PROBABILISTIC MULTIMEDIA OBJECTS (MULTIJECTS): A NOVEL APPROACH TO VIDEO INDEXING AND RETRIEVAL IN MULTIMEDIA SYSTEMS.

M. R. Naphade, T. Kristjansson, B. Frey and T. S. Huang

Beckman Institute for Advanced Science and Technology
University of Illinois at Urbana-Champaign
milind@ifp.uiuc.edu, kristjan@uiuc.edu, bfrey@uiuc.edu, hunag@where.csl.uiuc.edu

ABSTRACT

This paper proposes a novel scheme for bridging the gap between low level media features and high level semantics using a probabilistic framework. We propose a framework, in which scenes can be indexed at a semantic level. The fundamental components of the framework are sites, objects and events. Detection of presence of an instance of one of these influences the probability of the presence of instances within other classes. Detection of instances is done using probabilistic multimedia objects: Multijects. Indexing using Multijects can handle queries posed at semantic level. Multijects are built in a Markovian framework. Two ways of building the Multijects from low level features fusing features from multiple modalities are presented. A probabilistic framework is also envisioned to encode the higher level relationship between Multijects, which enhances or reduces the probabilities of concurrent existence of various Multijects. An actual implementation is presented by developing Multijects representing higher level concept of "Explosion" and "Waterfall". The models are evaluated by using the Multijects to detect explosions and waterfalls in movies. Results reveal, that the Multijects detect the aforementioned events with greater accuracy and are able to segment the video into scenes which have explosions and waterfalls.

1. INTRODUCTION

Multimedia Indexing and Retrieval presents a challenging task of developing algorithms that fuse information from multiple media to support queries. The state of the art in content based video indexing and retrieval techniques includes [1], [2] , [3]. While [3] is oriented towards effective browsing support [1] uses various features like color texture, shape and motion for video indexing and retrieval. There has been work in indexing for specific domains [4],[5], limiting the scenario to broadcast news, basketball videos etc. There are three approaches to the task of Video Indexing and Retrieval. In the first, tools can be developed for facilitating browsing. In the second, queries based on example are supported. The third broad category is to provide tools for queries based on high level semantics without providing the system with examples. In the first scenario, the focus is on detecting shot boundaries, structure the video, cluster shots into story units and present the user with hierarchical video along with useful summaries [3]. The focus is to let the user have a hierarchical picture about what goes on in the video and leave it to the user to search based on this structure. In the second scenario, algorithms focus on defining a similarity measure for comparing a query video with a database of target videos [6].

The third and final approach presents the most challenging problems of defining a system which will bridge the gap between low level features and high level semantics. This paper introduces early ideas, which fall in the third category of domain independent video indexing by providing a framework which supports indexing and thereby retrieval at a semantic level. An additional issue is the use of multiple media for indexing and retrieval. Recent work [7] uses video as well as audio in an attempt to derive the structure of videos. Recent work also proposes the use of closed caption and video.In this paper we present a novel framework for fusing multiple media i.e. video and audio for indexing.

While each of the above three approaches are complementary, there has been much work on structuring and browsing. Video Retrieval based on similarity goes a step further towards retrieval. However it is not always feasible to search for the necessary clip by browsing and examples may not be available that can be used to search based on similarity. This therefore highlights the need for our approach which allows the user to search a video database without browsing or querying by example. The first step towards this is to provide the user with a set of objects which are searchable. The onus is then on the indexing engine to

This research was supported by the NCSA ISAAC project and a grant from the Arnold and Mabel A. Beckman foundation. Milind R. Naphande and Trausti Kristjansson were supported by the NCSA ISAAC project; Trausti Kristjansson is a Fulbright Scholar; Brendan Frey is a Beckman Fellow.

process the videos in the database and segment them using the objects in a menu. The user can then query on any object, event or site from the menu. This is the idea central to our work presented in this paper. We propose a novel approach to multimedia indexing. We combine low level features from audio and video streams and develop multimedia objects or Multijects.

Multijects are probabilistic objects, which map low level features to high level semantics. Indexing and retrieval is based on these Multijects. We illustrate this approach by developing a Multijects for the events "Explosion" and "Waterfall". The Multijects are then used for indexing all videos in the database and segmenting sections of videos which contain explosions and waterfalls. We develop Multijects and compare their performance with single medium event detectors. It can be seen from the results that the Multijects outperform single medium based detectors by a considerable margin.

The paper is organized as follows. In section 1 we present the Introduction. In section 2 we review existing video indexing techniques. In section 3 we present a novel approach to indexing and retrieval which is based on our concept of the Multijects. We also explain the method of building Multijects. In section 4 we present a scheme which will integrate Multijects and support semantic queries. In section 5 we present results of our current implementation. In section 6 we present conclusion and directions for future research.

2. REVIEW OF EXISTING VIDEO INDEXING TECHNIQUES

This section looks into the state of the art in domain independent video indexing techniques. As mentioned earlier all the current system fall in one of two broad categories.

The first category involves the structuring of video for efficient browsing and summaries. A central step in most of these approaches is to segment the video in terms of basic units called shots.A state of the art high performance shot boundary detection algorithm can be found in Naphade et al. [8]. After segmenting the video into shots, shots of similar content are grouped together based on a set of rules to form story units. Cluster of shots are categorized as dialogs etc. The table of contents is then built based on this hierarchical structure enabling efficient video browsing. To facilitate browsing, summaries are also provided. Most techniques for generating summaries are based on selection of atypical frames in videos as representative frames for shots.

The second category involves video retrieval based on similarity between a query video and a target video. This query by example approach can be seen in [1], [2] and [6]. The video shot is characterized by a set of features including color, texture, shape and motion information. Region segmentation and tracking is used to enhance the performance. A database is then created of these regions with homogeneous characteristics. Queries are supported by example when the user presents to the system a video which is processed. The extracted features from the query are matched with the ones in the database. A variation of this approach is to compare the objects only found in key frames [3]. Compressed domain processing of video is popular in both categories. While all the algorithms in the above mentioned categories bring the user closer to what he is searching for, they place restrictions. In the efficient browsing paradigm, the user is unable to query the database. In the video retrieval by similarity approach it is very difficult sometimes to find an example which can be provided to the system, without which the database cannot be searched. This therefore limits the widespread use of video retrieval engines based on similarity. Another concern is that similarity is hard to define and quantify.

Neither of these approaches satisfies plain text based query where a user wants to find something like a "Car" or a "Building". The reason is, that there is a gap between low level image and video features like color, texture, shape, motion etc. and high level concepts like "Car". While it is difficult to bridge this gap for every high level concept, multimedia processing under a probabilistic framework facilitates, bridging this gap for a number of useful concepts.

3. MULTIJECTS: PROBABILISTIC MULTIMEDIA OBJECTS

Having acknowledged the need for high level indexing for retrieval, we now present a novel approach for the same. The central theme to this approach is the concept of Multijects or Multimedia Objects, which we define as follows.

A Multiject has a semantic label and summarizes a time sequence of low level features of multiple modalities in the form of a probability. Examples of Multijects are "Person diving into water", "Person skiing", "Bird flying", "Aeroplane taking off", "Rocket Launch", "Boat sailing", "Explosion", "Waterfall", "Beach", "Indoor", "Gunshot", "Outdoor", "Sunset", "Desert", "Snow Clad Mountains" etc. A Multiject has three main aspects. The first aspect is the semantic label, such as the ones in the list above. Later on in this paper, we implement a Multiject for the event "Explosion". Indexing is thus done, done by noting the segments of videos where such a Multiject occurs. During retrieval, all the video shots which are labeled with "Explosion", can be viewed.

The second aspect of a Multiject is, that it summarizes a time sequence of low level features. In this first system, this scope of summarization is one shot. Because the multiject can model an event in the media

streams, and is associated with a label, the Multiject maps the low level features to this high level semantic label.

The detection of a certain Multiject can increase or decrease the probability of occurrence of other Multiject. For example if the Multiject "Beach" is detected with a very high probability, then the probability of occurrence of the Multiject "Yacht" or the Multiject "Sunset" increases. This is the third aspect of Multijects, i.e. their interaction in a network.

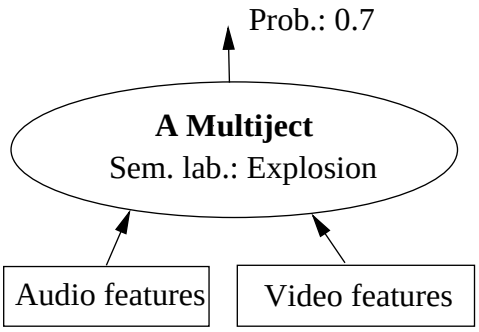

Figure 1: A Multiject

3.1. Multinets

We propose the *multinet* (multiject network) as a way to represent higher-level probabilistic dependencies between multijects. Fig. 2 shows an example of a multinet. In general, all multijects derive probabilistic support from the observed multimedia data (directed edges), as described in the previous section. The multijects are further interconnected to form a graphical probability model associating a real valued weight with each undirected edge in the graph. The weight indicates to what degree two multijects are correlated *a priori* (before the data is observed). In Fig. 2, plus signs indicate the multijects are correlated *a priori* whereas minus signs indicate the multijects are anticorrelated.

A plus sign on the connection between the bird multiject and the waterfall multiject indicates that the two multijects are somewhat likely to be present simultaneously. The minus sign on the connection between the bird multiject and the underwater multiject indicates that the two multijects are unlikely to be present simultaneously. The graphical formulation highlights interesting second-order effects. For example, an active waterfall multiject supports the underwater multiject, but these two multijects have opposite effects on the bird multiject. After obtaining robust multiject models, we will use labeled multimedia data to estimate the weights of the multinet using the generalized expectation maximization algorithm [9],[10]

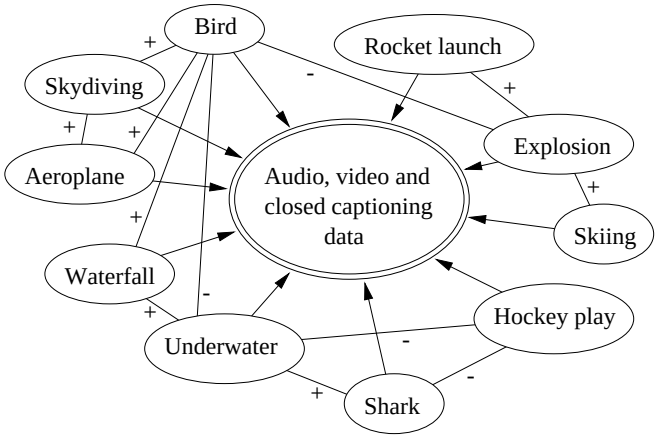

Figure 2: A Multinet: Multijects are combined in a network that captures the co-occurrence probabilities of multijects.

3.2. Building a Multiject

Having described various aspects of a Multiject, we now propose a scheme to build Multijects using low level features and a graphical probabilistic framework. We propose modeling each modality in a Multiject with a hidden Markov model [11] or a gaussian mixture model. For video we use the histogram and its variations to generate our feature set. The features used in video include the three channel linearized color histogram. Histograms are obtained in the HSV color space since it is perceptually closer to human vision discrimination for color objects. To trap information about the gradient, the difference in the three channel histogram is also used as a feature set. This is done by obtaining the difference between the histograms of successive image frames. The video feature set is 48 feature long with 24 features from the three channels with eight features from each channel and 24 features from the difference of the histograms of successive frames. This 48 features comprise the video feature vector. The audio features were calculated using a 20 bin filter bank with evenly spaced bins in the range from 0 to 22kHz. Features were computed at 20 millisecond intervals. Having obtained these two feature vectors for the two media, we propose two approaches to combine the features from multiple modalities to come up with the Multiject.

3.3. First approach Hierarchical Hidden Markov Models (HHMMs)

For our initial experiment, we chose to build Multijects for "Explosion" and "Waterfall". We first independently train a Hidden Markov model based on the feature vectors of the video stream and another based on the feature vectors of the audio stream. The Video and Audio HMMs for the "Explosion" Multiject has 3 states while the Waterfall HMM has only 1 state. The training set data was labeled independently for

the audio and video and the HTK Toolkit was used for training the models. After training the Video HMM and Audio HMMs independently, we obtained a video HMM and an audio HMM. The Viterbi algorithm was then used to find the best possible state sequence given the trained HMMs and the feature vectors for both the audio and the video. The optimal state sequences found by the Viterbi algorithm in the video and audio are then used as the input to a supervisor HMM. The supervisor HMM fuses the modalities. Its observations are the states of the media HMMs i.e. the Audio and Video HMMs. Since the observation rate of the video and audio are not the same, the state of the media HMMs are sampled at a fixed rate to produce the observation sequence for the supervisor HMM. The supervisor HMM was then trained and after being trained, it is this supervisor HMM along with the media HMMs which emits the probability of the occurrence of a Multiject. While the labeling of the audio is done without reference to the video and similarly for the video, the labeling for the supervisor HMM was done by taking into account the multi-modality and by observing the beginning and end of the occurrence of a given Multijects in the joint audio-visual data.

The supervisor HMM encodes the correlation of states in the two modalities. Thus this is a fast greedy bottom up algorithm which results in a an accurate approximation of the Multiject occurrence probability.

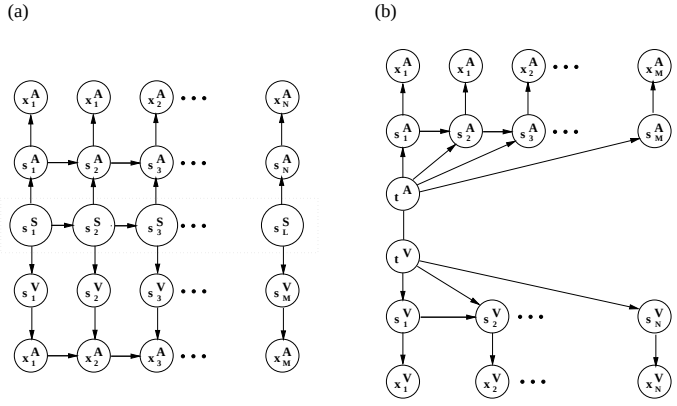

Figure 3: (a) Heirarchical Multiject: The state sequence withing the dotted box represents the supervisor HMM sequence. (b) Event Coupled Hidden Markov Model

3.4. Event-coupled hidden Markov models

Another approach to using the time relations of events in the audio and video streams, that relate to the same physical event, is to model the temporal constraint explicitly. This is the motivation behind Event-coupled hidden Markov Models. In this approach the same media models are used as in the former approach, however, instead of using the states of the media hmms as observations for a supervisor HMM, we model the distribution of the time difference of transition into the first state of the respective models. In other words, we model the time difference of the onset of the events. Figure 3 (b) shows a graphical representation of the ECHMM. In exchange for not modeling coupling that takes place after the onset of each event, we gain the ability to perform exact inference.

A graphical model that describes event-coupled audio and video HMMs is shown in Figure 3 (b). The observation sequence and the state sequence for audio are $\{x_1^A, \ldots, x_M^A\}$ and $\{s_1^A, \ldots, s_M^A\}$ respectively while those for the video are $\{x_1^V, \ldots, x_N^V\}$ and $\{s_1^V, \ldots, s_N^V\}$ Notice that the the two modalities may be sampled at different rates. In order to determine if an event occurs in a video sequence, we calculate the log probability ratio between the event model and an anti-event model. This is similar to what is done in some word spotting schemes. A more detailed discussion of ECHMMs and results of using them can be found in [12].

4. RESULTS

In this section we present early results of the supervisor type Multiject and demonstrate the accuracy of the Multijects over a testing set. Results for the ECHMMs have been reported in [12]

Our dataset consisted of 33 video clips, 19 of them containing explosion of different kinds and 14 containing waterfalls. There are in all 52 explosions containing over 6745 frames and 27 waterfalls containing over 7704 frames. The total dataset consists of several thousand frames with and without explosions and waterfalls. Some sequences were not usable for training the audio HMMs (because of music and other editing effects) and vice versa. The training and testing sets for the supervisor HMM was an intersection of the usable audio and video sets. In addition to the above mentioned explosion and waterfall Multijects, we also trained Multijects that modeled non-explosion and non-waterfall segments. These models serve to refine the boundary in the feature spaces between event and non-event.

In order to test the performance of our multijects, we divided our data set into a training set of 16 explosion, and 16 waterfall sequences, and and a testing set of 13 explosion and 9 waterfall sequences. The performance is determined by the temporal overlap of the automatic segmentation over a hand labeled ground truth. Thus the first row in Table 1 shows that 94.38% of the automatic labeling in segments hand labeled as explosion, were correctly classified, 0.11% were classified as waterfalls, and so on.

From the results we can see the benefit of combining multiple media, and of the Multiject based approach of video indexing and retrieval.

The inferior results of the single media stem partly

	explosion	waterfall	anti-explosion	anti-waterfall
explosion	94.38	0.11	5.52	0.0
waterfall	0.0	86.12	0.0	13.88

Table 1: Confusion matrix for waterfall and explosion for Multiject.

	explosion	waterfall	anti-explosion	anti-waterfall
explosion	84.58	0.0	15.42	0.0
waterfall	0.0	75.48	0.0	24.52

Table 2: Confusion matrix for waterfall and explosion for video.

from the fact that the ground truth for the combined modalities is not exactly the same as for the individual media. This is especially true for audio, where, for example explosion rumble can be heared after the initial visual flash. However, it is the combined ground truth that is important. This in fact highlights the need for a multiple media based indexing approach.

5. CONCLUSIONS AND FUTURE RESEARCH

In this paper we introduce a novel approach to video indexing and retrieval using probabilistic multimedia objects or Multijects. Multijects have been developed for the events of "Explosion" And "Waterfall" using global features in audio and video. We propose two new algorithms for fusing multiple modalities and observe that both outperform individual modalities based detectors. A set of such multijects can be developed and a menu can be provided to the user from which any of the object or event or site can be searched on all the database videos. We have successfully demonstrated their power in semantic video indexing and retrieval where the onus is placed on the indexing end and retrieval can be fast. We also have presented the concept of a network of multijects called a multinet and the power of multijects and multinets for automatic semantic indexing. We hope to increase the effectiveness of this approach by developing more multijects and also a multinet. Towards this end we plan to incorporate more features in our feature vector, include region segmentation to enhance the performance of multiject creation process for more difficult multijects. We plan to develop a complete framework for video indexing and retrieval based on the concept of multijects and multinets and our early results are very encouraging. By doing this we bridge the gap between low level features and high level semantics and present a dynamic and powerful tool for video indexing.

	explosion	waterfall	anti-explosion	anti-waterfall
explosion	55.54	0.07	44.06	0.34
waterfall	0.0	86.12	41.67	11.98

Table 3: Confusion matrix for waterfall and explosion for audio.

6. REFERENCES

[1] D. Zhong, S. F. Chang, "Spatio-Temporal Video Search using the Object-Based Video Representation", *IEEE Intl. Conf. on Image Processing*, Vol 1, pp. 21-24, Oct 1997, Santa Barbara, CA.

[2] Y. Deng and B. S. Manjunath "Content Based Search of Video using Color, Texture and Motion", *IEEE Intl. Conf. on Image Processing* , Vol 2, pp. 13-16, Oct 1997, Santa Barbara ,CA.

[3] H. Zhang, J. A. Wang and Y. Altunbasak, "Content-Based Video Retrieval and Compression: A Unified Solution", *IEEE Intl. Conf. on Image Processing*, Vol 1 pp. 13-16, Oct 1997, Santa Barbara ,CA.

[4] D. D. Saur, Y. P. Tan, S.R. Kulkarni, P. J. Ramadge, "Automated analysis and annotation of basketball video", *Proc. of SPIE,* Vol. 3022, pp 176-187, 1997.

[5] M. G. Brown, J. T. Foote, G. Jones K. Jones S. Young, "Automatic Content-Based Retrieval of Broadcast News" *ACM Multimedia 95*, San Francisco, U.S.A 1995.

[6] M. M. Yeung and B. Liu, "Efficient matching and clustering of video shots", *International Conference on Image Processing 1995* Vol. 1 pp. 338-341

[7] J. Nam, A.E. Cetin, A.H. Tewfik "Speaker Identification and Video Analysis for Heirarchical Video Shot Classification" *ICIP 97* Vol. 2, pp. 550-555,

[8] M. R. Naphade, R. Mehrotra, A. M. Ferman, J. Warnick, T. S. Huang, A. M. Tekalp, "A High performance Shot Boundary Detection Algorithm using multiple cues" *to be presented at the International Conference on Image Processing, 1998*

[9] P. Dayan, G. E. Hinton, R. M. Neal and R. S. Zemel 1995. The Helmholtz machine. *Neural Computation* **7**, 889–904.

[10] B. J. Frey 1998. "Graphical Models for Machine Learning and Digital Communication", *MIT Press*, Cambridge, MA.

[11] L. R. Rabiner, "A Tutorial on Hidden markov Models and Selected Applications in Speech Recognition", *Proceedings of the IEEE* Vol. 77 No. 2 Feb 1989

[12] T. Kristjansson, B. Frey and T.S.Huang, "Event-coupled hidden Markov models", *submitted to the Internatinal Conference of Neural Information Processing Sytsems, NIPS'98*, http://ifp.uiuc.edu/ trausti/papers/echmms_nips98.ps