MEETING VIDEO RETRIEVAL USING DYNAMIC HMM MODEL SIMILARITY

Dar-Shyang Lee, Jonathan J. Hull, Berna Erol

Ricoh California Research Center, 2882 Sand Hill Road, Menlo Park, CA94025, USA
{dsl,hull,berna}@rii.ricoh.com

ABSTRACT

Overcoming the semantic-feature gap and adapting to context are two main challenges in content-based retrieval. The problem is even more complicated for unstructured videos such as automated recordings of meetings. To address this problem, we propose a model-based approach to meeting retrieval with user controlled weighting for dynamic similarity comparison. Each video is represented by an HMM, and the similarity between videos is determined by comparing the corresponding models. Users can control the relative importance of temporal and static features by adjusting a weighting parameter in a way similar to content-based image retrieval. Experimental results demonstrate the feasibility and versatility of this approach.

1. INTRODUCTION

The most challenging aspects of content-based retrieval are bridging the gap between semantics and low-level features, and adapting to context dependency of the retrieval task. Typical solutions today utilize machine learning and relevance feedback to address these issues. A critical component underlying this approach is a similarity criterion that can be dynamically adjusted to correctly reflect the context and semantics. For image-based retrieval, visual similarity is commonly determined by a combination of color, texture, and shape features weighted according to a proportion either manually specified by the user or automatically learned through a training process.

The same definition of *visual similarity* has been used to measure video content similarity without much consideration for the temporal information. Videos are often treated as a set of static key frames where standard definition for visual similarity can be applied. Although some systems take into account the temporal alignment of key frames [4][10], very little attention has been paid to the meaning of *spatial-temporal* similarity as an adjustable measure. The problem can be illustrated using meeting video retrieval as an example. What does it mean for two meetings to be similar? Would two brainstorming sessions involving different people be more similar than two different types of meetings involving the same people? The answer is almost certainly context dependent.

Although numerous similarity models and retrieval techniques for query-by-example have been proposed for professionally generated videos, they are inadequate for automated meeting recordings due to the relatively low visual and semantic structure present in these videos. Instead, there is richer information in the participants and their interactions. Identifying meeting participants and meeting types, among other things, will not only provide useful retrieval cues, it can also improve speech transcription accuracy through appropriate choice of lexicon and speech style [1].

Under these considerations, we propose an approach to meeting video retrieval where similarity is dynamically computed by comparing their model representations with an adjustable parameter. Our framework uses hidden Markov models (HMM) to represent meetings as stochastic processes involving *interactions* among a number of *participants*, two prominent characteristics of meetings. A weighted combination of the divergence between the stationary and temporal components of the distributions represented by two HMMs determines the final similarity score. Adjusting the weighting effectively compares meetings solely based on participant information at one extreme, and entirely based on patterns of turn-taking at the other, or anything in between.

This work has several novel aspects. First, it uses an HMM model-based approach for meeting video retrieval. A fundamental component of this approach is a way to measure the dissimilarity of two HMMs, for which we proposed a new method for efficiently computing the divergence bound. In addition, the formulation leads to an adjustable similarity measure which allows context adaptation. The proposed solution extends the dimension of adjustable parameters over to the temporal domain, which has not been addressed before.

The rest of the paper is organized as follows. In Section 2, we describe the overall system and details of the meeting HMM representation. We also introduce background information on HMM dissimilarity measure and outline our approach based on the upper-bound on divergence from model parameters. The system is applied to dynamically compare meeting similarities based on participants or style criterion. The experiments are discussed in Section 3. We review relevant work in Section 4 and conclude in Section 5 with a summary of our findings.

2. BASIC FRAMEWORK

Our basic approach is to represent every meeting with an HMM, and evaluate its relevance to a query meeting based on the similarity of their models, taking into account the relative importance of features specified by the user. Details of the model representation, similarity measure and relevance ranking are described in the following subsections.

2.1 HMM Representation for Meetings

A meeting is characterized by a sequence of feature vectors, which could be any text, audio or visual features computed over a small time interval. This sequence is modeled by an HMM, $M = \{Q, \boldsymbol{\pi}, \mathbf{A}, B\}$, consisting of a set of states Q and associated

0-7803-9332-5/05/$20.00 ©2005 IEEE

initial, transition and emission probabilities, respectively. HMMs provide a compact representation that captures both global statistics of the features and temporal transitions of the feature distribution. Two of the most important attributes that characterize meetings are the participants and the style of interaction. HMMs trained on audio feature vector sequence capture both types of information, with emission probabilities reflecting the characteristics of the speakers in the feature space, and the transition probabilities representing the style of interaction.

In order to train the continuous density HMM such that the states roughly correspond to distinct speakers, we use the segmental k-means algorithm. First, unsupervised speaker clustering is performed on speaker location data to identify n clusters $C_1 \cdots C_n$. In our recording settings, sound source localization is computed and smoothed to generate audio segments [9]. The audio directions not only provide good segmentation boundaries for individual speakers, they provides a simple and yet sufficiently effective clustering of speakers when they are not moving. Once the clusters have been identified, each cluster C_i is represented by a corresponding state in the HMM. The transition probabilities can be calculated based on the cluster ID. The MFCC vectors are extracted from the voice segment of each speaker, followed by vector quantization to produce K codebook vectors. These codebooks are used to define a mixture density function for the emission probability of that state.

2.2 Model Similarity Measure

In order to evaluate the relevance of meetings, a similarity measure for comparing HMMs is needed. In general, these meeting HMMs are ergodic and do not have a constant number of states, as illustrated in Figure 1. We define a measure based on KL-Divergence (KLD) with a weighting scheme for user to adjust the relative importance between speaker similarity and interaction style.

Computing the divergence between two HMMs is non-trivial. The most general solution involves Monte Carlo simulation, evaluating a long random sequence generated by one model under the other [8]. This solution is commonly simplified to compute the likelihood of one observed sequence under the other model [1], which requires storing the original vector sequence. Closed-form approximations exist only for special

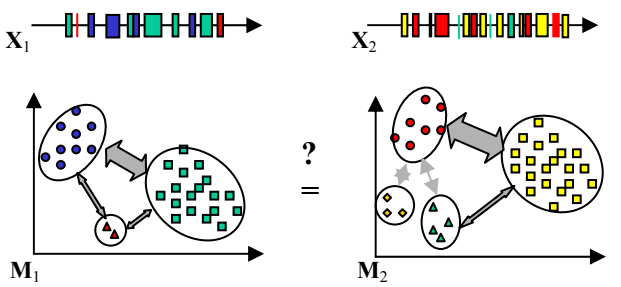

Figure 1 – Meetings are represented by HMMs with a number of speaker states (shade coded) with emission probabilities defined over feature space, and transition probabilities (shown as arrows) among the states. Meeting similarity is determined by comparing two HMMs.

architectures [13] or symbol density functions [5].

Our solution is based on the divergence upper-bound derived for state-aligned ergodic HMMs with Gaussian mixture density emission probabilities [3]. It was shown that the KL-Divergence, D, between two HMMs as computed by an infinitely long observation sequence is bounded by the following equation

$$D(M_1 \| M_2) \le \sum_{i=1}^{n} v_1^i (D(\mathbf{a}_1^i \| \mathbf{a}_2^{m(i)}) + D(\mathbf{b}_1^i \| \mathbf{b}_2^{m(i)})), \qquad (1)$$

where $\mathbf{v}_1$ is the stationary vector of M_1, and $m(i)$ is the state correspondence mapping from M_1 to M_2. The divergence for the transition probabilities is expressed as

$$D(\mathbf{a}_1^i \| \mathbf{a}_2^j) = \sum_{k=1}^{n} a_1^{ik} \log \frac{a_1^{ik}}{a_2^{jm(k)}} \qquad (2)$$

To compute divergence for the emission probabilities, a similar upper-bound derived for Gaussian mixtures is applied

$$D(\mathbf{b}_1^i \| \mathbf{b}_2^j) \le D(\mathbf{w}_1^i \| \mathbf{w}_2^j) + \sum_{k=1}^{K} w_1^{ik} D(g(\mathbf{c}_1^{ik}) \| g(\mathbf{c}_2^{jk})) \qquad (3)$$

where $g(\mathbf{c}_1^{ik})$ is the Gaussian centered at $\mathbf{c}_1^{ik}$, the k-th codebook vector of state q_1^i of M_1. Eq.(1)-(3) defines the basic solution for measuring HMM dissimilarity. Simply stated, the divergence between two HMMs is bounded by the divergence of the transition and emission probabilities between corresponding states.

However, several issues remain to be addressed. First of all, the quality of the bound depends critically on the assumed state alignment. For example, two identical HMMs can produce zero divergence bound with the correct state alignment, and a vastly different result if the states are misaligned. All possible permutations of state alignments, an $O(N^2N!)$ operation, must be evaluated to guarantee the best result. Although meetings usually involve a small number of active participants, this is still an expensive process. Since computing the divergence of transition probabilities of two corresponding states in Eq.(2) depends on the alignment of all other states, dynamic programming cannot be applied.

We resolved this by considering only *self-transitions* versus *out-going* transitions when comparing transition probabilities. More precisely,

$$D(\mathbf{a}_1^i \| \mathbf{a}_2^j) \approx a_1^{ii} \log \frac{a_1^{ii}}{a_2^{jj}} + (1 - a_1^{ii}) \log \frac{1 - a_1^{ii}}{1 - a_2^{jj}} \qquad (4)$$

There is an intuitive interpretation for this modification. By merging all out-going links, we preserve the essential difference between a monologue versus an interactive conversation while reducing the variations of specific speaker ordering. With this modification, the lowest KLD bound can be computed as a shortest path problem using dynamic programming.

A similar problem occurs when computing KLD for two mixture densities. Assuming all components have equal weights and variances, we approximate this by the averaging distances of the closest codebooks between the two states

$$D(\mathbf{b}_1^i \| \mathbf{b}_2^j) \approx \frac{1}{K} \sum_{k=1}^{K} \min_{1 \le r \le K} \{ \| \mathbf{c}_1^{ik} - \mathbf{c}_2^{jr} \| \} \qquad (5)$$

Finally, when models have different number of states, the state mapping function $m(i)$ would be undefined for unmatched

states. Unfortunately, no bounds can be derived for Eq.(1) in those situation. In our solution, a cost of v_1^i is added. When states correspond to speakers, this cost discriminates between one meeting with person A, B, C from another with A, B, C and D. The presence of D must be accounted for even when the states A, B, and C match perfectly.

2.3 Dynamic Relevance Evaluation

The formulation in the previous section suggests an intuitive solution for adjustable weighting by user input. Since the divergence can be computed separately for the transition and emission probabilities, applying a weighting to each portion determines the relative importance between the static and temporal components of the models. Therefore, we define

$$S(M_1,M_2;\alpha)=(1-\alpha)\cdot J_{\mathrm{A}}(M_1,M_2)+\alpha\cdot J_{\mathrm{B}}(M_1,M_2) \quad (6)$$

$$J_A(M_1,M_2)=\sum_{i=1}^{n_1} v_1^i D(\mathbf{a}_1^i \,\|\, \mathbf{a}_2^{m_*(i)})+\sum_{j=1}^{n_2} v_2^j D(\mathbf{a}_2^j \,\|\, \mathbf{a}_1^{m_*'(j)}) \quad (7)$$

$$J_B(M_1,M_2)=\sum_{i=1}^{n_1} v_1^i D(\mathbf{b}_1^i \,\|\, \mathbf{b}_2^{m_*(i)})+\sum_{j=1}^{n_2} v_2^i D(b_2^j \,\|\, b_1^{m_*'(j)}) \quad (8)$$

Function $m_*(i)$ is the best alignment mapping from M_1 to M_2, and $m_*'(j)$ is the reverse mapping. The divergence is computed in both directions to make the measure symmetric. Parameter α controls the relative importance between similarity in overall static distribution, J_B, and similarity in temporal characteristics of distribution changes, J_A. When $\alpha=1$, the system should retrieve meetings with similar participants regardless of the types of meeting. When $\alpha=0$, similarity is judged only by meeting style. Any values in-between should weigh these factors appropriately with $\alpha=0.5$ being the default comparison that weighs both components equally. The appropriate weighting can be adjusted interactively by users to refine their queries or learned using relevance feedback. Figure 2 shows an example interface with these three preconfigured setting where each highlighted link triggers a different search.

3. EXPERIMENTAL RESULTS

The system was tested on recordings of group meetings and presentations captured by an automated meeting recording system. Since evaluation of meeting relevance is subjective, it is difficult to perform a quantitative analysis. We designed an experiment to verify the intended behavior of the system. A control set of four recordings totaling 6 hours of video were selected to represent 4 distinct categories. Two were weekly meetings for group A and B, and two were talks given by a member of group A and B. The group meetings ($\mathcal{D}$) tend to be highly interactive with equal participation; while presentations ($\mathcal{P}$) tend to be dominated by a single speaker. We denote the four categories as {A$\mathcal{P}$,B$\mathcal{P}$,A$\mathcal{D}$,B$\mathcal{D}$}. Each recording class is time-compressed by randomly sub-sampled 10 times at two second granularity with a probability of 0.3 to create a total of 40 test meetings.

Figure 2 – Example interface of a meeting database. Highlighted links retrieve similar meetings under different criterion.

A 20-dimensional feature vector was used, composed of order 10 MFCC and first derivatives computed over 25ms windows on voiced segments. A rough approximation to unsupervised speaker identification is provided by clustering on speaker locations provided by our automated recording system [9]. We assumed the number of speakers was known. Using the cluster ID, the audio vectors were labeled and filtered using k-nearest-neighbor to reduce noise. Then 16 codebooks were extracted for each cluster using LVQ. Transition probabilities were estimated from the sequence labels.

We evaluate the performance with different α setting using each test meeting as query. Since the mean average precision would be favorably biased on a small dataset, we measure the *R*-Precision, the number of correct recalls in the top *R* ranked items, where *R* is the number of relevant items in the database [12]. With $\alpha=0.5$, all 4 categories are well separated, as shown in Table 1.

$\alpha=0.5$	A$\mathcal{D}$	A$\mathcal{P}$	B$\mathcal{D}$	B$\mathcal{P}$	Total
R=9	87.8%	92.2%	92.2%	100%	93.1%

Table 1 - *R*-Precision with equal weighted similarity.

When $\alpha=0$, we expect the presentations in class A$\mathcal{P}$ and B$\mathcal{P}$ would be merged. Therefore, given a query in one of the two classes, the ideal ranking should place the other 19 cases ahead of other meetings belonging to A$\mathcal{D}$ and B$\mathcal{D}$. The precisions for *R*=19 are shown in Table 2. The precision for combined classes A$\mathcal{P}$ and B$\mathcal{P}$ is 100%, and 76.3% for A$\mathcal{D}$ and B$\mathcal{D}$. The overall precision is 88.2%. By ignoring speaker characteristics, the system dynamically created a presentation class and a group meeting class.

$\alpha=0$	{A+B}$\mathcal{P}$	{A+B}$\mathcal{D}$	Total
R=19	100%	76.3%	88.2%

Table 2 – *R*-Precision based on style similarity.

Finally, we applied $\alpha=1$ to find meetings with similar participants. It is obvious that it should at least retrieve the 4 respective classes in the top ranks. The question is whether it will group A$\mathcal{D}$ with A$\mathcal{P}$, and B$\mathcal{D}$ with B$\mathcal{P}$, since only one member in a group dominated in a presentation. We computed the precision based on a 4-class grouping as well as a merged two-class grouping, shown in Table 3. The system correctly recalled the instances of the original 4 classes reasonably well, averaging 88.1% precision in the top 9 ranked meetings. If we consider the joined classes of group A and group B, a 70.7% precision is achieved in the top 19 recall.

$\alpha = 1$	A$\mathcal{D}$	A$\mathcal{P}$	B$\mathcal{D}$	B$\mathcal{P}$	Total
R=9	87.7%	87.8%	85.6%	91.1%	88.1%
$\alpha = 1$	A{$\mathcal{D}$+$\mathcal{P}$}		B{$\mathcal{D}$+$\mathcal{P}$}		Total
R=19	73.2%		68.1%		70.7%

Table 1 – *R*-Precision using speaker similarity.

Overall, the system behaved consistently with our expectation. As is the case with image retrieval based on visual similarity, a dynamically adjustable similarity criterion is a key element for building a more intelligent retrieval system based on relevance feedback or other learning paradigm to bridge the semantic gap.

4. RELATED WORK

In recent years, numerous approaches toward content-based video retrieval have been proposed. The area of research most relevant to this work is video query-by-example where the goal is to find videos in a database similar to a given video clip. Several systems have been developed based on different definitions of similarity and structural analysis techniques. One of the earliest video retrieval systems QBIC is based on a direct extension of techniques developed for static images [6]. The standard approach today usually incorporates temporal ordering on the frame-based distance [4], or involves more complex audiovisual analysis within a segmentation hierarchy [10].

The algorithm proposed in this paper solves issues not addressed by previous work. First of all, the reliance on visual and semantic patterns provided by the domain context makes previous solutions inadequate for unstructured contents captured by an automated system. Meeting videos typically do not have the level of visual variations required for shot segmentation and key frame extraction, which are necessary elements in these techniques. Moreover, most meetings lack the rigid semantic structure of news broadcasts and salient motion patterns of sports videos explored by earlier work. Another difference is that previous work considers temporal information in a local window. While a measure of local temporal similarity is useful for detecting a heated discussion or a person walking to a whiteboard, the overall style of interaction within a meeting is conveyed by the global characteristic of interaction. For unstructured video such as meetings, the statistical tendency of temporal transitions is more important than the exact ordering of the sequence. Therefore, an approach based on local sequence alignment such as dynamic time-warping is inappropriate. None of the previous work considered a global measure of similarity of temporal characteristics for meetings and allow context adaptation by modifying the similarity criterion.

Content-based retrieval is a multifaceted problem. An effective system will undoubtedly involve more than one search strategy. For retrieval of meetings, a variety of analysis techniques exist that extract keywords, topics, location, participant identity, etc [14]. There are also techniques that utilize turn-taking or other sequence information for meeting categorization. However, they typically train classifiers, such as HMMs, to identify predetermined categories or sequence events [7][11]. No model comparison was involved and dynamic categorization was not possible. In [2], an HMM was constructed to model each video frame, and model dissimilarity was used on a sequence of frames to identify segments and key frames. None of these earlier works has used HMM for sample representation, and provided a similarity measure for model comparison in a query-by-example context, especially an adjustable one based on user context.

5. CONCLUSIONS

We described a model-based system for meeting video retrieval that allows dynamic relevance ranking using an adjustable similarity criterion of participant information and interaction styles. HMMs provided an effective representation of speaker characteristics and turn-sequence information. An algorithm was developed to efficiently compute the divergence bound of HMMs. The similarity measure can be adjusted to weigh the relative importance between stationary and temporal component in model comparison. Experimental results showed that dynamic groupings consistent with our intuition can be formed and retrieved through a control parameter.

6. REFERENCES

[1] S. Burger, V. MacLaren, H. Yu, "The ISL Meeting Corpus: The Impact of Meeting Type on Speech Style," ICSLP, Sept. 2002.

[2] D. DeMenthon, L. J. Latechi, A. Rosenfeld, M.V. Stuckelberg, "Relevance Ranking of Video Data Using Hidden Markov Model Distances and Polygon Simplification," ICVIS 2000, pp.49-61.

[3] M. N. Do, "Fast Approximation of Kullback-Leibler Distance for Dependence Trees and Hidden Markov Models," IEEE Signal Processing Letters, Apr. 2003.

[4] N. Dimitrova and M. Andel-Mottaled, "Content-based Video Retrieval by Example Video Clip", SPIE vol.3022, pp.59-71, 1997.

[5] M. Falkhausen, H. Reininger and D. Wolf, "Calculation of Distance Measures between Hidden Markov Models," EuroSpeech, pp. 14871490, 1995.

[6] M. Flicker, et. al. "Query by Image and Video Content: the QBIC System", Computer, vol.28(9), pp.23-32, 1995.

[7] D. Gatica-Perez, I. McCowan, M. Barnard, S. Bengio, H. Bourlard, "On Automatic Annoation of Meeting Databases," ICIP, Sept. 2003.

[8] B. H. Juang and L.R. Rabiner, "A Probabilistic Distance Measure for Hidden Markov Models", AT&T Technical Journal, vol.62, no.2, pp. 391-408, Feb. 1985.

[9] D. S. Lee, B. Erol, J. Graham, J. J. Hull, N. Murata, "Portable Meeting Recorder," ACM Multimedia, pp.493-502, 2002.

[10] R. Lienhart, W. Effelsberg, R. Jain, "VisualGREP: A Systematic Method to Compare and Retrieval Video Sequences", SPIE, vol.3312, 1998.

[11] S. Reiter and G. Rigoll, "Segmentation and Classification of Meeting Events using Multiple Classifier Fusion and Dynamic Programming", Proc. ICPR, 2004.

[12] TREC Video Retrieval Evaluation, http:// http://www-nlpir.nist.gov/projects/trecvid/.

[13] M. Vihola, M. Harju, P. Salmela, et. al., "Two Dissimilarity Measures for HMMs and their Application in Phoneme Model Clustering", ICASSP, pp. 933-936, 2002.

[14] A. Waibel, H. Yue, H. Soltau, et. al. "Advances in Meeting Recognition", Proc. of Human Tech. Conf., 2001.