International Journal Of Computer Science And Applications Vol. 1, No. 2, August 2008
ISSN 0974-1003

Intelligent Agent for Retrieving Video Data

B V Patel
Computer Technology Department
Shah and Anchor Kutchhi Polytechnic
Chembur, Mumbai, India
patelbv@acm.org

B. B. Meshram
Department Of Computer Technology
Veermata Jijabai Technological Institute
Matunga, Mumbai, India
bbmeshram@vjti.org.in

ABSTRACT

In this paper, we present the framework for retrieving video based on the audio content of the video. We have done the extensive literature survey for the comparative study of various video retrieval algorithms. We also describe the proposed video retrieval algorithm, its performance analysis and browsing for multimedia digital libraries. The proposed goal based agent will index multimedia file so that the users should be able to search, access, examine and navigate among video as effectively as they can.

Categories and Subject Descriptors

I.2.M [**Artificial Intelligence**]: Miscellaneous.

General Terms

Algorithms, Performance, Design, Human Factors, Experimentation.

Keywords

Speech Processing, Audio Indexing, Video Databases.

1 INTRODUCTION

In recent decades, multimedia retrieval has attracted plenty of attention due to its data's rich content. Among media types, video presents the most complex data, including a sequence of frames (or feature vectors), audio, motion, etc. With ever more heavy usage of video devices and advances in video processing technologies, the amount of video data has grown rapidly and enormously for various usages, such as advertising, news video broadcasting, personal video archive, medical video data, and so on. Interestingly, the popularity of WWW enables enormous video data to be published and shared. Web search engines provide users convenient ways for indexing videos of their interests. Due to the high complexity of video data, retrieving the similar video content with respect to a user's query from a large database requires: (a) effective and compact video representations, (b) efficient similarity measurement, and (c) efficient indexing on the compact representations. This gives rise to the problem of combining various streams of information coherently for various tasks like search, organization for better browsing, which is quite challenging as there is a large amount of data exits on the web which contain audio, visual and text information together.

There are various methodologies exists today to retrieve the multimedia document based on various features like image search and text search which are the most common feature that are used for searching.

This paper is organized as follows:

In second section, we describe the related work. The third section deals with proposed algorithm. Section four describes the proposed method. Fifth section presents the experiment results. Finally we conclude in section six.

2 RELATED WORK

A huge amount of literature exists on combing visual and the text steam together for the task of video understanding and search. Tamara Berg *et al* [8] show how to construct a face dataset from a collection of automatically gathered news video/image and captions. The general task of attaching keywords to images itself has received considerable attention in [12, 13, 15]. These methods use variations of multiple instances learning which is a way to build classifiers from bags of labeled examples. Belongie *et al* [9] demonstrate examples of joint image-keyword searches. Barnard and Johnson [7] show one can disambiguate the scenses of annotating words using the images. The central theme of all these works is that text and images contain complimentary information and one can combine them together to solve problems which are otherwise hard to solve by themselves. In the domain of video, Aranjelovic and Zisserman [6], show how to do automatic face recognition in feature length films. Their system allows one to search for all occurrences of frontal faces in the movie given a small set of query images. Extending the detection to profile and three-quarter views, K.Mikolajczyk *et al* [14] give a temporal approach to reliably detect frontal and profile faces in a video. They use zero order dynamic model for appearance variation and use condensation filter to accumulate probabilities of face detection over time. The Condensation algorithm (Conditional Density Propagation) proposed by Izard and Blake [11] allows quite general representations of probability and the use of non-linear motion models more complex than those commonly used in Kalman Filters. Mark Everingham *et al* [10] show how to automatically name the faces in the video using the transcripts.

In the speech recognition community a number of researchers have examined a variety of ways to handle quickly transcribed data. Anand *et al* [28], describes an efficient repair procedure for quickly taken down transcripts. The focus was to compute word

level alignments of audio segments which can then be used as training data. The step however required manual alignment of the audio files to a set of transcripts which can be time consuming. The proposed method does not require this step. Yet another way to compute alignments between the text and the video would be to run a speech recognizer to obtain the speech which can then be used to align it with the transcript. However conventional ASR for large vocabulary is slow.

In Text Based Retrieval, user inputs the text of the image documents which are generally saved as a property of the video like video name, video date, owner, etc. In this retrieval, it is very much likely that the user must give the exact name of the video for which he is looking for. Results obtained from this are unsatisfactory. It is likely that if user wanted to search the video of computer may get the building of the company manufacturing the computers. Our Proposed method overcomes the problem of having knowledge of video property feature.

The Image Based Retrieval method requires sample image that user wants to search in the video or multimedia data. Here the video is divided in number of frames and then image comparison is performed. Problem with this method is user requires the image of the object he is looking in the video or database. Our method compares the audio content of the video, making easier for the users who do not have the image with them for the comparison.

Normally video or multimedia files on disk occupy high memory space which is the requirement of the above methods. User using generalized methods may not able to locate part of the video where his information is available. As the user is interested in his information and may not want to look only at the image or the whole movie. We present the novel approach to overcome this problem. In proposed system user will able to look at exact information from the video i.e. at what time or track given information exists in the multimedia file. It is also possible to search from the digital, multimedia databases rather then only from one video and present the user required information. Proposed method also reduces the working memory space requirement comparing to above traditional methods

3 ALGORITHMS

The following algorithms are used in the proposed method.

3.1 HMM

Every video or multimedia file contains the audio part. This audio can be processed and converted into text.

Audio Extraction:

Hidden Markov Model algorithm is used to recognition of the speech from the audio content of the video.

Following recognition formula determines the more likely sequence W*.

$$W^* = Arg\,Max_{W \in \mathcal{W}, S \in \mathcal{S}} \{P(W=W1|\Theta) \prod_{t=2,\ldots,T} P(W_t|W_{t-1},\Theta)\, P(S_1|W_1,\Theta)$$

$$\prod_{t=2,\ldots,T} P(W_t|W_{t-1},\Theta)\, P(S_1|W_1,\Theta) \prod_{t=1,\ldots,T} P(y_t|S_t,\Theta)\}$$

Where y_t is the acoustic observation depending on current state s_t.

W* should be found by trying all the possible sequence of the states and words. An exhaustive search is unfeasible for any realistic recognition problem. The computation cost is drastically reduced by resorting Viterbi-like algorithms.

Feature extraction of the HMM is performed by filtering the speech signal with the first order FIR filter whose transfer function in the z-domain is

$H(z) = 1-a.z^{-1} \quad 0 \le a \le 1$

A typical value of a is 0.95, which gives rise to a more than 20 dB amplification of the high frequency spectrum.

To extract sufficient statistics Y are from the speech samples X by simply applying the function Y=g(X). This is done using finding the DFT, the derivatives of the energy of the speech signals with the help of filter banks.

Spectral analysis reveals those speech signal features which are mainly due to shape of the vocal tract. Spectral features of the vocal tract are generally obtained as the exit of the filter banks, which properly integrate a spectrum at defined frequency range. A set of 24 band-pass filters is used as it simulates human ear processing. A computationally in-expansive method is to implement the filtering directly in the DFT domain.

The training HMM aims to maximize the likelihood function of the observation sequence $P(Y_T|\Theta)$ in a given HMM. First we initialize the model parameters with random values. Re-estimation of the model parameter is performed on the initial values and training observation sequence. This is repeated till $P(Y_T|\Theta)$ does not experience any improvements.

The speech utterance $Y_T = (y_1, y_2, \ldots, y_T)$ where y_t is the Y_T belongs to a process produced by a further underlying stochastic process characterized by a finite number N of states. States are not known. The probability of the utterance Y_T is

$$P(Y_T|A,\pi,B) = \sum_{S_T} \pi_{S1} \prod_{t=2}^{T} a_{st-1s} b_s(y_1)$$

Where $b_s(y_t)$ denotes the probability that y_t is produced when the state is s_t at time t. A is the set of Input. B is the set of output. The recognition process can be performed by determining the parameter set $\Theta = (\pi, A, B)$ such that above the probability of Y_t is maximized.

3.2 Filter

for I=1 to all words in index file do following

If $Word_I$ = $IndexWord_I$ then delete $word_I$ from the Index file.

3.3 Formation and updating of video database index.

for I=1 to all words in index file do following

Locate the location L of $Word_L$ in the file.

Insert the $Word_I$ at Location L in the Database Index file with meta data information of the $Word_L$

3.4 Meta data retrieval from the video database index algorithm.

Using binary search locate the word user is searching in video data base index file.

If word is found extract related Meta data information.

Now apply the linear search and locate all the video where the word appears and extract related Meta data information of the same.

3.5 Time alignment algorithm.

Time alignment is done using following formula.

$P_f = P_i + P_t$

P_f is forwarding time of video track.

P_i is initial video track time generally 0.

P_t is the time where the retrieved word appears in the video obtained from the index file.

4 PROPOSED METHOD

Architecture of the proposed goal based intelligent system:

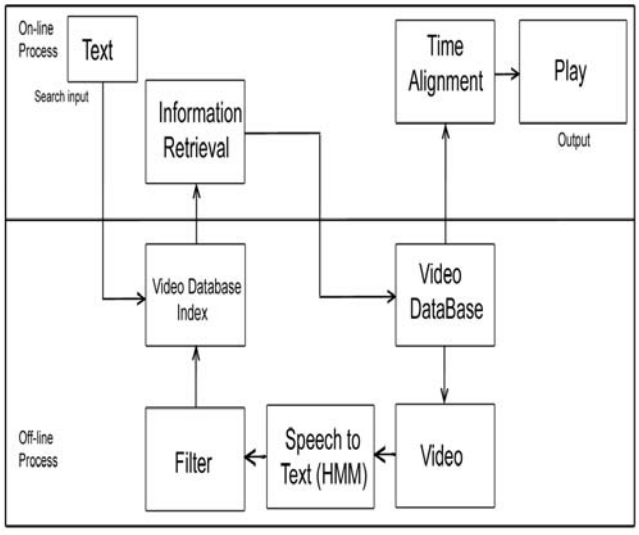

Figure 1. Architecture of proposed goal based intelligent system

4.1 Off-line process

HMM coverts the given video from video database to text. Additional information of the spoken word (time and name of video) is extracted and stored in text format which is then filtered. Video database index is updated with this text information.

Aim: To retrieve video from multimedia database.

Input: Video from video database.

Step 1. Store the video in the video database.

Step 2. Select the video required to be indexed.

Step 3. Audio from video is extracted and converted to using the HMM. Also record additional information like when the word is spoken.

Step 4. Filter the text generated from step 3.

Step 5. Filtered output is indexed and stored in the video database index.

4.2 On-line process

Input the text (audio word from video) which is to be searched in video database index file. Retrieve the mete data available in the video database index file and with the help of this locate the video from the video database and perform the time alignment and display the results and play the video based on user selection time.

Input: Text (This text is related to the audio of video)

Step 1. Search the text from video database index.

Step 2. Retrieve the metadata about video i.e. time, video name etc.

Step 3. With the help of this metadata information pickup video from the video database.

Step 4. Perform the time alignment of video.

Step 5. Play the video from starting time or from the spoken word (text) from the video.

User will be able to navigate through these files by above information and only the files which are selected by the user will have to be brought to the user instead of the whole database. This improves searching as we are searching only in the index file instead of whole database compared to other methods discussed above. This helps user in getting quick the improved and required result.

5 EXPERIMENT RESULTS

The proposed algorithms are implemented and tested in VB using the oracle as backend. The preliminary results are encouraging when tested on various topics of video data with run time of around 300 hours of video. The performance is shown below as a graph in figure 2.

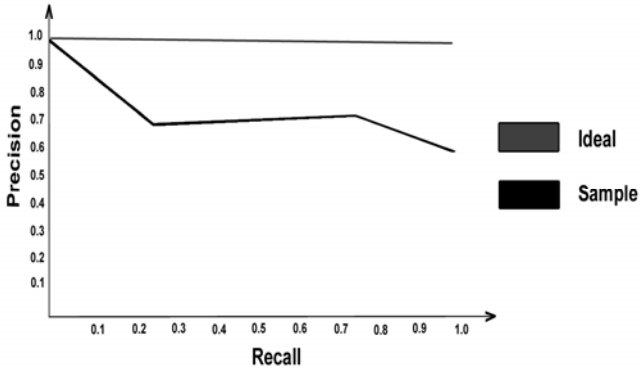

Figure 2. Performance of proposed system

Performance is measured based on the retrieved word and the total number of words in the database using following parameter.

Recall refers to the words relevant retrieved words and total numbers of words in database.

Precision refers to number of relevant and retrieved words and total number of retrieved words.

$$\text{Recall} = \frac{\text{No. of words retrieved and relevant}}{\text{Total no. of relevant words in database}}$$

$$\text{Precision} = \frac{\text{No. of words retrieved and relevant}}{\text{Total no. of retrieved words}}$$

6 CONCLUSION

The proposed method may work as an addition improvement to present video search methods to locate required video content from the web with user satisfactory results. We have presented novel method of indexing and retrieval of video/multimedia databases. Proposed method improves the present search, index and retrieval of video database. Initial experimental results are very much encouraging and can further improved with the combination of video summarization and automatic annotation of video/multimedia data.

7 REFERENCES

[1] B. V. Patel, B. B. Meshram 2007, Retrieving and summarizing images from PDF documents, In the proceedings of Int. conf. on Soft computing and intelligent systems, India-2007.

[2] B. V. Patel, B. B. Meshram 2006. Mining and clustering images to improve image search engines for geo-informatics database, In the proceedings of National Conference on Geo-informatics, India 2006.

[3] B. V. Patel, B. B. Meshram 2008. Intelligent agent for musical instruments Recognization. In the proceedings of Int. conf on Emerging Technologies and Applications in Engineering, Technology and Sciences, India 2008.

[4] B. V. Patel, B. B. Meshram 2008. Content based image retrieval, In the proceedings of National Conference on Emerging trends in information technology, India 2008.

[5] Subhransu, Razena, 2007. Fast unsupervised Alognment of video and text for indexing/names and faces.

[6] O. Arandjelovic and A. Zisserman. Automatic face recognisation for film character retrieval in feature-length films. CVPR2005.

[7] K. Barnard, M. Johnson, D. Forsyth. Word sense disambiguation with pictures. 2003. Workshop on learning word from non-linguistic data. USA.

[8] T.L. Berg, A. C. Berg, J. Edward, M. Maire, R. White, Y. W. The, E. Learned-Miller, D. A. Forsyth 2004. Names and Faces in the news, In Computer Vision and Pattern Recognisation 2004.

[9] C. Carson, S. Belongie, H. Greenspan, J. Malik. 2002. Image segmentation using expectation-maximum and its applicationto image querying. IEEE Trans. Pattern Analysis and Machine Intelligence 2002.

[10] M. Everingham, J. Sivic, A. Zisserman. 2006 Hello! My name is … buffy – automatic naming of characters in tv video. In proceedings of the British Machine Vision Conference.

[11] M. Isard and A. Blake 1998. Condensation condition density propagation for visual tracking. Int. J. Computer Vision

[12] O. Maron and T. Lozano 1998 A framework for multiple-instance learning. In the proceedings of International conference of machine learning. CA.

[13] O. Maron, A. L. Ratan 1998, multiple-instance learning for natural scene classification.. In the proceedings of 15th International conference of machine learning. CA.

[14] K. Mikola, R. Choudhury, C., Schmid. 2001. Face detection in a video sequence: A temporal approach. In CVPR 2001.

[15] Q. Zhang, S., Goldman, Em-dd, 2001. An improved multiple-instance learning technique. 2001

[16] A. S. Gordon, Kavita Ganesan 2005. Automatic story capture from conventional speech. ACM. K-CAP 2005, Canada

[17] M.K.S. Khan, Wasfi G. Al-Khatib, M. Moinuddin. 2004. Automatic Classification of speech and music using neural networks., ACM, MMDB 2004 Nov. 2004, USA.

[18] Hala ElAarag, Laura Schindler. 2006. A speech recognition and synthesis tool. ACM SE March, 2006USA.

[19] Malcolm Slaney., Kyogu Lee, 2006. Automatic Chord Recognition from Audio Using a Supervised HMM Trained with Audio-from-Symbolic Data. In Proceedings of AMCMM'06, October 27, 2006, Santa Barbara, California, USA.

[20] Charles L. Thompson Jr. David D. Langan, Michael V. Doran, 2005. Pianist Style: Can it be Measured and Recognized?. In proceedings of 43rd ACM Southeast Conference, March 18-20, 2005, Kennesaw, GA,USA.

[21] B. Schuller, G. Rigoll, M. Lang, 2003. In proceedings of HMM-based music retrieval using stereophonic feature information and framelength adaptation. IEEE International Conference on Multimedia and Expo 2003.

[22] Ana B. Benitez and Shih-Fu Chang 2002. Multimedia Knowledge Integration, Summarization And Evaluation, In Proceedings of Third International Workshop on Multimedia Data Mining MDM/KDD'2002 Edmonton, Alberta, Canada July 23rd 2002.

[23] Donald Byrd 2001. Music Notation Searching and digital libraries. In Proceedings of JCDL, June 2001, Roanoke, Virginia, USA.

[24] Jyh-Shing Roger Jang, Hong-Ru Lee, Ming-Yang Kao 2001. In the proceedings of Content-based Music Retrieval Using Linear Scaling and Branch-and-bound Tree Search, IEEE International Conference on Multimedia and Expo (ICME'01), 2001

[25] Huhns, Singh, Munidar P 1998. Readings in agents (Morgan Kaufman Publishrs, 1998).

[26] Tecuci, Gheorghe Building intelligent agents (Academic Press 1998).

[27] Earge, John M Music, sound and technology (Van Nostrand, Rinhold Companyy, 1990).

[28] A. Venkataraman, A. Stckle, W. Wang, D. Vergyri, V. R. R. Gadde, J. Zheng. 2000 An efficient repir procedure for quick transcriptions. In proceedings of ICSLP, 2000.

[29] Garofolo, J.S. et al.1999. The TREC Spoken Document Retrieval Track: A Success Story. In the Proc. of TREC-8, 1999.

[30] Ohtsuki, K. et al. 2002. Topic Extraction based on Continuous Speech Recognition in Broadcast News Speech. *IEICE Trans.* Vol.E85-D, No.7, pp.1138-1144, 2002.

[31] Hayashi, Y. et al. 2000. Searching Text-rich XML Documents with Relevance Ranking. In the Proc. of SIGIR2000 Workshop on XML and IR, 2000.

[32] Hearst, M.A. 1997. TextTiling: Segmenting Text into Multiparagraph Subtopic Passeges , Computational Linguistics, Vol.23, No.1. pp.33-64, 1997.

[33] Bessho, K. 2001 Text Segmentation Using Word Conceptual Vectors (in Japanese) . Trans. of IPSJ, Vol.42, No.11, pp.2650-2662, 2001.

[34] Taniguchi, Y. et al. 1997. PanoramaExcerpts: Extracting and Packaging Panoramas for Video Browsing", In the Proceedings of the Fifth ACM Multimedia Conference, pp.427-436, 1997.

8 APPENDIX

The results are shown as below. Search for the text. In the figur-3 the text is microsoft. The following video files containing microsoft are displayed.

Figure 3. Search result

Select video file from menu and play it is shown in figure-4

Figure 4. User selection play option for viewing video