenevPS., and Atick J. Local feature analysis: A general statistical theory for object representation. Network: computation in neural systems, 1996; 7(3): 477-500
31. Fan W, Wang Y, Liu W, et al. combining null space-based Gabor features for face recognition. In 17th International Conference on Pattern Recognition, 2004; 1: p. 330-333.
32. Lee KC, Ho J, Kriegman DJ. Acquiring linear subspaces for face recognition under variable lighting. IEEE Transactions on pattern analysis and machine intelligence, 2005; 27(5): 684-698.
33. Jiang D, Hu Y, Yan S, et al. Efficient 3D reconstruction for face recognition. Pattern Recognition, 2005; 38(6): 787-798.
34. Zuo W, Zhang D, Yang J, et al. BDPCA plus LDA: a novel fast feature extraction technique for face recognition. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 2006; 36(4): 946-953.
35. Song F, Zhang D, Wang J, et al. A parameterized direct LDA and its application to face recognition. Neuro computing, 2007; 71(1): 191-196.
36. Sahoolizadeh AH, Heidari BZ, Dehghani CH. A new face recognition method using PCA, LDA and neural network. International Journal of Computer Science and Engineering, 2008; 2(4): 218-223.
37. Kinage KS, Bhirud SG. Face recognition based on two-dimensional PCA on wavelet sub-band. International Journal of Recent Trends in Engineering, 2009; 2(2): 51-54.
38. Yang M, Zhang, L. Gabor feature based sparse representation for face recognition with gabor occlusion dictionary. Computer Vision–ECCV 2010; 448-461
39. Tang P, Gong Q, Ni L, et al, Face feature extraction and recognition using contourlet transform and coupled subspace analysis. In 5th International Conference on Biomedical Engineering and Informatics (BMEI), 2012; p. 270-273.
40. Zhang H, Luo S, Yoshie, O. Facial expression recognition by analyzing features of conceptual regions. In 12th International Conference on Computer and Information Science (ICIS), 2013; p. 529-534.
41. Yu H, Liu H. Combining appearance and geometric features for facial expression recognition. In Sixth International Conference on Graphic and Image Processing, International Society for Optics and Photonics, 2015, p. 944308-944308.
42. Ren CX, Lei Z, Dai DQ, et al. Enhanced local gradient order features and discriminant analysis for face recognition. IEEE transactions on cybernetics.

Tamil Priya D received the M.E degree in Computer Science and Engineering from Anna University, India. She is Assistant Professor in the Department of Information Technology, VIT University, India. Her research interests include image processing, data mining, computer vision and multimedia, semantic Indexing, and Machine learning.

Divya Udayan J received the PhD degree in Internet and Multimedia Engineering from Konkuk University, South Korea. She is Associate Professor in the Department of Information Technology, VIT University, India. Her research interests include semantic modeling, image processing, recognition and classification, augmented reality and HCI.