Video Event Classification Based on Image Search Engine

[1]Monika B. Khakhariya, [2]Prof. Tosal M. Bhalodia,

[1]Student, [2]Assistant Professor,

[1]MECE, Atmiya Institute of Technology and Science (AITS), Rajkot, Gujarat

____

***Abstract*—The bulk of video in the web is huge and analytic specific allotment of a video can be accomplished application a agreeable or argument based video indexing, grouping, analytic and retrieval approaches. To apprehend agreeable based video searching, in this plan we adduce arena based video comment for the identification and labeling of contest and altar in a video with a anecdotic text.Video comment requires a ability abject to ascertain semantic acceptation of contest and altar in the video. Manual and semi supervised video comment approaches abort as both crave ability for the actual identification and labeling of video concepts. Comment requires a abundant accord of abstraction annex and relatedness processing to accord anecdotic account to a arena in the video.**

***IndexTerms*— Video annotation, Joint Group Weighting Learning (JGWL),Near-duplicate segment**

____

I. INTRODUCTION

Due to the atomic advance of user-generated videos aggregate on the Internet and the abeyant demands in assorted multimedia applications, such as the Web video indexing, customer agreeable management, and open-source intelligence analysis,thetaskofmultimediaeventrecognitionhasreceived accretion absorption in contempt years. The multimedia accident acceptance involves the automated acceptance of circuitous contest from a ample set of airy videos. This assignment is acutely arduous due to four factors: (a) multimedia contest are higher-level circuitous descriptions of multimedia data, which cover several semantic concepts of objects, scenes and animal actions; (b) contest generally accept significant intra-class variations and inter-class similarities on both the visual appearance and semantic concepts; for example, both the events of 'birthday party 'and 'we dinning ceremony' may contain objectssuchascandlesandcakesasshowninFig.1,butthese altar ability not arise in all the videos that accord to the sameeventcategory;(c)videosareoftentakenbyconsumersin airy environments with altered recording devices; asaresulttheyusuallycontainsignificantvisualvariationsand noise[1];and(d)therearedtypicallyonlyfewpositiveexamples for complicated events. Many efforts are fabricated on evaluating the efficacy of low-level appearance [2], [3]. However, as contest are generally characterized by affinity in semantics rather than visualappearance,recentapproachesstarttousehigh – levelsemantic concepts to abetment inthe recognition of events.

One of the popular video retrieval approaches adopted by large multimedia search engines such as YouTube is using video descriptions embedded in the meta tags and titles of each video document. In 2014, considering YouTube video, 300 hours of videos are uploaded every minute and everyday hundreds of millions of hours YouTube video are watched [1]. To facilitate searching processes, YouTube provides manual annotation service which is tiresome, costly and time consuming. A lot has been done on video annotation but most of the works did not go beyond annotating objects and events in a shot. Video annotation is different from image annotation as users want not only object description but also events, object motions have to be annotated and presented at scene level rather than key frame level. In video, an object may have several role in a scene unlike a key frame. The main challenge in video annotation is providing descriptive object and event annotations within a scene while keeping scene level frame dependencies and spatiotemporal correlations between objects.

II. RELATED WORK

In this A generic video annotation framework in [9] considers every shot on an individual basis whereas expansion a video. This framework contains 2 layers of similarity analysis; the primary layer computes similarity between videos in a very datas et and also the question video, whereas the second is employed to differentiate and merge similar annotation texts from similar videos. Video signature, a compact illustration of spatio-temporal options, is generated so as to perform video similarity and later to search out free-text annotations from similar videos. These video signature is generated exploitation SIFT options [10] extracted at some flight points in a very video. A video matching is finished by activity a similarity comparison between eve ry and each flight purpose of the question video and a video in a very dataset normalized with the minimum range of trajectories found in either of the 2 videos.

Once the similarity is computed, Stanford information processing Log-linear PartOf-Speech Tagger is employed to spot objects exploitation nouns, and actions exploitation verbs. The authors used WordNet [11] in bridging word distinction in annotation and calculate similarities between terms found in numerous videos. Finally, ConceptNet [12] is employed to derive the particular conception relationship between objects flight of key frame. This generic framework considers similarity between single shot videos wherever there's no would like for scene analysis and conception dependency. the final word result expansion in such videos is a lot of or less like expansion a picture. Hence, this generic framework depends on a trial annotation instead of a full video annotation. AN automatic video annotation through search and mining is projected in [13] to use overlap within the content of reports video to mechanically annotate similar videos. a picture feature combined with text, and SVM conception options area unit accustomed hunt for connected videos. This work is predicated on one shot video as a result of with one text question it should not be doable to explain an outsized video. Once connected set of videos area unit generated transcripts of those videos area unit deep-mined as AN input for the annotation method.

A term frequency vector is made for every video transcript to pick those with maximal range of incidence to be used for annotation. the matter here is however similar those searched videos would be. The results of annotation is extremely enthusiastic about the similarity between videos and, it's unlikely to induce full length annotated videos terribly the same as the input video. compared with transcripts, subtitles area unit a lot of out there in documentaries and films. A framework for cluster primarily {based} image retrieval and video annotation is projected in [14] that uses region based wave remodel similarity methodology to match question video with preannotated videos. A video frame is split into 4X4 region and a feature vector is taken into account victimization the center of mass purpose of the wave reworked information at the region. Once the feature vector is built, similarity is calculated victimization earth movement distance live between identical blocks of frames of every video.

The basic downside of this analysis work is, it takes the full annotation text of comparable video into the POS tagger and directly uses those known set of objects and events within the sentence merging module. during this case, not all known events and objects might presumably be within the question video and also the temporal arrangement of these events and objects isn't glorious. although a video is split into frames and annotation is predicated on key frames of comparable videos, the approach disregards context; i.e. an occurrence found in one video frame, might not continuously have an analogous context in another video frame. As a result, unrelated set of events is also generated from such annotation works. metaphysics and rule learning primarily based video annotation and retrieval is projected in [15]. The metaphysics contains ideas, idea instances and their linguistic relationship from WordNet and rule learning is finished from the metaphysics. Once a rule is learned, it's applied to the metaphysics that contains instances, obtained victimization linguistics classifiers, to mechanically extend the video annotation. for instance, if metaphysics contains ‘‘airplane’’, ‘‘sky’’ and ‘‘ground’’ instances, the observer detectors “airplane take-off ” is that the associated. These detectors are created victimization the Viola and Jones algorithmic rule (provided by OpenCV) and color-based pel classification with a support vector machine, to observe and localize objects. Then, the spatiotemporal evolution of the looks of ideas is decided.

III. PROPOSED WORK

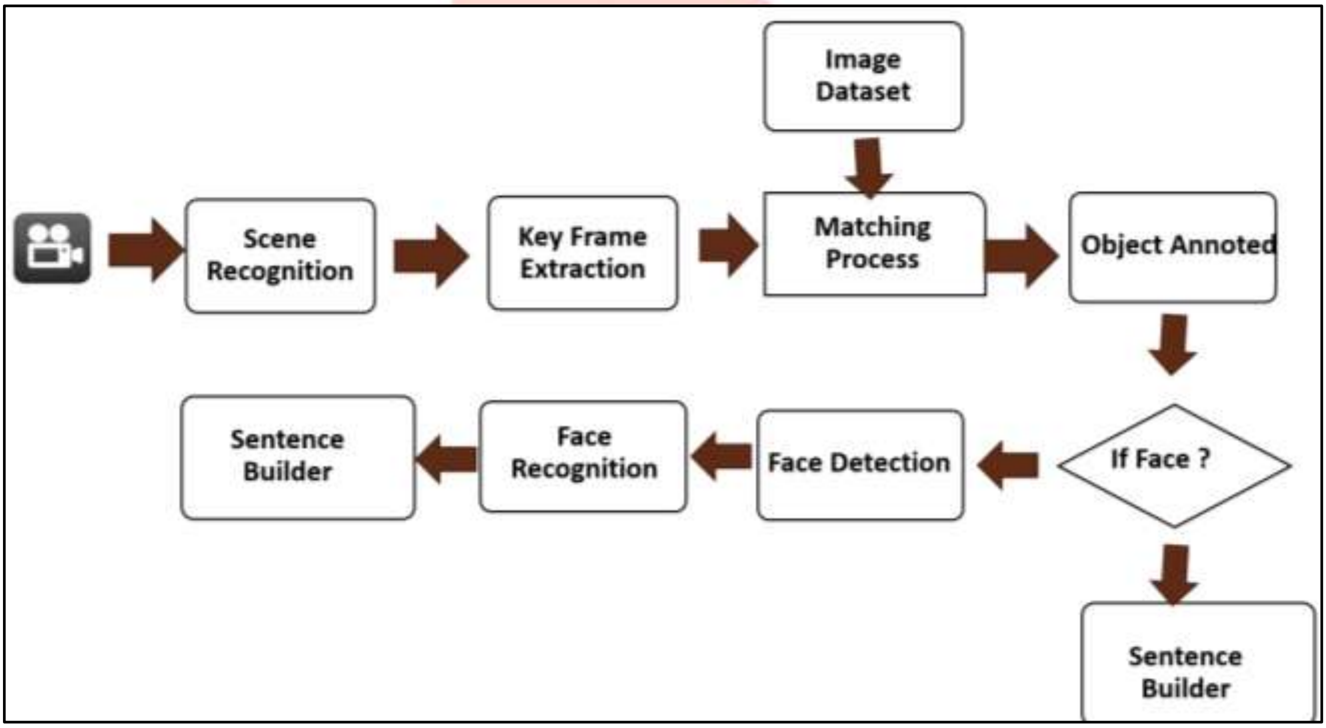


Fig 1. Flow of the proposed approach

Step 1: Input video

Take a cricket video for the pre-processing as input.

Step 2: Convert video into frames

The input video is converted into the number of frames.

Step 3: Key Frame Generation

Key frames ar those frames that best mirror the shot contents to represent the shot. The extracted key frames ought to covers the maximum amount similar content of the shot as doable and shut the maximum amount redundancy as doable. The options used for key frame extraction will embody colors (particularly the color histogram), edges, shapes, optical flow, MPEG motion descriptors, motion vectors, MPEG distinct cos constant, camera activity etc.

Algorithm: Key Frame Generation

- Read all the frames.
- **For** each shot frame **do**
  Take 1st, last and Middle frame of each shot
  Store to another folder.

Step 4: Scene Recognition

We square measure finding shot boundary detection for various transitions. initial input are going to be video of any motion-picture show then it'll transfer to shot Boundary detection then in this cut fade and dissolve detection are going to be done that helps to spot the shots in motion-picture show. In every shot we tend to extract three frames for key frame extraction.

In each shot first frame last frame and middle frame are going to be take into account as key frames. Those all key frames square measure info for retrieval section. thus for higher retrieval it's required that there ought to be sensible shot boundary detection. Identification of actual cut fade and dissolve can offer higher retrieval.

In this module we tend to square measure finding scene detection for various transitions like abrupt transitions and gradual transition. Abrupt transition simple|isstraightforward|is simple} to find however gradual transitions aren't easy to spot as a result of these isn't eye catching sequence. solelycolor histograms may be used for cut detection. Edge amendment magnitude relation is employed for distinguishing cut, fade and dissolve. variance is employed for distinguishing fade and dissolves.

For cut detection we tend to square measure combining distinct circular function Transformation, Edge amendment magnitude relation and HSV (Hue Saturation Value) bar graph. CT (discrete circular function Transformation) represents frame as total of sinusoids and frequencies. It works on N X N blocks. These all options aren't decent one by one for any detection. Thus we tend to square measure combining these all options for higher transition detection.

Thus we are going to mix these all n create one new parameter for detection. Correlation is employed for matching method that done consecutive frame by frame.

Colour changes won't occur quickly among an effort. It'll solely occur between consecutive shots. Thus solely colour feature may be used for cut detection. If colour histograms between 2 consecutive frames square measure larger than threshold then there'll be shot amendment occur.

There's risk that there'll be no same edge between first and second frame. The quantity of input and output edges may be counted [3]. If each the frames belong to an equivalent shot then the quantity of input and output edges is additional or less same. Thus if distinction is extremely giant then shot amendment may be happens. If there square measure n pixels within the frame and therefore the in edges square measure Xin and therefore the out edges square measure Xout for frames f and f − one severally.

$$ECR = \max\left(\frac{X_{in}}{n}, \frac{X_{out}}{n}\right)$$

Fade and Dissolve transitions square measure combination of repetitive amendment in constituent intensities. Mistreatment window one will observe frames for his or her intensity amendment. If adequate numbers of pixels show repetitive amendment in their intensity inside the window, we will assign current frame to be enclose dissolve transition [12].

A bar graph could be a operate that calculates the quantity of observations that be every of the dislodge classes called bins. we discover bar graph for individual H, S & V elements of image we've got return to the important HSV bar graph that is nothing however the mix of all 3 individual Histograms.

If we have a tendency to let n be the overall variety of annotations and k be the overall variety of bins. This equation for individual H, S & V elements of image and that we return to the resultant HSV bar graph that is nothing however the permutation of all 3 individual Histograms.

$$n = \sum_{i=1}^{K} m_i$$

We are using single feature that can be used for every transition. That feature combines different methods for cut fade and other gradual transition.

<u>Algorithm for Scene Recognition</u>

```
Step-1: Frame Conversion
Step-2: Computer HSV histogram
Step-3: Find Discrete Cosine transformation
Step-4: Compute edge detection
Step-5: Find entropy
Step-6: Calculate standard deviation
Step-7: Go to step-2 while end of the frames
Step-8: Correlation of parameters
Steo-9: Find unique parameter with combination of this all features.
        New parameter= 1 / (hsvhistocorr+graycorr+histocorr+edgecorr)
Step-10: Find Mean and Standard deviation of new parameter
Step-11: Calculate Thresholding
        Tb = Mean + a * standard deviation where a is constant
Step-12: Start from 1st frame
Step-13: If new parameter of frame  > Tb
          Strong cut detected
          Else If new parameter of frame > (mean + std)
          Weak cut detected
```

```
Step-13 If entropy of frame is zero
        Fade detected
Step-14 If accumulated difference of frame  > Tb
        Gradual transition detected
Step-15 Go to step-13 while end of the frames
```

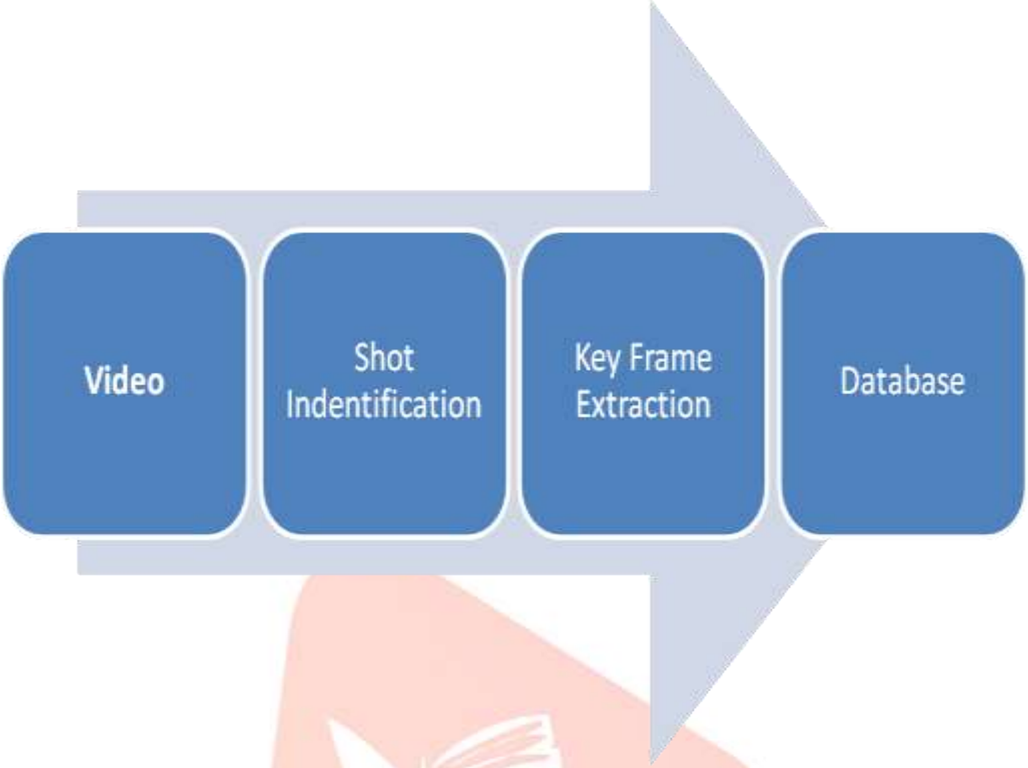


Fig 2 Proposed Framework for Shot boundary detection

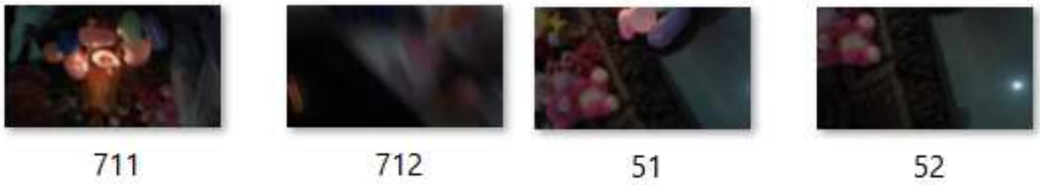


Fig 3 Cut Detection

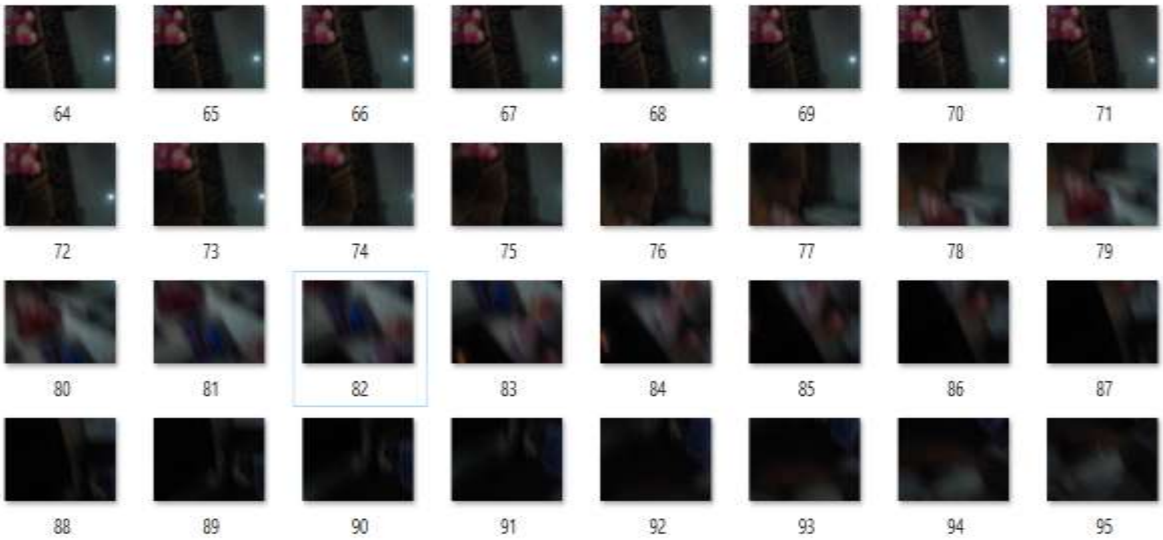


Fig 4 Fade Detection

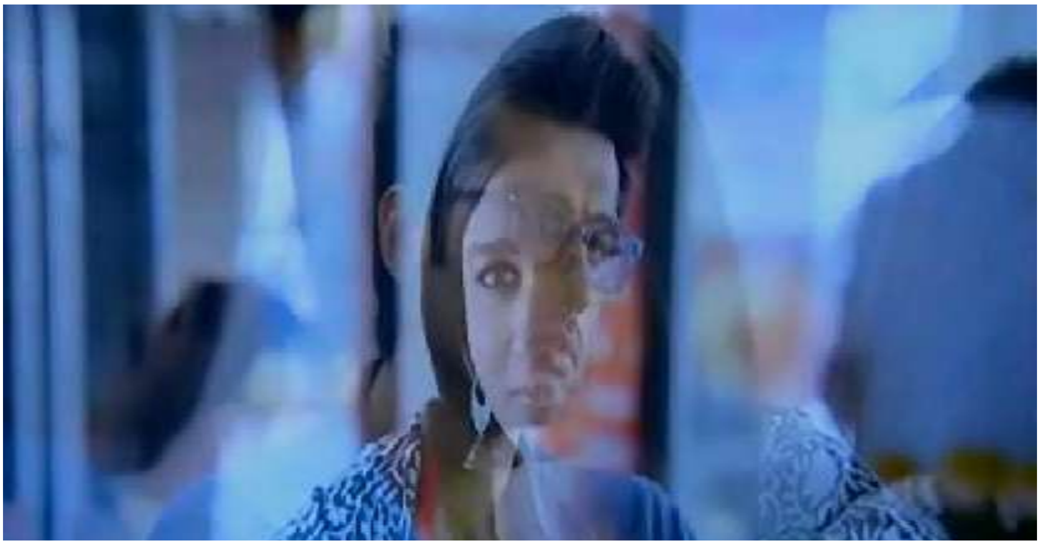


Fig 5 Dissolve Detection

Step 5: Matching Process

After feature extraction matching method may be a key component. mistreatment Visual feature like color texture form we'll match key frames of our input video with image information set. however there's a problem that if there's face then it'll match with person class however it'll not discover similar face thus for that we'll use face recognition for higher matching method,

Another issue is that generally nearly ninety fifth image is same however object is completely different than there'll be false result for that we have a tendency to area unit mistreatment object detection and matching.

Target detection refers to the employment of high spectral resolution remotely perceived pictures to map the locations of a target or feature (often a plant species of interest) with a specific spectral or abstraction signature. Target detection or feature extraction encompasses a broad vary of techniques, together with measurements derived from individual bands and a lot of complicated ways designed to acknowledge distinct options by form, hyper spectral signature, or texture. Targets of interest area unit usually smaller than the element size of the image (sub element target detection) or area unit mixed with different non-target cowl varieties among a element, requiring techniques like spectral mixture analysis to discover the target species. Hyper spectral pictures area unit helpful in target detection as a result of they contain an oversized contiguous set of spectral bands, usually enumeration within the tons of to thousands, and supply giant quantities of high spectral resolution information. employing a hyper spectral image, the spectral properties of the target, like distinction, variability, similarity and discriminability, are often accustomed discover targets at the sub element level. The user specifies spectral finish members, that area unit the reflectivity spectra of the "pure" targets that occur across the landscape, and image process computer code is employed to characterize the extent of the target across the landscape. the choice of spectral finish members is comparable to the thought of distinguishing coaching areas in supervised classification, however the spectral finish member will then be used at a sub element level to discover the species of interest. Spectral finish members area unit usually generated within the field employing a field spectroradiometer. Then the image is processed mistreatment classification algorithms to discover the locations of the target species.

- ❑ We have taken each genre's 100-200 images for database for matching
- ❑ We are also using IMDB Images as database.
- ❑ Among them if object matched then its genre's will be annoted
- ❑ Matching will be based on shape color texture by sub pixels matching

Algorithm for Matching Process

```
❑ Load Image Database
❑ For each key frame
    ❑ For each data base frames
        ❑ RGB to Grey Key frame and Image Database frame
        ❑ Edge detection of both the frames let a and b
        ❑ Filter for a & b
        ❑ One to one pixel (Sub pixel matching) using correlation
        ❑ if it equals to 1
            ❑ Then image has matched object
❑ For each object matched image matched image name on database will be written in text file.
❑ Similar Objects name will be discarded
```

Step 6: Face detection

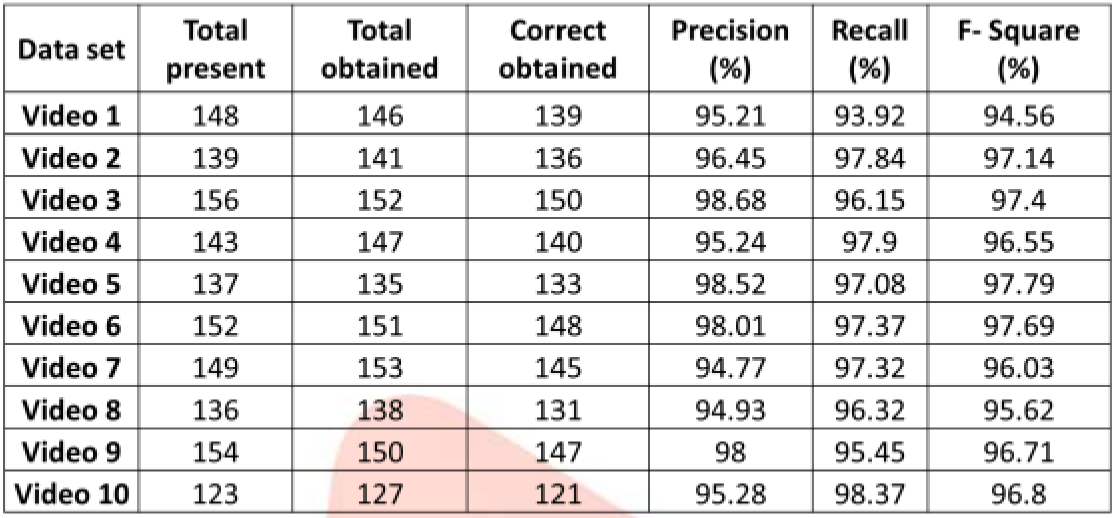
```
If Face Detected?
    Go to step 7
Else
    Go to step 8
```

```
Step 7: Face Matching
If any face matched with stored known database?
        Add label in text file
Else
        Go to step 8
```

Step 7: Indexing and labelling

Most relevant labels will be given in this step. Labels which are indexed 1-10 will be comes many times in every frames during matching process according to that we will give appropriate naming to it.

Table 1 Scene Recognition Analysis



Data set	Total present	Total obtained	Correct obtained	Precision (%)	Recall (%)	F- Square (%)
Video 1	148	146	139	95.21	93.92	94.56
Video 2	139	141	136	96.45	97.84	97.14
Video 3	156	152	150	98.68	96.15	97.4
Video 4	143	147	140	95.24	97.9	96.55
Video 5	137	135	133	98.52	97.08	97.79
Video 6	152	151	148	98.01	97.37	97.69
Video 7	149	153	145	94.77	97.32	96.03
Video 8	136	138	131	94.93	96.32	95.62
Video 9	154	150	147	98	95.45	96.71
Video 10	123	127	121	95.28	98.37	96.8

Table 2 Object detection Analysis

Images	Total present	Total obtained	Correct obtained	Precision (%)	Recall	F- square
bus	100	98	97	98.98	97	97.98
bears	100	101	97	96.04	97	96.52
car	100	99	95	95.96	95	95.48
cat	100	98	95	96.94	95	95.96
dog	100	99	97	97.98	97	97.49
flower	100	103	99	96.12	99	97.54
lakes	100	98	97	98.98	97	97.98
sunset	100	97	96	98.97	96	97.46
tigers	100	99	94	94.95	94	94.47
trees	100	98	95	96.94	95	95.96
waterfalls	100	99	97	97.98	97	97.49
castles	100	97	96	98.97	96	97.46
human	100	98	97	98.98	97	97.98

IV. CONCLUSION

As transmission growth there's vast quantity of videos that square measure unlabeled and if some person have gb's of videos except for naming them he or she should read that videos then provide correct naming. Thus for that the sphere of annotation has been introduced. During this paper we've mentioned all the techniques and trends that square measure within the field of videos annotation. Video has several content in it like visual options, foreground options, background options, objects audio feature and lots of a lot of, from that videos extracting helpful content for matching with the labelled pictures has difficult job. Thus for that we've studied numerous papers and from that we are able to propose a sturdy formula for any kind of videos like movies, cartoons, news, sports, parties etc.

We have implemented scene recognition which works very well in our approach then implemented key frame extraction for reduction of time. After that we have try to develop system which is used to detect object from various key frames. Our object detection is works very well in less informative image but it works not that much good for real time images of videos so we will try to improve matching of object in various key frames. Another approach is our matching process detects that it is a face so face detection has been done but it is of whom? We also have to detect that so we have implemented face recognition algorithm for better results. It gives good result for indoor videos but if person is far from camera or he/she is not showing their exact face to camera then it will be difficult to recognize. In this area there is more to work.

REFERENCES

[1] Wang, Han, and Xinxiao Wu. "Finding Event Videos via Image Search Engine." *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*. IEEE, 2015.

[2] Wang, Han, Xinxiao Wu, and YundeJia. "Video annotation via image groups from the web." *IEEE Transactions on Multimedia* 16.5 (2014): 1282-1291.

[3] Duan, Dingbo, and Jian Ma. "Automatic video annotation by motion recognition." *Progress in Informatics and Computing (PIC), 2014 International Conference on*. IEEE, 2014.

[4] Rouvier, Mickael, et al. "Audio-based video genre identification." *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 23.6 (2015): 1031-1041.

[5] Chou, Chien-Li, et al. "A novel video annotation framework using near-duplicate segment detection." *Multimedia & Expo Workshops (ICMEW), 2015 IEEE International Conference on*. IEEE, 2015.

[6] Sun, Shih-Wei, et al. "Automatic annotation of web videos." *2011 IEEE International Conference on Multimedia and Expo*. IEEE, 2011.

[7] Tong, Wenjing, et al. "CNN-based shot boundary detection and video annotation." *2015 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting*. IEEE, 2015.