A Machine Learning based Approach to Video Summarization

Varun Luthra*, Jayanta Basak† Prof. Santanu Chaudhury*,and K.A.N.Jyothi*,
*Electrical Engineering Department
Indian Institute of Technology,
Hauz Khas, New Delhi
Email: schaudhury@gmail.com
jyothi1984@gmail.com
† IBM India Research Lab,
New Delhi

***Abstract*—In this paper we have proposed a novel machine learning based approach to video summarization problem. Semi-supervised learning algorithm has been used to generate the summaries. Manually generated summaries (ideal summaries) serves as the labeled samples for the semi-supervised learning. Both visual and aural feature vectors taken over a set of videos are clustered and the individual video sequences are represented as vector-quantized time series. Then state transition machine based representation has been generated for both the complete class of videos and the labeled samples. A new information theoretic measure has been proposed for the goodness of a generated summary, which reduces the summarization process to finiding the sequence of frames for which the value of goodness measure is maximum.**

I. Introduction

The rapidly growing amount of digital video in todays network gives rise to the problem for efficient browsing and management of video data. To solve these problems, *Video summarization*, which offers concise representation of original video clips by showing the most representative sysnopsis is gaining more attention. There are two fundamental types of video summaries: *static video abstract*, which is a sequence of key frames and *dynamic video skimming*, which is a collection of dynamically-composed audio-video sub-clips, and in both cases the goal is to find the most interesting or important video segments that capture the essence of the original clips.

Some work has been done in the past in the field of video summarization in various prespectives. For static video summary, most early work selects key frame images by random or uniform sampling, like the Mini Video systems [Taniguchi 95]. Later work tends to extract key frame images by adapting to the video content. A mosaic based approach is suggested in [Lee 97]. In [Rui 99], the authors analyzed the video structure after video segmentation, and then got a tree structured Video-Table-Of-Contents(V-TOC).

For video skimming generation, in the VA abstract system [Leinhart 97], key movie segments are selected to form a movie trailer. The informedia system [Kanade 97] selects the video segments according to the occurence of important keywords in the corresponding caption text. Another effort in this field is the attention model by Ma *et al*. [Ma 02] that aims at bridging the gap between low level features and human perception by analyzing viewers attention to generate the summary. Sundaram and Chang [Sundaram 01] uses Kolmogorov complexity as a measure of video shot complexity, and compute the video summary according to both video shot complexity and additional semantic information under a constrained optimization formulation. Divakaran et al. [Divakaran 03] have used MPEG-7 motion activity and audio descriptor to generate the video summary. Parshin et al. proposed another approach for video summarization using user defined constraints and preferences in [Parshin 04]. Zhang *et al*. proposed a different approach to video summarization using graph modelling and motion attention [Zhang 05].

Though a lot of work has been done in the past, not much effort has been put in development of machine learning based algorithms for the purpose of summarization. Generating a perfect summary for a given video requires good understanding of video semantic content. It may be difficult to capture the semantics of a video using most of the existing techniques. Machine learning provides a way to capture the hidden semantic contents of a video sequence. We present a *semi supervised learning based algorithm* to generate the summaries for videos known to belong to a particular class of videos. Our algorithm assumes that the user will have a similar interest pattern over a class of videos. We have tested our technique extensively over two different class of videos and the subjective Precision-Recall graphs have also been drawn to find the effectiveness of our algorithm.

II. Video Sequence Representation

Conventional approaches to video summarization have used frame-based features to generate a representation for the videos. In this work, we have tried to capture the information content present in the transition between the frames. Various aural and visual features have been used here to capture various aspects of similarities.

A. Feature Extraction

1) Audio features: The most common audio classes in videos are speech, silence, music and the combination of later three. These classes can be well distinguisged by using Short Time Energy (STE), Zero Crossing Rate (ZCR) and Fundamental Frequency functions. The Short Time Energy function (*STE*) basically distinguishes speech and music and Short Time Zero Crossing Rate (*STZCR*) is used to seperate voiced speech from unvoiced one. Whereas, Short Time Fundamental Frequency (*STFF*) seperates audio into harmonic and nonharmonic classes. This way all the speech components are well distinquished using these three features.

Both *STE* and *STZCR* are calculated for every overlapping window of 511 samples of the audio signal with an overlap of 35 samples at either end of the window at a sampling rate of 44100 samples/s. The *STFF* of the audio segment is estimated over an overlapping window of 2048 samples with an overlap of 284 samples. When no fundamental frequency is estimated, the *STFF* is set to zero. Once these features have been extracted, different audio classes are characterized using statastical property of variance over overlapping windows of 140 feature samples with an overlap of 40 samples at either end of window. Thus we obtain a feature with a sample for every second of the audio. .

2) Visual Features: Color histogram, Edge histogram and Texture similarities are the three visual features that are used in the current study. In this work we employ the similarity between the histograms of two consecutive images as one of the features characterizing the transition. We use the histogram intersection technique to measure the similarity between fames. Let 'h' and 'g' be the two histograms. Then, similarity is defined as

$$sim(h,g) = \frac{\sum_{i=1}^{n} min(h(i), g(i))}{max(|h|, |g|)} \quad (1)$$

where, $'h'$ and $'g'$ are the two frames in the video sequence for which similarity measure has to be computed, $|h|$ and $|g|$ gives the magnitude sum of each histogram and *n* is the size of the histogram. The same equation has been used to find the color histogram, edge histogram and texture similarity measure between the two frames.

Following the extraction of audio and visual features we use feature fusion to integrate the two to form one combined audio visual feature vector. The audio features (with the exception of short time fundamental frequency) are produced at the rate of 1 sample / sec. whereas video features are obtained at 25 samples / sec. In order to bring them both to the same temporal scale we take the median value of visual features for every 25 features. This gives us a feature vector that is generated at a rate of 1 sample / sec. Such a feature vector combines both the audio as well as the visual cues in the video sequence while taking into account their interdependence.

B. Clustering of feature vectors

Once the feature vectors for all the videos in the dataset have been computed, self-organizing-map (*SOM*) approach proposed by Kohonen has been used to cluster the data. One of the reasons for choosing Kohonen clustering approach rather than conventional techniques like k-means etc. is that Kohonen technique creates a network that stores information in such a way that any topological relationships within the training set are maintained.

C. Sequential Representation of videos

Once we have clustered the feature vectors for the complete dataset, we go on with generating the sequential representation of individual videos. For each video, all the frame-transitions are taken and seen to which cluster it belongs. This way we have cluster numbers for all the frame-transitions in a video.

D. State Machine Representation

In our summary generation approach, we make an attempt to capture the semantics associated with a class of videos and hence we must have a mathematical representation for the complete class of videos.

Hence, we formulate the state machine representation for the class and also give complete derivations for transitions probabilities associated with the state machine. In our state machine represenation, states are the various clusters formed using the Kohonen clustering and we compute the transition probabilities for the state machine using the Bayesian approach. Our state transition machine can be considered as a first order markov process in which the probability of transition to the next state depends only on the current state.

1) Computing State Transition Probabilities: From the video sequence representation, we have cluster numbers for each of the transitions in the video sequence. Let us say that transition from frame 1 to frame 2 belongs to cluster number 'i' and that the transition from frame 2 to frame 3 belongs to cluster number 'j'. So, when we go from frame 1 to frame 3, we actually have a transition from cluster 'i' to cluster 'j'. In the same way, we count the number of transitions from each cluster to every other cluster for all the videos in the dataset. If N_{ij} is the number of transition from cluster 'i' to cluster 'j', then P_{ij}, transition probability from state 'i' to state 'j' is given by equation 2.

$$P_{ij} = \frac{(N_{ij} + \alpha_{ij})}{(\sum_{j=1}^{m} N_{ij} + \sum_{j=1}^{m} \alpha_{ij})} \quad (2)$$

Where, α_{ij} are the parameters associated with the dirichlet prior.

III. Summary Generation

A. Use of Ideal Summaries

By ideal summary, we mean a manually generated summary that the user feels captures the essence of the video. Such summaries are given as an input to the learning system. These ideal summaries are used to formulate a state transition machine (Q) and estimate the corresponding probabilities in the same way as we did for generating the state machine for the complete class of videos. These two state machines are in turn used to measure the goodness of any generated summary. So, the summary generation task reduces to finding the summary for which this goodness measure is maximum. The block diagram for the overall summary generation system is shown in the Fig 1.

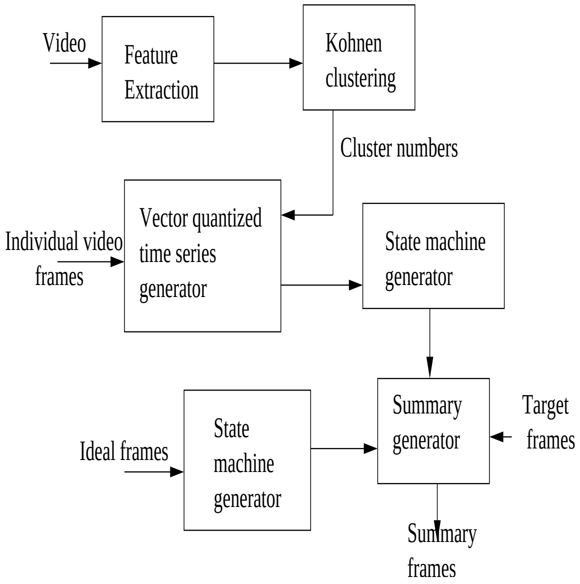


Fig. 1. Block diagram of summarization system

B. Finding goodness measure

Let us assume a m frame video for which the summary has to be generated. The given video is known to belong to a class and we have the matrices P and Q for this class of videos. For the sake of convenience, let us also assume that we have to generate a 4 frame summary (i.e the essence of the whole video can be seen in just 4 frames). So our task is to find the indices [p,q,r,s] of the four summary frames. Assume that the cluster numbers associated with these frames are [i,j,k,l]. Let us define I, the goodness measure for the generated summary as:

$$
\begin{aligned}
\boldsymbol{I} \;=\; & q_{c_i c_j} * \log \frac{q_{c_i c_j}}{p_{c_i c_{i+1}} * p_{c_{i+1} c_{i+2}} * \cdots * p_{c_{j-1} c_j}} \\
+ \; & q_{c_j c_k} * \log \frac{q_{c_j c_k}}{p_{c_j c_{j+1}} * p_{c_{j+1} c_{j+2}} * \cdots * p_{c_{k-1} c_k}} \\
+ \; & q_{c_k c_l} * \log \frac{q_{c_k c_l}}{p_{c_k c_{k+1}} * p_{c_{k+1} c_{k+2}} * \cdots * p_{c_{l-1} c_l}} \qquad (3)
\end{aligned}
$$

where, c_i is the cluster number associated with frame number i, p_{jk} is the transition probability from cluster number j to cluster number k of P matrix and q_{jk} is the transition probability from cluster number j to cluster number k of Q matrix.

For a summary consisting of a fixed number of frames, this measure basically evaluates the semantic closeness associated between the frame transitions. The value of the goodness measure is expected to be good for the sequence of frames whose frame transitions are relatively more probable for the given set of exam- ple summaries. That means the sequence of frames whose frame transitions best captures the semantics of the video corresponds to the required summary. Now, the problem left is to find out the sequence of frames that optimize goodness measure. We need an algorithm for the purpose of finding out this sequence of frames. Genetic Algorithms is one such class of algorithms.

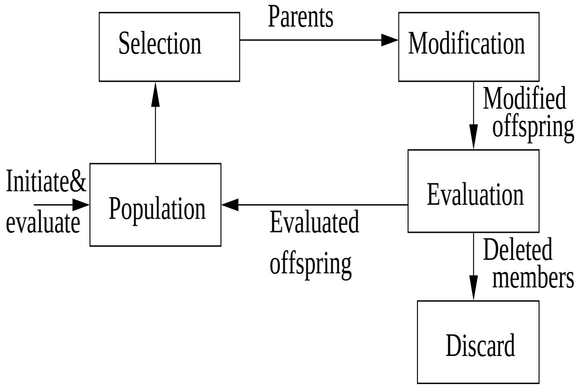


Fig. 2. GA evaluatory cycle

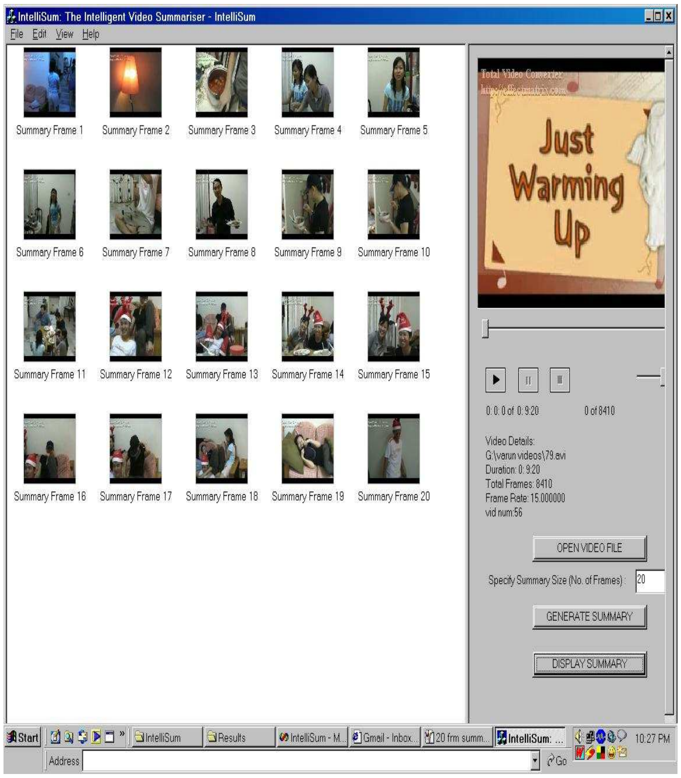
C. Genetic Algorithm

Genetic Algorithms encode a potential solution to a specific problem on a simple chromosome-like data structure and apply recombination operators to these structures so as to preserve critical information. A block diagram depicting the GA evoluationary cycle isshown in Fig 2.

In the present study, GA has been used to find the sequence of frames that optimize the goodness measure defined in equation 3. Here, we represent the sequence of frames by bit strings in a simple chromosome like data structure and the bit string should have a property that even after crossover and mutation operations it should represent a valid sequence of frame numbers. Once we have choosen the current generation population members, next step is that choosing the members from the current generation for the next generation according to their fitness. Roulette wheel selection method has been used for this purpose. Then reproduction of chromosomes will be done by Crossover and Mutation operations. In crossover, two chromosomes are used to generate two new chromosomes whereas, in mutation single chromose will be mofified to generate a new chromosome. Good values for crossover and mutation rates are usually around 0.7 and 0.01 respectively. Then, for the purpose of evaluation, the chromosome is first converted into a sequence of summary frame numbers. These numbers are then plugged into the goodness measure equation 3. The value for the goodness measure returned by this is taken as the value of the chromosome. These steps are repeated until a fixed number of iterations are done. The best chromosome thus found is converted to the sequence of frames which is presented to the user as the best found summary.

IV. Experimentation and Results

After the algorithm was designed and complete system was implemented, extensive testing was done to establish the effectiveness of our system. We have tested our system extensively over two different class of videos viz. home-shot party and Soccer videos. We have first tested our algorithm using only visual features and then tested by using aural

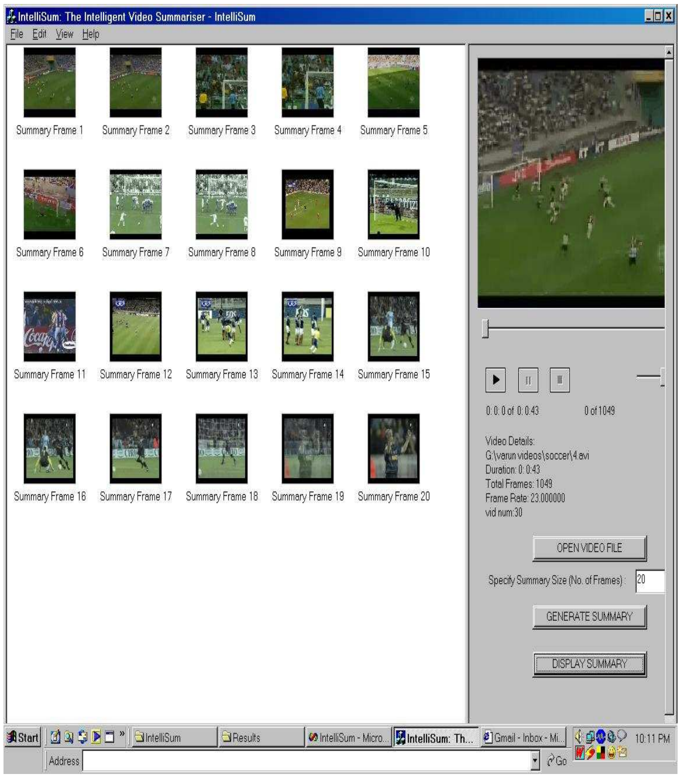


0: 0: 0 of 0: 9:20   0 of 8410

Video Details:
G:\varun videos\79.avi
Duration: 0: 9:20
Total Frames: 8410
Frame Rate: 15.000000
vid num:56

Specify Summary Size (No. of Frames) : 20

Fig. 3. GUI interface showing summary results for a home-shot party Video



0: 0: 0 of 0: 0:43   0 of 1049

Video Details:
G:\varun videos\soccer\4.avi
Duration: 0: 0:43
Total Frames: 1049
Frame Rate: 23.000000
vid num:30

Specify Summary Size (No. of Frames) : 20

Fig. 4. GUI interface showing summary results for a soccer Video

features along with visual features. It is observed that there is an improvement in the results when we consider aural features along with the visual features. This improvement can be well seen from the subjective verification. The GUI interface showing results of 20-frame summaries for home-shot party videos as well as for soccer videos using combined aural and visual features have been given in Figs 3 and 4 respectively.

Summary verification is an extremely subjective task and the quality of result will vary widely from subject to subject. We have used subjective precision and recall measures to asses the summary quality subjectively. To validate our results, we have performed ten-fold cross validation, five-fold cross validation and two-fold cross validations. The Subjective precision-Recall graphs of ten-fold cross validation for both homeparty and soccer video data set for the two cases i.e with and without using aural features are shown in Figs 5 and 6 respectively.

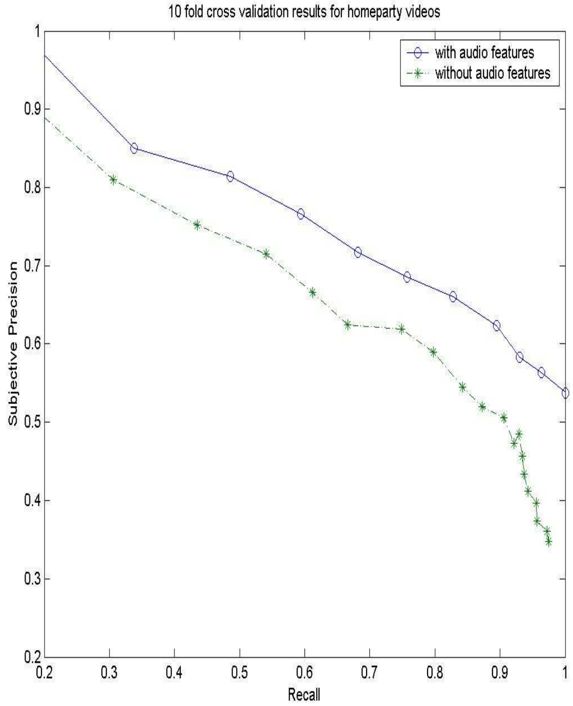


Fig. 5. Subjective assesment of summary results for home-party video

V. Conclusion

We have presented a Video Summarization system for summarizing videos known to belong to a class of videos. We have used semi-supervised learning to capture the high level semantics associated with a set of videos and to map them to low level features. Though, in past semi-supervised learning has been used for a number of applications like Face recognition, Gesture Recognition, Medical systems e.t.c., but this work is the first attempt at using a semi-supervised

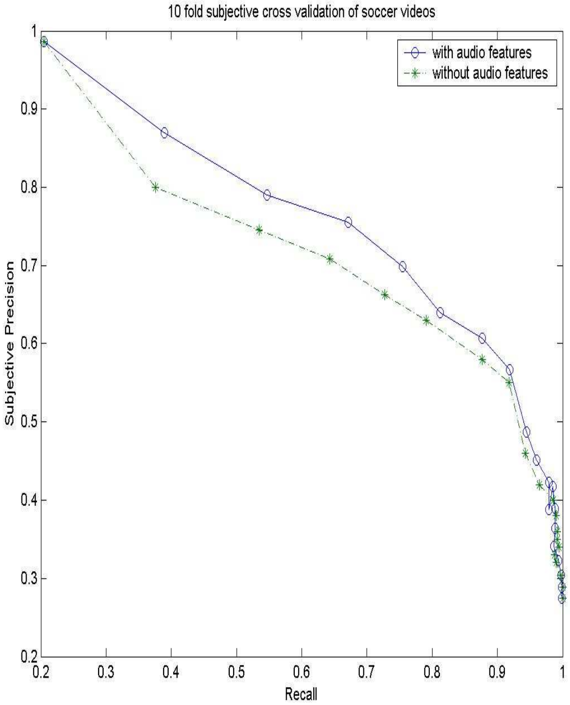


Fig. 6. Subjective assesment of summary results for a soccer Video

learning algorithm for the purpose of video summarization.

Acknowledgment

This work has been supported by the grant for DST project E-Video: Video Information Processing with Enhanced Functionality.

References

[1] Y. Taniguchi,A. Akustu,Y. Tonomura and H. Hamada, *An intuitive and efficicent access interface to real-time incoming video based on automatic indexing*, in proceedings of the third ACM international conference on multimedia, 1995, pp. 25-33.

[2] M. Lee, W. Chen, C. Lin,C. Gu and T .Marloc, *A Layered video object coding system using sprite and affine motion model*, IEEE transactions on circuits and systems for video technology, vol 1,pp. 130-145, 1997.

[3] Y. Rui,T.S. Huang and S. Mehrotra, *Constructing table of content for videos*, ACM multimedia systems jouranal, Special issue multimedia systems on video libraries, vol 7,no.5,pp. 359-368, 1990.

[4] R. Leinhart, S. Pfeiffer, and W. Effelsberg, *Video abstracting*, Communication of the ACM, pp.55-62, December 1997.

[5] M. Smith, and T. Kanade, *Video skimming and characterization through combination of image and language understanding techniques*, Proceedings of Computer Vision and Pattern Recognition, CVPR, 1997.

[6] Y.F. Ma, L. Lu, H.J. Zhang, and M.J. Li, *A user attention model for video summarization*, In Proceedings of ACM Multimedia, pp. 533-542, 2002.

[7] H. Sundaram and S.F. Chang, *Constrained Utility Maximization for Generating Visual Skims*, IEEE Workshop on Content-Based Access of Image and Video Library, 2001.

[8] Ajay Divakaran, *Video summarization using motion and audio descriptors*, Mitsubishi Electric Research Laboratories, 2003.

[9] Vyacheslav Parshin and Liming Chen, *Video Summarization Based on User-Defined Constraints and Preferences*, in proceedings of RIAO, 2004.

[10] C.W. Ngo, Y.F. Ma and H.J. Zhang, *Video Summarization and Scene Detection by Graph Modeling*, in IEEE Trans on Circuits and Systems for Video Technology, 2005.