Communications on Applied Electronics (CAE) – ISSN : 2394-4714
Foundation of Computer Science FCS, New York, USA
Volume 2 – No.1, June 2015 – www.caeaccess.org

A Study of Content based Multimedia Retrieval Systems

Joy Bhattacharjee
Research Schola
Uttar Pradesh Technical University
Lucknow, Uttar Pradesh, India

Syed Qamar Abbas, Ph.D
Director
Ambalika Institute of Technology and Management
Lucknow, Uttar Pradesh, India

ABSTRACT
With the advent of Arpanet the potential capabilities of information sharing was very well recognized and gradually the world is gifted with the ocean of Information I.e. the Internet .There is continuous inflow of information , the information may be of any kind be it texts, audio, images or it may be the audio visual information. Thus, the users are flooded with information. Due to this flooding, it is also important to see that right information may be provided to the right person at right time or 3R's, but many a times it does not happen. This paper studies various approaches by noted researchers, academician and their contributions to the multimedia retrieval systems. This paper also tries to bring out some pros and cons of the different approaches.

Keywords
Content Based Multimedia Retrieval System, Recall, Precision, Hausdorff Distance ,Multijects etc.

1. INTRODUCTION
The multimedia retrieval system has great potential and millions of people access multimedia every now and then. As per the statistics released on public domain by Google, the users of Google, YouTube and Android users are , Google have 1.2 Trillions users per year [2], You Tube has over one billion users and nearly 300 hours of Videos are uploaded every minutes[3], Android has millions of customers in 190 countries in the world[4] . The Content Based Video Retrieval is the extension of Content Based Image Retrieval Systems. The Content Based Image Retrieval Systems was first conceptually Developed by Kato[1].This paper tries to discuss the various approaches by the authors in a chronological order and tries to find out the advantages and disadvantages of various approaches . The pros and cons are analyzed in terms of time, recall and precision.

The popular search engines like Google, Yahoo Crawl the Web and listings are made and people are made to select from the listings available to them. The Multimedia searches are based on the text tagging on the video. The Query is submitted and then the Query is matched to the tags associated with each video. These results gives listings of video that may be not be relevant to the users and the users have no choice but to choose from the listed video. So the problems associated with the above tag based searches was replaced by Content based searches. This avenue of searching techniques paved the way for the On-Demand video , Online Tv Shows or popular scenes in a particular video/(s) etc .So there is potential growth of Content Based Multimedia Retrieval Systems due to specific and Varied tastes of searches in a particular video or in a different Video. This High Potentiality of Content based Video Retrieval was not untouched by researcher and or academician and they continuously thrived to give excellent researches which is being studied in this paper

2. CONTRIBUTED WORKS ON CONTENT BASED MULTIMEDIA RETRIEVAL SYSTEMS
This section discusses the contribution made by the respective researcher and academicians.

2.1 The first work which is to be discussed is by RM Bolle et.al [5]. This paper concentrates on Video analysis on Label Sequence Semantics. The authors concentrated on edited video footage rather than considering raw footage. The fact of the matter is that every video have a time line, which tells a story in continuity, but it consists of different Scenes, which rolls out in front of our eye, which gives the impression of a story .The scenes consists Shots which in turn consists of frames or it can be called as Key Frame also. This is the common Structure of any simple video. Every video analysis may or should consists of Segmentation ,in Content Based Image retrieval a segmentation is that where there is abrupt change in texture ,color etc, but in video retrieval systems the segmentation can be defined as where the scene changes. The analysis can be done on two cases (a) it can be between shot analysis ,(b) within shot analysis.

If the segment needs better visual representation then the between shot processing is done to high-level video structure that gives semantics for automatic annotation of video segments. The approach is the Hierarchical Video Decomposition. Clusters are formed when labels have the same locale while the shots having different locale tends far apart to form a cluster . Actually a video is formed in a small number of locale and in each locale small numbers of cameras are used.

The basis of cluster formation is the color histogram. The histogram is the tones attached to RGB component, there is a numerical value attached to tones of RED, GREEN, BLUE. The Histogram is useful in the study of Images in different aspect and now days on video. Hence, the color extracted from the shots becomes a basis of histogram, the distance is measured between these colors, and clusters are formed.

The other basis of clustering can be the textures of shots, similar motion characteristics, spatial moment's Audio features etc. The effect of labeling and clustering Is that suppose an half an hour video consists of 300 shots it is reduced to 20 units so the user now need to examine 20 instead of 300 shots. The indexing also becomes also easier.

The subsequence is classified as "dialogues","actions and "others". There can be repetition or lack of repetition. The user can look into the degree of it .A dialogue refers to actual conversation like montage presentation of two or more concurrent process, which has to be shown sequentially one after the other. The parallel events are interspersed by so-called establishing shots or shots of other parties, it can be called as noise shots or noise labels. The repetition can be

constructed and parsed to the shot label. In this paper the author has taken a video sequence of 15 shots, label sequence is taken, and noise is identified Suppose the video has labeled as given below

$\{A, B, A\}, X, Y, Z\{A, B, A, B, A\}, B, C\{D, E, F, E, D\}, E, G, H, I$.

The equally label shots may contain same object or background the reason to believe is that because the labels are made by the visual content of the shots. Now considering the above label sequence it is stated that since the first label sequence is left out because of starting sequence the next label A,B,A,B,A,B is dialogue , the next label DEFEDE is also a dialogue but it is having a noise "F".

When there is a contrast in visual data content, there is a fast movement of shots in a video and this can be called as action. In such a sequence, there is typically little or no recurring of shots taken from same object of the same person and same locale. Such a sequence of shots constitutes an action event.It is reported that dialogue and event constitute 50 to 70 percent of a video by MM Yeung et.al [6].

Advantages -: The advantages can be applied on navigating, searching, browsing and viewing. Such representation, provide nonlinear access to video and give quick views of the visual content. This gives rapid nonlinear viewing of video

Disadvantages The said approach do not talk about the precision or recall of the content retrieved. The visual pattern that has been defined in video may or may not match with that users want to search. Does not say about match when textual pattern is combined with visual pattern.

2.2 The next paper, which is going to be discussed, are the contribution made by Wen-gang cheng et.al [7]. In this paper author used a Shot Cluster Tree. The time adjacent and visual adjacent contents are grouped into shot groups and then shot groups are clustered into shot clusters with simple agglomerative hierarchical clustering method. After hierarchical clustering method is applied there is formation of different levels and these levels constitute the shot cluster tree .The approach can be better represented by the following figure that has been given the above contributor .The diagram consists of video which is divided into shot cluster 1,2 ….n then these clusters are forming the groups and at last comes the shots numbers. Therefore, it can be said it follows the hierarchical structure hence it is named as hierarchical structure tree. At the top it is the video and at the leaves there are clusters and number of clusters that are forming the group and group forming the shot cluster and which constitutes the video The shot cluster tree is represented in the diagram given below.

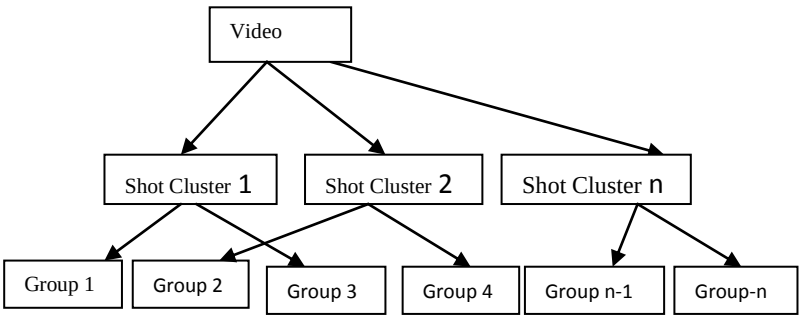

Fig-1 Structure of Shot Cluster Tree [7]

Many methods are used for shot boundary detection. There are five types of classification for automatic shot boundary detection 1),pixel based ,2)statistics based ,3)transform based, 4)feature based and 5)histogram based.

The author proposed a sliding shot window. In sliding shot window the each shot is compared to the shots in that window. Suppose a a shot window has length L(L is the number of shots in that window) the method is simple it is just to calculate the similarity between the current shot with next L shots(the next L shots is the destination shots) and the window moves according till the end is met. The author uses the color for the key frame the similarity with the key frame is measured between shot i and shot j where Fi and fj are key frames are done by $\text{Sim(i,j)}=\sum_{k=1}^{N} \min(HFi(k), HFj(k))$ where HF_i and HF_j are normalized histogram for the two key frames and N is the number of bins used in the histogram. The similarity between a shot i with group Gk is defined by the author as $\text{Sim SG(I,k)}= \max_{j \in Gk}(\text{Sim(i,j)})$ where j is a shot group G_kThe similarity was defined as $\text{SimGG(k,l)}= \max_{i \in Gk, j \in gi(}\text{Sim(i,j)})$.

Advantages : it can identify shots of similar content. It can also access non linear video sequences.. It supports Query by content and as well as Query by Example.

Disadvantages : The proposed approach does not define the accuracy and the time factor. It also does not clarify how to decide the Key Frame. The proposed algorithm also does not say how to decide the Thresh hold value and how to decide on which thresh hold the clustering the clustering will end.

2.3 The next contribution which is going to be discussed is contribution made by Sameh megrhi et al[8]. In this paper the author uses the feature selection for content based video retrieval systems. The derived features are compared with recorded feature with Hausdorff Distance. Feature selections are used in this technique because it best feature describes the reliability and efficiency of the video content. Selection of feature is important because if best feature is not selected then it can lead to bad recall and precision and can have longer execution time. The recall can be defined as the ratio between the total number of relevant video retrieved and total number of relevant video, which can be mathematically expressed as

$$\text{Recall} = \frac{\text{Total number of relevant video retrieved}}{\text{Total number of relevant video}}$$

The precision can be defined as

$$\text{Precision} = \frac{\text{Total number of relevant video retrieved}}{\text{Total number of video retrieved}}$$

The paper uses the Hausdorff distance, which was named after Felix Hausdorff. The Hausdorff is used in metric space in mathematics. A metric space can be stated as a metric space is a space for which all distances measured for the elements of set are defined. It is a ordered set of (Y, d) where X is the set and d is the metric on Y. It can be represented by the function d:Y xX → R. To outline metric space is beyond the scope of this paper. Now Felix Hausdorff gave a outlet of metric space. If the distance between a point **a** in set **X** and a point **b** in set **Y** is Hausdorff distance when **a** of set **x** is on supremum from **b** of set **Y** and **a** of set X on infimum and **b** of set Y is on supremum. Therefore, the distance between the two sets of points is the Hausdorff distance.

The same concept has been applied in this paper where the author has used the metric between the shots using Hausdorff metric. The basic Spatio – temporal Feature that are used in this paper are let there be a Pixel p in a frame F then the position vector associated with pixel p is the defined as P_F = F(x,y) where F is the frame and (x,y) is co-ordinate position of pixel. Then the next frame will be F+1 the pixel p moves u in the direction x and v in direction y. Then the displacement vector will be given as P_{F+1} = (F+1, x+u , y+v). Then the vector generated by motion of the pixels is given by (1,u,v) of a color patch can be estimated by classical optical flow techniques. The projection of the motion vector in the (t,x) and (t,y) planes define two motion angles α_x and α_y between these planes and two lines L_x and L_y supporting the motion vector. Let O be the point at the center of a video frame group , and let O_x and O_y be the projections of O onto planes (t,x) and (t,y) as defined .

Therefore for the pixel p = (t, x,y) in the frame stack , two distances are defined D_x ranging from O_x to L_x and D_y ranging from O_y to L_y . Therefore the position component D_x and D_y of the motion vectors for the pixel p can be calculated as 7 D feature vector where the Feature Vector (FV) can be calculated as FV = (C_1, C_2, C_3, α_x, D_x, α_y, D_y) the definition is due to D. Dementhon et. al [9]. After the Feature Vector is computed the next step is to cluster these process, the most stable pixels are selected . Then the Basic Spatio Temporal feature matrix A is computed as A=(m*n) where m represents the number of stable pixels in a group of frames and n=7 is the dimension of Feature Vector. Then A is represented as

$$A=\begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} & a_{q6} & a_{17} \\ a_{21} & a_{22} & a_{23} & a_{24} & a_{25} & a_{26} & a_{27} \\ a_{31} & a_{32} & a_{33} & a_{34} & a_{35} & a_{36} & a_{37} \\ a_{m1} & a_{m2} & a_{m3} & a_{m4} & a_{m5} & a_{m6} & a_{m7} \end{pmatrix}$$

where, the first 7 D ,, Feature Vector FV in the feature Matrix A is FV_1={ a_{11} a_{12} a_{13} a_{14} a_{15} a_{16} a_{17}}.The author uses Hausdorff distance, as Euclidian distance is very sensitive to noise. The Hausdorff distance is implemented in video retrieval is given by. Let d be the distance between the Query and the video dataset is given by.

(Hausdorff Distance)HDd(Q_v,V D_s)=max{h(Q_v, V D_s),h(V D_s, Q_v) . where Q_v is the Query video and V D_s is the video databases.Now the lastly comes the discussion based on advantages and disadvantages of the contribution made by Sameh Megrhi et.al

Advantages: The recall and precision works well on the small video databases. It searches for the salient objects rather than scene. It has more accuracy on small video database The .acceleration rate Spatio temporal segmented object database feature extraction to Spatio temporal feature extraction is very much nearer to 50%

Disadvantages: The feature extraction consumes much time although parallel algorithm was introduced to reduce some amount of time. The algorithm was only for small database but I does not say about when the database is large enough. The recall increases with the number of videos but precision decreases when the number of video increases .

2.4 In this section the discussion on the contribution made by Milind Ramesh Naphade et.al[10].The retrieval of video is based on multi nets or multijects using probabilistic feature and using Bayesian Belief Network. Multiject model as defined by the author is the probabilistic multimedia object which summarizes time sequence of features extracted from multiple multimedia. Multijects can of three categories like Objects , sites and events. The objects can be real world objects like ship, bike, airplane etc the sites can be sky , water mountain etc and events can be classified as game of cricket, vehicles moving etc. These are the low features but with these features, inferences can be drawn. Semantic concepts[17] are related to each other. The author states a very interesting fact about semantic relationship and these semantic relationships can be used to draw interesting inferences. This states that certain multijects used can be high probability of certain other multijects like it has been stated in this paper is that car , road and sky will give high probability of a road or High way and this may not give the water body or campus of any building. The conceptual figure as given by Naphade et.al is as given below.

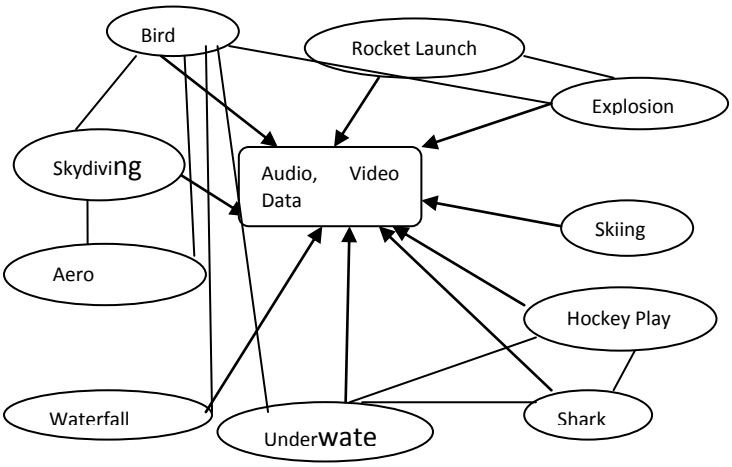

Fig-2 Conceptual Figure of Multiject Model[10] Figure by Milind Ramesh Naphade et.al.

This figure explains that there is strong interaction between Airplanes or there can be strong interaction between Underwater and shark but there is a likely very weak interaction or no interaction between underwater and playing of Hockey Game or Shark Playing Hockey Game.

The process applied by the author is Preprocessing and Feature extraction. In this process video clips are segmented by the multiple cues [11], then spatio- temporal segmentation is done using [12] these two algorithm is applied separately so that homogeneous regions are obtained for color and motion. For a large video, the author uses the artificial cuts are introduces every two seconds. The huge segmentation thus obtained by the above two methods are then merged morphological operation and the morphological operation is done using the basis of coherent motion and weak edges. Each regions having similar segmentation were then processed to extract a set of feature characterizing the visual properties like the color, texture, motion and structure of regions. How on the basis of colors there is a brief approach by the author.

Color-: A normalized linear three channel of HSV histogram is used H(Hue) Saturation(S) and V (Intensity), 12 bins for each is selected.

Texture-:The Gray Scale evaluating technique was using Gray Level Co –occurrence Matrix(GLCM) . The GLCM is a

statistical measure which is in use to measure texture analysis. . Let intensity values of a pair of pixel p(I,j) is denoted by

$P(i,j,d,\theta) = \frac{P(i,j,d,\theta)}{N(d,\theta)}$ where P(.) is the GLCM for the

Displacement vector d and orientation θ and N(.)Is the normalizing factor. The computation of GLCM of the channel V using 32 Gray levels and 4 orientations corresponding to θ values 0, 45, 90 and 135 degrees. The current pixel with adjacent pixel is at a distance of 1 unit so the d is equal to 1(d=1) for each of the four matrices six statistical feature of GLCM is evaluated. The features taken are contrast, Energy, Entropy, Homogeneity, Co relation and Inverse Difference Moments[18].

Structure -: To capture the structure within each region, a Sobel operator with 3x3-window size is used. Using this method an 18-bin histogram of edge direction is obtained.

Motion-: The inter-frame affine motion parameter for each region tracked by spatio- temporal segmentation algorithm is used as motion feature.

Color Moments -: For the three channels H S and V first order Moments and Second order moments are calculated. All 98 features are extracted to represent to represent the visual properties of the region of which 84 features (color ,texture, structure, and moments) are used for sites.

The author then preprocesses the audio track due to limitation of this paper this part is excluded.

Now this section focuses on how the multijects is based on video .Let $\overrightarrow{x_j}$ be the feature vector for region j. The two hypotheses are defined by the author as H0 and H_1. Under each hypothesis, it is assumed that the feature vector is drawn from distinct probability distribution.$H_0 : \overrightarrow{x_j} \sim P_0(\overrightarrow{x_j})$ & $H_1 : \overrightarrow{x_j} \sim P_1(\overrightarrow{x_j})$ Where $P_0(\overrightarrow{x_j})$ represents the concept is absent I,e null hypothesis and $P_1(\overrightarrow{x_j})$ represent the concept is present. The true and null hypothesis is presented in form of diagonal Gaussian component (GMM), objects, and events Hidden Markov Model is used (HMM).The author took MPEG streams of decompressed nature to perform shot boundary detection ,spatio temporal video region segmentation and tracking subsequent feature extraction. Initially the author took about 1800 frames were used for training and segmentation and another 9400 frames were taken for testing. The pixel size of 176X112 .The detection performance is on average 81 percent near about and False Alarm average was 15.88 percent. To get the frame level the contributor has fused region level features. For Binary Classification of each concept in each region the the binary variable R_{ij} is given and $_{Rij}$ is defined by -:

R_{ij} = 1 if concept i is present in region j
0 otherwise

Then Bayes Rule is pplied for any concept is present or absent in any region.

$$P = \left\{R_{ij} = 1 \middle| \overrightarrow{xj}\right\} = \frac{P\left\{\overrightarrow{xj} \middle| R_{ij} = 1\right\}}{P\left\{\overrightarrow{xj} \middle| R_{ij} = 1\right\} + P\left\{\overrightarrow{xj} \middle| R_{ij} = 0\right\}} \quad eq-1$$

The multijects used by the author are region level detectors . For fusing regional information at frame level the frame level features are adapted. In frame F_i if there are N number of concepts it can be expressed as a set F_i ε { 1,2…N} where N is the number of concepts. Let there be M number of regions in a frame. Then it can be expressed as

$\chi = \left\{\overrightarrow{x_1}, \overrightarrow{x_2}, \cdots, \overrightarrow{x_n}\right\}$Again the fusion of region is defined as -;

$P(F_i = 0|\ \chi) = \prod_{j-1}^{M} P\ (R_{ij} = 0|\ \overrightarrow{x_j})$

$P\ (F_i = 1|\ \chi) = 1 - \prod_{j-1}^{M} P\ (R_{ij} = 0|\ \overrightarrow{x_j})$ eq-2

Both the equation given by Minind et.al. .The binary random variables F_i, denoting the presence /absence of multijects at frame level have been derived from region level features for each concept i separately. However, the represent semantic concept sin a movie. Since they convey meaning within context , the must interact. Sine the random variables F_i do not capture this interaction , so another set of variables are taken as T_i, i ε { 1,2, … N} Ti and Fi represent the same concept. The dependence between the various multijects at frame level is modeled in the conditional distribution $P(F_i \mid T_1, \ldots T_N)$. The next step is the training of data and before the training probability is being found out and the probability is is evaluated on the basis of $P(F_i = 0|\ \chi) = \prod_{j-1}^{M} P\ (R_{ij} = 0|\ \overrightarrow{x_j})$

$P\ (F_i = 1|\ \chi) = 1 - \prod_{j-1}^{M} P\ (R_{ij} = 0|\ \overrightarrow{x_j})$ .During the inference phase in a two layered Bayesian network two layered approach is taken where layer 0 is taken as feeder into the network and in the layer 1 inference is drawn on layer 1 .If there is interaction then 1 otherwise 0 given by $P(T_i = 1 \mid F_1, \ldots F_N)$ and $P(T_i = 0 \mid F_1, \ldots F_N)$. To improve the detection performance the author has used the Neyman Pearson Criterion . The Neyman pearson lemma [13] states that when two Hypothesis are competing then only one Hypothesis is selected at the cost of other at some threshold value. The basis on which the one hypothesis is selected is done by the likely hood ratio test on some conditional probability. To improve

Both the equation given by Minind et.al. .The binary random variables F_i, denoting the presence /absence of multijects at frame level have been derived from region level features for each concept i separately. However, they represent semantic concept sin a movie . Since they convey meaning within context, the must interact. Since the random variables F_i do not capture this interaction , so another set of variables are taken as T_i, i ε { 1,2, … N} Ti and Fi represent the same concept. The dependence between the various multijects at frame level is modeled in the conditional distribution $P(F_i \mid T_1, \ldots T_N)$. The next step is the training of data and before the training probability is being found out and the probability is is evaluated on the basis of $P(F_i = 0|\ \chi) = \prod_{j-1}^{M} P\ (R_{ij} = 0|\ \overrightarrow{x_j})$

$P\ (F_i = 1|\ \chi) = 1 - \prod_{j-1}^{M} P\ (R_{ij} = 0|\ \overrightarrow{x_j})$ .During the inference phase in a two layered Bayesian network two layered approach is taken where layer 0 is taken as feeder into the network and in the layer 1 inference is drawn on layer 1 .If there is interaction then 1 otherwise 0 given by $P(T_i = 1 \mid F_1, \ldots F_N)$ and $P(T_i = 0 \mid F_1, \ldots F_N)$. To improve the detection performance the author has used the Neyman Pearson Criterion . The Neyman pearson lemma [13] states that when two Hypothesis are competing then only one Hypothesis is selected at the cost of other at some threshold value. The basis

on which the one hypothesis is selected is done by the likely hood ratio test on some conditional probability. To improve detection performance the author has placed a bound on the false alarm probability and then maximize the detection rate subject to this constraint maxδ $P_D(\delta)$ subject to $P_F(\delta) \leq \alpha$. Where α is bound on the false alarm rate . This was achieved by Likelihood ratio test. Now the author has drawn a ROC curve I,e receiver operating Characteristics at frame level is obtained by -:

$$\frac{P(F_i = 1|\chi)}{P(F_i = 0|\chi)} > \tau \text{ the range of } 0 \leq \tau \leq \infty \; i \in \{1, \ldots N\}$$

Where N is the number of multijects

Advantages -: This approach can be applied on meaning ful filtering of content. Large amount of Multijects can be handled using this framework. The authors have proposed a flexible architecture semantic video indexing.

Disadvantages: There should be ability of multinet to seamlessly integrate multiple data simultaneously and there is also a need to develop to retrieve a video having dynamic events in a video.

This paper has best tried to include the important features of the paper by Milind Napahde et. al .

2.5 This section describes the paper by Ayesha Slahuddin et.al [14]. In this the author has retrieval technique by Particle Swarm Optimization. Let's first describe in brief about Particle Swarm Optimization (PSO). This is technique to navigate the space of search within certain parameter to maximize the objective. The name was due to James kennedy and Russel C. Eberhart[15][19]. Particle Swarm Optimization the name is drawn from the real world as group of birds together fly for the food and individual bird is also is search for food. Optimization is the mechanism by which finds the maximum and minimum value of a function or a process. Like if a function f is maximized then –f will be minimized . Mathematically it is expressed as f: $R^n \rightarrow R$ and then to find $\hat{X} \varepsilon R^n$ such that $f(\hat{X}) \leq f(x), \forall X \varepsilon R^n$ Similarly , a maximization task is defined as Given f: $R^n \rightarrow R$ find$\hat{X} \varepsilon R^n$ such that $f(\hat{X}) \geq f(x),), \forall X \varepsilon R^n$. The domain R^n is referred as search space each element of R^n is the candidate solution in the search space with $(\hat{X})$ is the optimal solution, whereas the n denotes the number of dimensions of the search space, or the number of parameters involved in the optimization process. The f denotes the objective function, which maps search space to function space. The in this paper uses the Particle swarm optimization technique for the optimization of the search video. In this techniques also the video is divided into frames and that video is extracted which best matches with the query image and the frame of the video. The each particle of the swarm is the two dimensional representation i,e the video number and the frame number. After this fitness is calculated, fitness refers to how similar is the each particle video frame with that of query image. For calculating of the fitness the author has tried three similarity measure, these are (a) Convolution, (b) Co-relation Coefficient,(c) Score from SIFT algorithm. Once the fitness score is evaluated of each particle then this information is spread across the swarms global best particle in the search space. Where global best particle represent the best match .The author has given how the fitness score is being evaluated

If Current $_{particle}$ > Particle $_{best}$ then

$Particle_{best} \leftarrow current_{particle}$

End if

2.5 This section describes the paper by Ayesha Slahuddin et.al [14]. In this the author has retrieval technique by Particle Swarm Optimization. Let's first describe in brief about Particle Swarm Optimization (PSO). This is technique to navigate the space of search within certain parameter to maximize the objective. The name was due to James kennedy and Russel C. Eberhart[15][19]. Particle Swarm Optimization the name is drawn from the real world as group of birds together fly for the food and individual bird is also is search for food. Optimization is the mechanism by which finds the maximum and minimum value of a function or a process. Like if a function f is maximized then –f will be minimized . Mathematically it is expressed as f: $R^n \rightarrow R$ and then to find $\hat{X} \varepsilon R^n$ such that $f(\hat{X}) \leq f(x), \forall X \varepsilon R^n$ Similarly , a maximization task is defined as Given f: $R^n \rightarrow R$ find$\hat{X} \varepsilon R^n$ such that $f(\hat{X}) \geq f(x),), \forall X \varepsilon R^n$. The domain R^n is referred as search space each element of R^n is the candidate solution in the search space with $(\hat{X})$ is the optimal solution, whereas the n denotes the number of dimensions of the search space, or the number of parameters involved in the optimization process. The f denotes the objective function, which maps search space to function space. The in this paper uses the Particle swarm optimization technique for the optimization of the search video. In this techniques also the video is divided into frames and that video is extracted which best matches with the query image and the frame of the video. The each particle of the swarm is the two dimensional representation i,e the video number and the frame number. After this fitness is calculated, fitness refers to how similar is the each particle video frame with that of query image. For calculating of the fitness the author has tried three similarity measure, these are (a) Convolution, (b) Co-relation Coefficient,(c) Score from SIFT algorithm. Once the fitness score is evaluated of each particle then this information is spread across the swarms global best particle in the search space. Where global best particle represent the best match .The author has given how the fitness score is being evaluated

If Current $_{particle}$ > Particle $_{best}$ then

$Particle_{best} \leftarrow current_{particle}$

End if

2.5 This section describes the paper by Ayesha Slahuddin et.al [14]. In this the author has retrieval technique by Particle Swarm Optimization. Let's first describe in brief about Particle Swarm Optimization (PSO). This is technique to navigate the space of search within certain parameter to maximize the objective. The name was due to James kennedy and Russel C. Eberhart[15][19]. Particle Swarm Optimization the name is drawn from the real world as group of birds together fly for the food and individual bird is also is search for food. Optimization is the mechanism by which finds the maximum and minimum value of a function or a process. Like if a function f is maximized then –f will be minimized . Mathematically it is expressed as f: $R^n \rightarrow R$ and then to find $\hat{X} \varepsilon R^n$ such that $f(\hat{X}) \leq f(x), \forall X \varepsilon R^n$ Similarly , a maximization task is defined as Given f: $R^n \rightarrow R$ find$\hat{X} \varepsilon R^n$ such that $f(\hat{X}) \geq f(x),), \forall X \varepsilon R^n$. The domain R^n is referred as search space each element of R^n is the candidate solution in the search space with $(\hat{X})$ is the optimal solution, whereas the n denotes the number of dimensions of the search space, or the number of parameters involved in the optimization process. The f denotes the objective function, which maps search space to function space. The in this paper uses the Particle swarm optimization technique for the optimization of the search video. In this techniques also the video is divided into frames and that video is extracted which best matches with the query image and the frame of the video. The each particle of the swarm is the two dimensional representation i,e the video number and the frame number. After this fitness is calculated, fitness refers to how similar is the each particle video frame with that of query image. For calculating of the fitness the author has tried three similarity measure, these are (a) Convolution, (b) Co-relation Coefficient,(c) Score from SIFT algorithm. Once the fitness score is evaluated of each particle then this information is spread across the swarms global best particle in the search space. Where global best particle represent the best match .The author has given how the fitness score is being evaluated

If Current $_{particle}$ > Particle $_{best}$ then

$Particle_{best} \leftarrow current_{particle}$

End if

$D_e(p,q) = 0$ when match is exactly which is the fittest score and no score otherwise.

Now this paper discusses the advantages and disadvantages of the proposed approach.

Advantages: The author has taken about 20 videos and each video consists of nearly about 3500 frames hence the are about 70,000 frames to be compared .Using Particle Swarm Optimization technique this comparison has been brought down to 700 I,e 1% of the search space is used . The Comparison between the target frame and input image is drastically reduced. This is done small video database.

Disadvantages: The search space is reduced when the video database is small but it does not say about when the repository is large. Secondly, this paper does not say about the precision I.e. with how much precision the target frame is achieved .Thirdly it does say anything about the recall

3. CONCLUSION

This paper has best tried to include the important points of the contributors and strived hard to include the advantages and the Limitations of each of the paper included. There is ample research that is going on in the field of Content Based Multimedia Retrieval systems. There are many approaches also. There is a dearth need for retrieval systems for large video databases. The future of Multimedia retrieval could be a systems with much higher precision and recall rate.

4. ACKNOWLEDGEMENT

A sincere thanks to all researchers, academician whose paper have been cited. A Sincere Thanks to my supervisor Dr. Syed Qamar Abbas who guided me to complete my survey work.

5. REFERENCES

[1] Toshikazu Kato "Database Architecture for Content Based Image Retrieval "Proc. SPIE 1662, Image Storage and Retrieval Systems, 112 (April 1, 1992)

[2] www.internetlivestats.com/google-search-statistics/

[3] https://www.youtube.com/yt/press/statistics.html

[4] http://developer.android.com/about/index.html

[5] Content Based Digital Video Retrieval "RM Bolle ,Bl Yeo and MM Yeung " IBM Thomas J Watson Research Center,USA, on International Broadcasting Convention 12-16 September 1997 Conference Publication N0 447, IEEE 1997.

[6] MM yeung and B.L Yeo "video content characterization and compaction for digital library application", SPIE Storage and Retrieval for Image & video Databases,Vol SPIE3022 pp45-58 Feb 1997.

[7] Wen –Gang Cang , De Xu " Content based Video retrieval using Shot cluster Tree" Proceedings of the second International Conference on Machine Learning and Cybernetics , Xi'an,2-5 November 2003.

[8] Sameh Megrhi, Wided Souidene, Azeddine "Spatio – Temporal Salient Feature Extraction for Perceptual Content Based Video Retrieval" from University of Paris Color and Visual Computing Symposium-2013

[9] D. DEMenthon and Doermann, "Video retrieval of near duplicates using nearest neighbor retrieval of spatio-temporal descriptors", , vol 30 , pp 229-253, 2006.

[10] Milind Ramesh Naphade and Thomas S .Huang Fellow IEEE "A probabilistic Framework for Semantic Video indexing, Filtering, and Retrieval "IEEE Transaction on Multimedia Vol 3 No , 1 March 2001.

[11] M. Naphade, R. Mehrotra, A.M. Ferman, J. Warnick , T.S Huang and A.M. Teklap, " A high Performance shot Boundary detection algorithm using multiple cues" in Proc Fifth IEEE international Conference on Image Processing Vol 2 , Chicago ,II pp 884-887, Oct 1998.

[12] D. Zong and S. F chang , "Spation-temporal video search using the object based video representation" in Proc.International Conference on Image Processing , Vol. 2, Santabarabara CA pp21-24 Oct 1997.

[13] H. V. Poor, An Introduction to Signal Detection and Estimation, 2nd edition. New York: Springer-Verlag, 1999.

[14] Ayesha Salauddin ,Alina Naqvi, Kainat Mujtaba and Junaid Akhtar ," Content based Video Retrieval using Particle Swarm Optimization", in Proc. of International Conference on Frontiers of Information Technology pp 79-83 , 2012.

[15] James Blondin, Particle Swarm Optimization, 2009.

[16] Rafeal C Gonzalez and Richard E. Woods, Digital Image Processing.

[17] S.F Chang, W. Chen and H Sundaram," Semantic visual templates Linking features to semantics" in Proc Fifth IEEE International Coference Image Processing, Vol 3 Chicago pp531-535,Oct 1998.

[18] R Jain, Rkasturi and B Schunk, Machine Vision, Cambridge, MA MIT Press/Mc Graw Hill , 1995.

[19] Bae-Muu Chang, Hung –Hsu Tsai and Wen Lin Chou " Content Based Image Retrieval Bases on Image Features and Particle Swarm Optimization" Business and Information 2012