Automatic Identification used in Audio-Visual indexing and Analysis

A. Satish Chowdary[1], N.Tirupathi[2], K. Nageswara Rao[3], K. Nagamani[4]

[1]M.Tech CSE, Mother Terissa Inisitute of Science & Technology, A.P., India.
[2]M.Tech CSE, Mother Terissa Inisitute of Science & Technology, A.P., India.
[3]Assoc.Professor & Head, Dept.of CSE, Mother Terissa Inisitute of Science & Technology, A.P., India.
[4]Asst.Professor, Dept.of CSE, Mother Terissa Inisitute of Science & Technology, A.P., India.

Email: satish.alla@gmail.com

Abstract

To locate a video clip in large collections is very important for retrieval applications, especially for digital rights management. We attempt to provide a comprehensive and high-level review of audiovisual features that can be extracted from the standard compressed domains, such as MPEG-1 and MPEG-2. This paper presents a graph transformation and matching approach to identify the occurrence of potentially different ordering or length due to content editing. With a novel batch query algorithm to retrieve similar frames, the mapping relationship between the query and database video is first represented by a bipartite graph. The densely matched parts along the long sequence are then extracted, followed by a filter-and-refine search strategy to prune some irrelevant subsequences. During the filtering stage, Maximum Size Matching is deployed for each sub graph constructed by the query and candidate subsequence to obtain a smaller set of candidates. During the refinement stage, Sub-Maximum Similarity Matching is devised to identify the subsequence with the highest aggregate score from all candidates, according to a robust video similarity model that incorporates visual content, temporal order, and frame alignment information. This new algorithm is based on dynamic programming that fully uses the temporal dimension to measure the similarity between two video sequences. A normalized chromaticity histogram is used as a feature which is illumination invariant. Dynamic programming is applied on shot level to find the optimal nonlinear mapping between video sequences. Two new normalized distance measures are presented for video sequence matching. One measure is based on the normalization of the optimal path found by dynamic programming. The other measure combines both the visual features and the temporal information. The proposed distance measures are suitable for variable-length comparisons.

Keywords – *Audio and Video, Multimedia, Speech Processing, Topic Segmentation, Topic Identification.*

1. Introduction

Newspapers, magazines, radio, television, World Wide Web information is of strategic importance for business and governmental agencies as well as for citizens. The exponential evolution of multimedia makes it difficult to overview the opulence of information, and that's why important pieces of information have to be filtered and processed automatically. Content analysis of video is to extract meaningful information such that efficient classification, indexing, retrieval, and filtering are possible. One crucial step for such tasks is to define similarity / dissimilarity measure between two video sequences.

The common techniques rely on key-frames since classical methods developed in content based image retrieval can be applied on these still-frames. In [1], a fast video signature based on randomized algorithms is proposed to approximate the video similarity defined as the percentage of clusters of similar frames shared between two video sequences. In [2], block-based minimum variances are used to create video hash values. However, temporal information is ignored in both of the above methods. A template-frequency model which makes uses of the temporal dimension is proposed in [3]. Another similarity measure between shots is developed by using dominant color histograms and spatial structure histograms [4]. Nowadays, information is mainly obtained by manually analyzing (reading, listening and watching) large audio and video databases and current broadcast multimedia sources (such as broadcast TV, radio or Internet streams). After having assigned topics to the incoming news and stories, only the items of interest or items regarding a specific request will be selected and further processed. The use of automatic methods for selective dissemination of information would enable such monitoring companies to cover a much larger variety of media sources by working more cost efficiently and providing 24 hours availability.

Recent advances in computer, telecommunications, and consumer electronics industries have brought huge amount of multimedia information to a rapidly growing audience. More and more digital audio and video data are made available over the Internet. Traditional TV broad cast is moving into the digital and interactive era. People are starting to get high-speed network connections via DSL and cable modem. Multimedia content provides rich information to consumers, but also poses challenging problems of management, delivery, access, and retrieval because of its data size and complexity.

In this paper, we present a shot-level video similarity measure based on dynamic programming. Note the temporal information such as shot durations is not affected by frame rate conversion or illumination changes. The proposed method can be used to locate and identify a video sequence in large collections. Unlike the technique in, where a frame-level dynamic programming is used to deal with frame misalignment, our new method uses shot level dynamic programming, where shot sequences are created in an illumination-invariant color space by clustering video frames in independent component analysis (ICA) subspace. In addition, two new normalized distances are introduced to calculate the dissimilarity. Optimal path is found by dynamic programming. The presented new method is robust to histogram processing, and frame rate conversion. The new distance measures are insensitive to the lengths of videos.

2. Background

2.1 Video Copy Detection

Extensive research efforts have been made on extracting and matching content-based signatures to detect copies of videos. Mohan introduced to employ ordinal measure for video sequence matching. Naphade et al. developed an efficient scheme to match video clips using color histogram intersection. Pua et al. proposed a method based on color moment feature to search video copy from a long segmented sequence. Hampapur et al. [10] examined several methods of using a sequence of frame features (ordinal, motion, or color signature) to leverage the characteristic of sequence-to-sequence matching. In their work, query sequence slides frame by frame on database video with a fixed length window. In addition to distortions introduced by different encoding parameters, Kim and Vasudev proposed to use spatiotemporal ordinal signatures of frames to further address display format conversions, such as different aspect ratios (letter-box, pillar-box, or other styles). Since the process of video transformation could give rise to several distortions, techniques circumventing these variations by globe signatures have been considered.

They tend to depict a video globally rather than focusing on its sequential details. Some properties that are likely to be preserved even with these variations (e.g., shot length information) were suggested to be generated as compact signatures and string matching technique could be used to report such a copy. This method is efficient, but has limitations with blurry shot boundaries or very limited number of shots. Moreover, in reality, a query video clip can be just a shot or even a sub shot. However, this method is only applicable for queries which consist of multiple shots.

2.2 Audio Segmentation

Several methods for audio segmentation have been proposed, like Akaike's Information Criterion (AIC) , the Bayesian Information Criterion (BIC) , the Consistent AIC (CAIC) and the Minimum Description Length (MDL). These and other methods have been compared in [5] and it has been shown that with optimal parameters, almost all algorithms perform comparably well.

The audio segmentation algorithm deployed in this work uses the BIC, which was among the best performing methods. The BIC follows the method of Tritschler and Gopinath [7] which will be described briefly. The algorithm takes a window of n audio features x1….xn, arbitrarily places a boundary at i position, resulting in two segments. It then decides whether it is more likely that one single model θ_1 produced the output x1…xn or that two different models θ_{21} and θ_{22} have generated the two segments' output x1..xi and xi+1 ….xn respectively. The decision rule to check if there is a boundary at point i is

$$\Delta BIC_i \overset{!}{<} 0 \qquad \text{with} \tag{1}$$

$$\Delta BIC_i = -\frac{n}{2}\log|\Sigma_w| + \frac{i}{2}\log|\Sigma_f| + \frac{n-i}{2}\log|\Sigma_s| \tag{2}$$

$$+\frac{1}{2}\lambda(d+\frac{d(d+1)}{2})\log n.$$

Σ_w denotes the covariance matrix of all window feature vectors $x_1,\ldots\ldots,x_n$, Σ_f *and* Σ_8 are the covariance matrices of the features of the first and second segment respectively, d is the feature vector dimension. According to theory, the penalty weight λ should equal 1, but practical applications show better results with $\lambda \neq 1$.

If for a point *i* $\Delta BIC_i < 0$*, then also for some points j surrounding i there will be* $\Delta BIC_j < 0$*. The algorithm decides that the boundary is at the point with the lowest* Δ *BIC value.*

To detect all audio segments of a news show, the window is shifted over all feature vectors with varying length n and varying i. After implementing the above described algorithm, it was noticed that sometimes segment boundaries are set too early, roughly one or two syllables before the speaker finishes. Instead of considering the point *i* at which the minimum of Δ *BIC* occurs as a boundary, the middle of those two points where the Δ *BIC* value crosses the 0 line was chosen. This modification improves the segmentation accuracy and reduces the number of boundaries appearing too early.

2.3 Shot Detection

The first step is to segment a video into a shot sequences using a method in our previous work [6], where illumination-invariant chromaticity histograms are used as raw features and an ICA based method is used to convert the 256-dimensional raw feature subspace into a two dimensional feature space, in which a dynamic clustering algorithm is employed to cluster video frames into shots.

2.4 Shot-level Feature Extraction

The normalized chromaticity histogram is selected as a shot-level visual feature. The illumination-invariant normalized chromaticity (*r*, *g*) is defined as: *r*=*R*/(*R*+*G*+*B*), *g*=*G*/(*R*+*G*+*B*). Histograms with 256 bins are generated in the normalized chromaticity color space for each frame of the video. During implementation, only *r* component is used for simplicity. Each shot is represented by a feature vector which is the mean vector of all video frames within the same shot. A shot sequence is then a vector sequence,□ **r**(*i*),*i* □□ 1,□□*NR*□, where **r**(*i*) represents the *i*th shot and *NR* is the total number of shots. Shot lengths (measured in time) are also calculated during feature extraction.

2.5 Normalized Distance Measure

The overall cost *D* can be used to measure the distance or dissimilarity between two video sequences. A desirable property for such a measurement is that the cost *D* should not depend on the lengths of the sequences. Therefore, a proper normalization of the total cost is necessary. For string matching, the problem has been addressed in [9] using *normalized edit distance*. However, it is computationally expensive. In practice, *D/Np* can be used to calculate the distance measure with a certain amount of normalization. Our first new simplified normalization measure *D*1 is defined as:

$$D_1 = D_0 / N_p,$$

where *D*0 denotes the original total cost, i.e., *D*0=*D* min, with *D* min.

For video sequence comparison, normalization of the total cost by the length of the path is essentially related to the number of shots since the length of the path is bounded between max(*NR*,*NT*) and (*NR*+*NT*). Note that one video sequence with more shots does not necessarily imply it is longer than the other. However, in terms of video similarity measure, people are often interested in how long the two video sequences “overlap” instead of how many shots (or key-frames) are similar. In another word, if we have two pairs of dissimilar shots, it is reasonable to penalize the longer sequences more, compared with the other pair with relatively shorter durations. Therefore, the *normalized edit distance* proposed in cannot be directly applied here since it is only penalizes the lengths of sequences without considering the duration of each symbol. We present the second new distance measure to integrate both visual features and shot durations for video sequence comparison as follows:

$$D_2 = \frac{\sum_{i=1}^{N_P} \left[d(\mathbf{r}(p(i)), \mathbf{t}(q(i))) \cdot \left| L_R(p(i)) - L_T(q(i)) \right| \right]}{\sum_{i=1}^{N_P} \left[L_R(p(i)) + L_T(q(i)) \right]},$$

where *LR*(*n*) is the duration for *n*-th shot in *R* and *LT*(*n*) is the duration for *n*-th shot in *T*. It is easy to show that *D*2 has an upper bound as follows:

$$D_2 \le \frac{\sum_{i=1}^{N_P} \max \cdot (L_R(p(i)), L_T(q(i)))}{\sum_{i=1}^{N_P} (L_R(p(i)) + L_T(q(i)))}.$$

This new distance measure *D*2 combines both visual feature and time information. For applications that do not require strong temporal information, the distance measure *D*1 can be used. While all the distances defined above can be used to measure the distance between two video sequences, the original total cost *D*0 highly depends on the length of the path. For *D*1, the values are within the range of [0, 1] since the cost is normalized by the length of the path. However, to compare large video sequences, even if the two sequences are very dissimilar, the value of *D*1 may still be very small because of the large length of the path. That makes it difficult to evaluate the variable-length comparisons or choose a suitable global threshold to identify videos. On the other hand, the value of distance measure *D*2 is numerically stable and at the same time has good discrimination ability, as will be shown by the numerical results.

3. Automatic Audio-Visual Topic Segmentation of Multimedia Documents

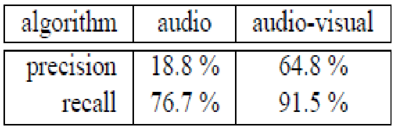
3.1 Video Similarity Search

The methods mentioned above only have been designed to detect videos of the same temporal order and length. To further search videos with changes from query due to content editing, a number of algorithms have been proposed to evaluate video similarity. To deal with inserting in or cutting out partial content, Hua et al. used dynamic programming based on ordinal measure of resample frames at a uniform sampling rate to find the best match for different length video sequences. This method has only been tested on a small video database. Chiu et al. also considered length difference in matching. Through time warping distance computation, they achieved higher search accuracy than the methods proposed. However, with the growing popularity of video editing tools, videos can be temporally manipulated with ease. This work will extend the investigations of copy detection not only in the aspect of potentially different length but also allowing flexible temporal order (tolerance to content reordering). Cheung and Zakhor developed Video Signature to summarize each video with a small set of sampled frames by a randomized algorithm. Shen et al. proposed Video Triplet to represent each clip with a number of frame clusters and estimate the cluster similarity by the volume of intersection between two hyper spheres multiplying the smaller density. It also derives the overall video similarity by the total number of similar frames shared by two videos. For compactness, these summarizations inevitably lose temporal information. Videos are treated as a “bag” of frames thus they lack the ability to differentiate two sequences with temporal re orderings, such as “ABCD” and “ACBD.” On the other hand, temporal characteristic naturally models a video as a trajectory in vector space with each frame as a feature point. Various time series similarity measures can be considered, such as Mean distance, DTW, and LCSS, all of which can be extended to measure the similarity of multidimensional trajectories and applied for video matching. Mean distance normalizes the sum of pairwise distances of feature vector pairs by sequence length.

However, it adheres to temporal order in a rigid manner and does not allow frame alignment or gap, and is very sensitive to noise. DTW can be utilized to address frame alignment by repeating some frames as many times as needed without extra cost, but no frame can be skipped even if it is just a noise. In addition, it is capacity limited in the context of partial content reordering. LCSS is proposed to address temporal order and handle possible noise by allowing some elements to be skipped without rearranging the sequence order, but it will ignore the effect of potentially different gap numbers. Edit distance can handle temporal order and frame alignment by matching two videos with the smallest number of added, deleted, or substituted frames. However, it implicitly assumes that similar frame mappings should be monotonically of the same direction and cannot cross each other, which may not be true for actually similar videos.

The extraction of topic information from a multimedia document (such as broadcast news) requires a correct detection of topic boundaries. The well-known methods for audio segmentation are mainly capable of detecting significant changes in the audio signal, thus indicating speaker turns or transitions from speech to non-speech segments. One approach to topic segmentation is to consider audio boundaries as topic boundaries. However, this will lead to an over-segmentation of the document, as there will be many more audio boundaries than topic boundaries. Our observations have shown that 76.7 % of the existing topic boundaries are detected, but 81.2 % of the boundaries are mistakenly inserted by such an audio segmentation algorithm (see Table 2).

If a multimedia document contains video information, this additional information might be helpful to detect correct topic boundaries. Eickeler and Muller [8] presented a novel approach to scene classification based on Hidden Markov Models which was extended in order to extract real *topic* boundaries. Our approach describes the topics of a news show using an HMM-based topic model that makes use of video as well as of audio features. Each show is thus modeled as a stochastic sequence of topics.



algorithm	audio	audio-visual
precision	18.8 %	64.8 %
recall	76.7 %	91.5 %

Table 2. *Topic segmentation performance.*

3.2 Video Segmentation

Eickeler and Muller [8] used only the visual track of a news show. They segmented it into content classes (Begin, End, Newscaster, Report, Interview, and Weather Forecast) and into edit effects (Cut, Dissolve, Window Change, Wipe). Each of these classes is modeled by a Hidden Markov Model (HMM). The models are combined to a flexible news show structure.

A feature vector consisting of 12 video features represents each image [8]. Among these features are the center, the velocity and variance of motion, intensity of motion, difference histogram and a feature which improves the detection of dissolve edit effects. All these features are based on luminance only. Three more values are added to the vector, giving the average value of the luminance (Y) and the two chrominance (U,V) components. The scene segmentation and classification of a show are the result of calculating the sequence of HMMs that most probably has generated the observed feature vector.

As the described approach does not allow detecting topic boundaries, the following extensions have been introduced.

- Video and audio features are combined into the HMM structure.
- An adapted video model is used which represents the topic structures within a news.

4. Conclusion

We have presented a system that automatically scans multimedia data like TV or radio broadcasts for the presence of specific topics. Each of the three main modules (speech recognition, topic segmentation and topic identification) shows a good performance. The speech recognition module achieved a performance of 18.7 % word error rate using a gender dependent tri phone system. This module can be considered as one of the most advanced German broadcast speech recognition systems.

The well-performing audio-visual topic segmentation module is, unlike many other approaches, able to detect real topic boundaries instead of just audio or video cuts. For topic identification, two algorithms were presented. The new innovative approach is more robust against transcription errors, e.g. mistakenly combined or split compound words, than the standard one, but recognition rates should be improved for a high number of potential topics. The standard approach shows better overall recognition rates.

Acknowledgment

We are thankful to our Guide **Miss. K . N agamani** Asst.Professor Dept.of CSE and **Mr. K. N ageswararao** Assoc.Professor & Head, Dept.of CSE for there valuable suggestion to complete our thesis paper in intime.

References

[1] S.-S. Cheung and A. Zakhor. Efficient video similarity measurement with video signature. *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, pp. 59-74, Jan. 2003.

[2] R. Lancini, F. Mapelli, and A. Mucedero. Automatic identification of compressed video. In *Proc. ICASSP'04*, vol. 3, pp. 445-448, May 2004.

[3] P. Muneesawang and L. Guan. Automatic relevance feedback for video retrieval. In *Proc. ICASSP'03*, vol. 3, pp. 1-4, Apr.2003.

[4] T. Lin, C.W. Ngo, H.J. Zhang, and Q.Y. Shi. Integrating color and spatial features for content-based video retrieval. In *Proc. ICIP'01*, Oct. 2001.

[5] Mauro Cettolo and Marcello Federico, "Model selection criteria for acoustic segmentation," in *Proc. of the ISCA ITRW ASR2000 Automatic Speech Recognition*, Paris, France, 2000, pp. 221–227.

[6] J. Zhou and X.-P. Zhang. Video shot boundary detection using independent component analysis. In *Proc. ICASSP'05*, Mar. 2005.

[7] Alain Tritschler and Ramesh Gopinath, "Improved speaker segmentation and segments clustering using the bayesian information criterion," in *Proc. EUROSPEECH*, 1999, vol. 2, pp. 679–682.

[8] Stefan Eickeler and Stefan M¨uller, "Content-based video indexing of tv broadcast news using hidden markov models," in *Proc. IEEE ICASSP*, 1999, pp. 2997–3000.

[9] A. Marzal and E. Vidal. Computation of normalized edit distance and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 15, no. 9, pp. 926-932, Sept. 1993.

[10] A. Hampapur, K.-H. Hyun, and R.M. Bolle, "Comparison of Sequence Matching Techniques for Video Copy Detection," Proc. Storage and Retrieval for Image and Video Databases (SPIE '02), pp. 194-201, 2002.

Authors Profile

Alla Satish Chowdary pursuing his M.Tech(CSE) in Mother Terissa Institute of Science & Technology, under the guidence of K. Nageswararao Head, Dept.of CSE, and K. Nagamani, Asst.Professor Dept.of CSE, at Mother Terissa Inistitute of Science & Technology, Affiliated to JNTU, Hyderabad, A.P., India. My research Interests are Data Mining and Distributed Databases.

N. Tirupathi pursuing his M.Tech(CSE) in Mother Terissa Inistitute of Science & Technology, under the guidence of K. Nageswararao Head, Dept.of CSE, and K. Nagamani, Asst.Professor Dept.of CSE, at Mother Terissa Inistitute of Science & Technology, Affiliated to JNTU, Hyderabad, A.P., India. My research Interests are Data Mining and Distributed Databases.