

Detecting topical events in digital video

Tanveer Syeda-Mahmood*
IBM Almaden Research Center
K57/B2, 650 Harry Road
San Jose, CA 95120
stf@almaden.ibm.com

S. Srinivasan
IBM Almaden Research Center
K57/B2, 650 Harry Road
San Jose, CA 95120
savitha@almaden.ibm.com

ABSTRACT

The detection of events is essential to high-level semantic querying of video databases. It is also a very challenging problem requiring the detection and integration of evidence for an event available in multiple information modalities, such as audio, video and language. This paper focuses on the detection of specific types of events, namely, topic of discussion events that occur in classroom/lecture environments. Specifically, we present a query-driven approach to the detection of topic of discussion events with foils used in a lecture as a way to convey a topic. In particular, we use the image content of foils to detect visual events in which the foil is displayed and captured in the video stream. The recognition of a foil in video frames exploits the color and spatial layout of regions on foils using a technique called region hashing. Next, we use the textual phrases listed on a foil as an indication of a topic, and detect topical audio events as places in the audio track where the best evidence for the topical phrases was heard. Finally, we use a probabilistic model of event likelihood to combine the results of visual and audio event detection that exploits their time co-occurrence. The resulting identification of topical events is evaluated in the domain of classroom lectures and talks.

Keywords

Topic of discussion events, multi-modal fusion, slide detection, topical audio events, query-driven topic detection.

1. INTRODUCTION

Despite the progress made in image and video content retrieval, making high-level semantic queries, such as looking for specific events, has still remained a far reaching goal. Yet, most practical applications embedding content-based retrieval require precisely a way to handle such queries. One such application is the domain of distributed or distance learning where querying the video content for events containing a topic of discussion is a desirable component of any

*Author for correspondence.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

ACM Multimedia 2000 Los Angeles CA USA

Copyright ACM 2000 1-58113-198-4/00/10...\$5.00

learning system. While a large number of studies have been done on event perception in various fields such as economics, perception, psychophysics, and artificial intelligence, the automatic detection of events has remained a challenging problem due to several reasons. First, the detection and understanding of events requires examining objects, actions, and their inter-relationships automatically. Secondly, events are often multi-modal requiring the gathering of evidence from information available in multiple media sources such as video and audio. Finally, identifying an event also requires identifying a duration over which it occurs (including the precise start and end times). Even with the best techniques for visual or audio scene analysis, event detection using the individual cues will continue to possess robustness problems due to detection errors. For example, the detection of objects and their relationships are well-known to be difficult problems in computer vision. Further, the localization inaccuracies with individual cue-based detection often lead to conflicting indications for an event at different points of time making their multi-modal fusion difficult. Previous work on the automatic detection of events has primarily focused on actions including event classification[12], and object recognition for capturing visual events. The automatic detection of auditory events, on the other hand, has been mainly limited to discriminating between music, silence and speech[15]. The notion of combining of audio-visual cues, though not for event detection, has also been explored by others. In general, the methods of combining cues have considered models such as linear combination[4] including Gaussian mixtures, winner-take-all variants[7], rule-based combinations[2], and simple statistical combinations[2].

In this paper, we address the problem of event detection in digital video by focusing on specific types of events, namely, topic of discussion or topical events. Topical events are defined here as points of time in a video where a specific topic as indicated in a given foil, was discussed. From a survey of the distance learning community, it has been found that the single most useful query found by students is the querying of topic of interest in a long recorded video of a course lecture. Such classroom lectures and talks are often accompanied by foils (also called slides) some of which convey the topic being discussed at that point in time. Figure 1c,f,i shows examples of such slides. When such lectures are video taped, at least one of the cameras used captures the displayed slide, so that the visual appearance of a slide in video can be a good indication of the beginning of a discussion relating to a topic. However, the visual presence alone may not be sufficient,

since it is possible that a speaker flashes a slide without talking about it, or can continue to discuss the topic even after a slide is removed. In this paper, therefore, we focus on detecting topical events by combining visual and audio cues derived from the image and textual content of foils.

To our knowledge, no work has yet been done on the detection of topical events using a combination of visual and audio search of foils. The work reported here makes several novel contributions in that direction. First, we present a novel method of topical visual event detection by identifying points of time in video where a foil was displayed and captured in the video stream. The identification of specific foils exploits the color and spatial layout geometry of regions on foils using a technique called region hashing. Specifically, an illumination-invariant description of color is used for robust foil localization. Region hashing, being pose-invariant assures robust recognition of foils. Next, we present a novel method of topical audio event detection based on the phrasal content of foils. In particular, by relying on the phrases listed on a foil as a useful indication of the topic, we search the audio track for places where the phrases were spoken. The search uses a combination of word and phonetic recognition of speech and exploits the order of occurrence of words in a phrase to return points in video where one or more sub-phrases used in the foil were heard. The individual phrase matches are then combined into a topical match for the audio event using a probabilistic combination model that exploits the contiguity of occurrence to group phrasal matches into audio events. Finally, we present a novel method of multi-modal fusion for overall topical event detection that uses a probabilistic model to exploit the time co-occurrence of individual audio and video events. The automatic topical event detection method reported here has been extensively tested to demonstrate effective topic detection as part of a distributed learning system for the indexing and browsing of a large number of teaching and training videos.

2. TOPICAL VIDEO EVENT DETECTION

We begin by addressing the problem of topical video event detection through the detection of foils in a video. There has been some work done in the multimedia authoring community to address this problem from the point of synchronization of foils with video. The predominant approach has been to do on-line synchronization using a structured note-taking environment such as Zenpads[1] to record the times of change of slide electronically and synchronize with the video stream. Current presentation environments such as Lotus Freelance or Powerpoint have features that can also record the change time of slides when performed in rehearse modes. In distributed learning, however, there is often a need for off-line synchronization of slides since they are often provided by the teacher after a video recording of the lecture has been made. The detection of foils in a video stream under these settings can be challenging. There are a multitude of ways in which foils appear depending on the camera geometry used in taping lectures. The resulting appearance of slides in video can vary greatly in color, and the slides themselves could appear anywhere in the video frame. Figure 1a,d,g show examples of different slide appearances possible in videos. A solution to this problem for constrained camera geometries was presented in[9]. There, a solution was proposed for a two-camera geometry, in which

one of the cameras was fixed on the screen depicting the slide. Since this was a more-or-less calibrated setting, the boundary of the slide was visible so that the task of selecting a slide-containing region in a video frame was made easy. Further, corners of the visible quadrilateral structure could be used to solve for the 'pose' of the slide under the general projective transform. Our approach to foil detection is meant to consider more general imaging situations involving one or more cameras, and greater variations in scale, pose and occlusions. We break the foil detection problem into two phases, namely, 1) detecting foil-containing regions in video frames, and 2) recognizing which of a given set of foils appears in a slide-containing region of a video frame.

Detection of foil regions based on color

We detect slide containing regions within video frames using the background color of slides. Since there is considerable color variation in the appearance of the slide in a video frame from its original electronic or hardcopy form (see Figure 1), this can be a difficult problem even in cases where a uniform color background is used. To enable a robust detection of slide using the background color, we adopted an approach to describing color based on surface color classes as reported in an earlier work[18]. A surface color class is the set of surfaces with same spectral properties but different spatial distributions. It was shown in [16] that the eigenvector of the projected clusters from samples of such surface classes is an illumination-invariant description of the surface color class. We model the background color of slides (the largest region on a foil that encloses all other regions is taken to be the background region) in terms of this description. For cases in which multi-colored backgrounds are used, we describe it by multiple surface color class descriptions.

To detect foils in video, we first process the video to group into shots using conventional histogram-based scene clustering methods. Each such shot is represented by a keyframe. To handle videos with fixed camera settings that generate very few shots, we also allow a regular sampling of video (for eg., once per second) to ensure at least twice as many keyframes as the number of slides used in the talk. The method of detecting color regions using the specific color class description described in [16] was then applied to detect regions in the keyframes that contain one or more of the background colors of the specified foil set.

Figure 1b,e,h shows background color detection in sample video frames shown in Figure 1a,d,g using the query slides of Figure 1c,f,i respectively. As can be seen, the detection works well even under considerable changes in color appearance. A detailed analysis of the results of slide detection are reported in Table 1 and are discussed under the evaluation of topical event detection in Section 5.

Recognition of foils in video

Even though color-based selection points to candidate regions, it cannot be used to identify which of the foils is depicted in the region, as most foils of a set tend to have the same (slide master) background. Also, due to the scale at which these foils are imaged, it is not possible to decipher the individual words on a slide using an OCR algorithm. In

addition, differences between successive slides can be small (eg. when a topic is continued) so that foil recognition requires a detailed modeling of spatial layout of the smaller regions constituting the foil. Such a modeling of spatial layout, however, must be pose-invariant, to account for effects of warping, rotation, and scaling that are often present due to the camera geometry used for taping the lectures. Finally, it should be robust to occlusion errors that are present often as speakers move in front of displayed screen, or when camera pans to the surrounding scene. A technique for recognizing objects by the spatial layout of regions called region hashing, was presented in an earlier paper[17]. In this paper, we apply region hashing to the problem of foil recognition. For this, we note that the foils or slides displayed on a screen can be modeled as planar regions in space, so that their transformation assuming orthographic projection, can be modeled as 2d affine distortion¹. For constrained camera geometries, perspective effects can also be modeled as described in [9]. To model the spatial layout of foil regions using region hashing, we exploit the well-known observation that the shape of a 2d pattern can be described in a pose-invariant fashion by recording the affine coordinates of features within object, computed with respect to a triple of basis features chosen as an object-based reference frame[8]. The relative location of a pair of foil regions can be specified precisely and in a pose-invariant fashion through affine intervals, i.e. the interval in which affine coordinate values lie. Thus ideally, a region of a video frame can be recognized as containing a specific foil if the affine intervals of corresponding region pairs are identical. In practice, due to occlusions, the affine intervals overlap rather than register exactly. To account for missing data which cause affine intervals to be missing altogether, we compute affine intervals w.r.t multiple basis triples and region pairs on a given foil, and store them in a suitable index structure. This pre-computation saves time during recognition of foils in video stream following the principle of geometric hashing[8].

The electronic foil images of a foil set are pre-processed to extract features. Specifically, curves are extracted from an edge map of a slide. Connected components of curves are used to form regions within the foil image. Consecutive features along curves are used to form basis points. The affine coordinates of all features of one region are then computed w.r.t. a basis triple of another region, and the range in which they lie are noted in the corresponding affine interval. The spatial layout of each foil is then represented as

$$\text{Object layout} = \{(R_i, C_{R_i})(R_j, C_{R_j}), \{Int_{ij}, B_{ik}\}\} | 1 \leq i, j \leq N \quad (1)$$

where N is the number of object regions, C_{R_i} is the color of the region R_i , and Int_{ij} is the affine interval information given by $\langle (\alpha_{jmin}, \beta_{jmin}), (\alpha_{jmax}, \beta_{jmax}) \rangle$ of features F_j of region R_j computed with respect to k th basis B_{ik} of Region R_i . For slides containing single color text, the color of the region is not distinctive information. On the other hand, for slides containing diagrams and images, color can be useful for pruning false matches. The affine interval information is consolidated and represented in an index structure called the interval hash tree. The details of the interval hash tree

¹The region hashing formalism relying on the overlap of affine intervals also covers in most cases the usual effects of perspective projection.

construction and search are reported in [17] and are skipped here for brevity.

Given a query foil-containing region in a video frame, an identical processing is done to generate the affine intervals with the exception that they are computed with respect to one basis triple per region. In our experiments we found the median basis triple of a curve to be a reliable choice. We then find evidence for overlap of query affine intervals of all query region pairs with a subset of database affine intervals, by indexing the interval hash tree used to store the affine intervals. Let the affine interval information retrieved for a query region pair $F_O = (R_{O_i}, C(R_{O_i}), R_{O_j}, C(R_{O_j}), \langle Int_{O_{ij}}, B_{O_{ijm}} \rangle)$ after such indexing be denoted by $\{R_k, C(R_k), R_l, Int_{kl}, B_{klm} \rangle\}$. We first discard region pairs if the corresponding region identities do not match i.e., $C(R_k) \neq C(R_{O_i})$ and $C(R_l) \neq C(R_{O_j})$, or their overlap is less than a certain threshold. The score of the basis retrieved B_{klm} is then incremented by the extent of interval overlap $\frac{2Int_{kl} \cap Int_{O_{ij}}}{Int_{kl} \cup Int_{O_{ij}}}$. We then select the top few basis, and declare their corresponding enclosing regions as matching region pairs, and the corresponding foils as candidate matching foils in the database. For each foil selected, we use the basis triple pair with the highest score in the region pair with the highest score as a candidate matching basis. Since these are three pairs of matching points, an affine transformation relating the object to its presence in the image is found and used to project the selected foil image at the foil-containing region in the video frame for verification. The sum of verification scores of all potential matches is used for normalization to render scores of matching positions as probabilities.

Examples

We now illustrate regions hashing for foil recognition through a few examples. Figure 1a,d,g shows examples of keyframes from sample videos. The recognized slide in each of video frames at the located slide region shown in Figure 1b,e,h are shown in Figure 1c,f,i respectively. More details on results of slide image matching for topical event detection are described in Section 5.

Topical video event detection

The above method of foil matching in video can now be applied to detect topical video events. Since the foil matching is performed on keyframes, the video event spanned by the foil topic based on image content is taken to be the duration between the scene changes of two consecutive foil matches. That is, let b_i, e_i be the shot duration corresponding to the keyframe F_i . Let i th match to the foil image Q_j be in frame F_k with probability of correctness p_k . Let F_l be the keyframe with the smallest $l > k$ where a match to foil $Q_m \neq Q_j$ was found (notice that the intermediate keyframes between F_l and F_k may be pure scene changes not depicting a slide). Then the topical video event duration of the i th match of foil image Q_j is (b_k, e_{l-1}) . Notice here that we are regarding different match instances to a foil as separate events, even though they may be related.

Video	Number of foils	video frames	total keyframes	keyframes showing foils		Correct matches	false or no matches
				Detected	Actual		
1.	13	59457	20	11	10	9	1
2.	24	143900	196	61	46	42	7
3.	6	19318	25	10	6	6	0
4.	11	23398	25	12	12	12	0
5.	10	17054	161	14	9	8	4
6.	10	19176	127	16	10	10	1

Table 1: Precision in visual event detection.

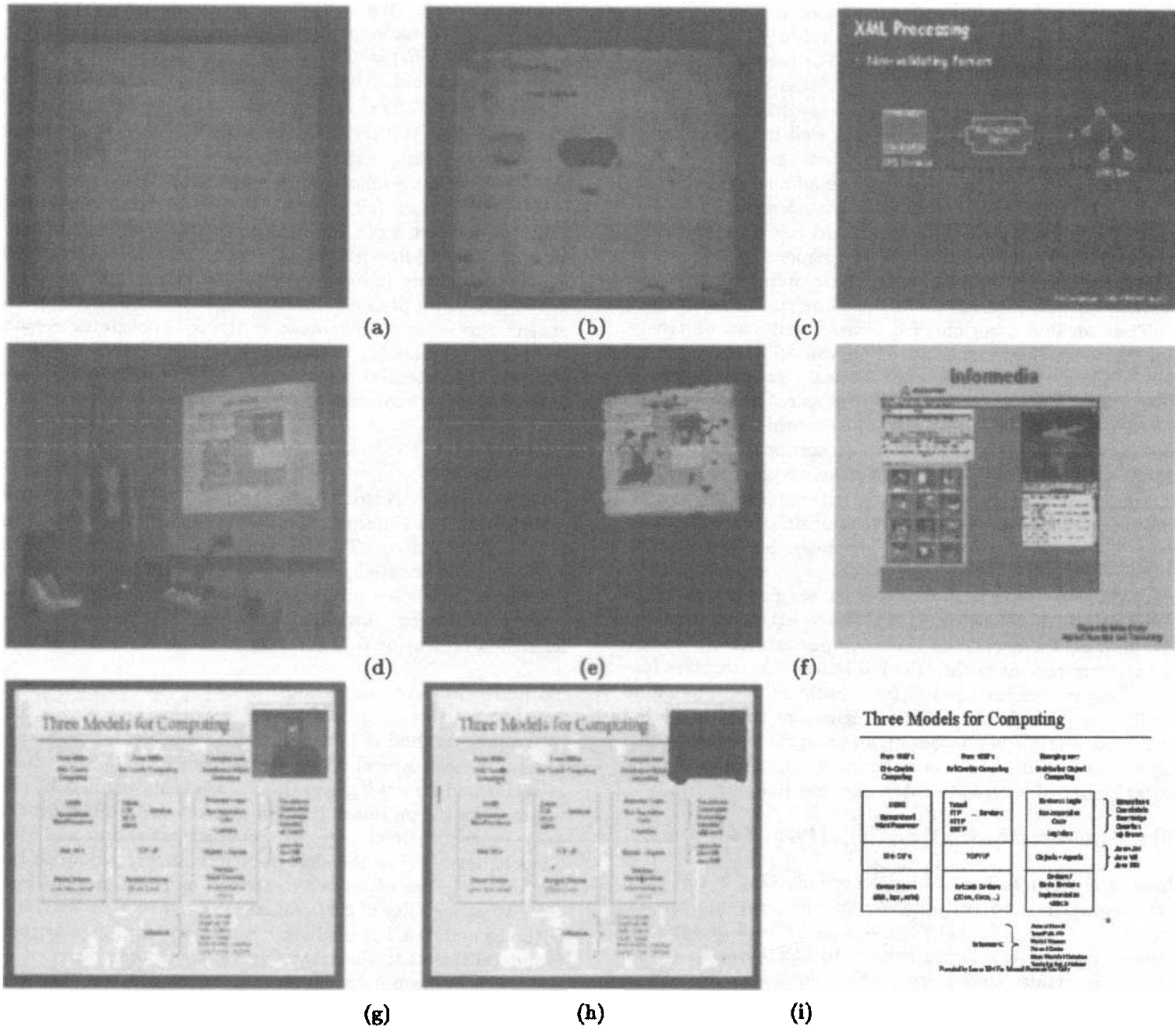


Figure 1: Illustration of foil detection and recognition. The first column shows the video frames, the second column shows the slide-containing regions detected based on background color, and the third column shows the corresponding recognized slide in the detected region.

3. TOPICAL AUDIO EVENT DETECTION

We now turn our attention to the detection of topical audio events in which topical phrases indicated on a foil were spoken. Detecting topical audio events is particularly useful in case of errors in foil recognition which can either cause portions of video discussing a topic to be entirely missed, or worse still, can point to the wrong point of time. It also becomes necessary in cases where the slide was not displayed, or displayed but not captured by the camera, so that only the audio can be relied upon to indicate topical information. The deduction of topics from transcribed audio or text documents is an area of intense exploration in the information and spoken document retrieval communities [6, 3, 5, 11]. Much of the research addresses topic discovery for large document collections, which can be referred to as intervideo topic discovery. These methods focus on the bottom-up detection of topics by exploiting the multiple occurrences of phrases (both local and global) to indicate their topical relevance. We take a different approach here by focusing on the problem of time-localized topical audio event detection (rather than topic detection) within a single video. Further, we take a query-driven approach in that we assume the desired topical event can be suitably abstracted in the topical phrases used on foils. For example, from the slide shown in Figure 2a we can guess that at the time the slide was displayed in the video, the speaker may have been discussing the details of an XML schema. Other slides such as those containing mostly graphics as shown in Figure 3a, need not give as clear an indication of the topic of discussion. In such cases, we can rely either on the visual cue, or other slides preceding or following the given foil to decipher the topic.

Thus a topical audio event in our case is defined as the set of contiguous points of time in an audio track where there is spoken evidence for the maximal number of textual phrases listed on a foil. It is detected by a 4-step algorithm that involves (i) word spotting of individual textual words on foils, (ii) consolidation of word matches into matches for text phrases on foils, (iii) grouping of matches to multiple text phrases on foils to identify candidate audio events, and finally, (iv) probabilistic ranking of the candidate events to identify most likely audio events. Each of these operations are described in detail below.

Word spotting for phrasal matching

The localization of points in time where topical phrases are spoken makes use of a word spotting algorithm that is based on a combined word and phonetic-based representation of audio. The details of word and phonetic-based retrieval algorithms are available in [13, 14], so that our discussion of them here will be brief.

We first analyze the audio track using a speech recognition engine (IBM ViaVoiceTM with a large (65,000) word vocabulary) to generate a word transcript. This is then filtered using a language model that imposes some sentence structure through tokenization and part-of-speech tagging. Thus for example, an original query word "growing" will be converted to "grow" during this stage. This is then followed by stop-word removal to prevent excess false positives during retrieval. To account for errors in word boundary detection, word recognition, and out-of-vocabulary words, we also ex-

tract a phone-based representation of the audio to build a time-based phonetic index. The phonetic query representation is generated by converting the original query words to equivalent phone sequences. For example, the original query "growing" is converted to the phone sequence "G-R-OW". Finally, the word and phoneme indexes are represented as a tuple $(w, \{t_w, p_w\})$ and $(s, \{t_s, p_s\})$ respectively, where w, s are the word and phoneme strings, $\{t_w, t_s\}$ are points in time where they occur, and $\{p_w, p_s\}$ are their respective recognition probabilities.

These indexes are now used to perform word spotting as follows. We convert the original query comprising of a sequence of words into word and phonetic representations as described above. We then use the word-based query representation to retrieve an ordered list of matches in time using the word index. We refer to this as word-based retrieval. Next, we use the phonetic query representation to generate an ordered list of matches in time using the phone index. We refer to this as phonetic retrieval. We merge the two ordered lists in time, and compute a combined score for each of the matches. We empirically arrive at a threshold value for the score, and retrieve only those matches with a score above the threshold value.

Our word spotting algorithm has good performance bounds as described in Table 3 which lists the precision and recall numbers for word-based, phonetic and combined word-phoneme retrieval for the experimental data set outlined in Table 2.

Topical phrase detection

We now discuss the localization of points in time where topical phrases extracted from lines of text on foils are heard. The lines of text in electronic foils are extracted automatically using OLE code for Powerpoint or Freelance slides as described in [10]. Matches for each of the words in a phrase (line of text) are obtained using the word-spotting algorithm described above. Although the precision of our word spotting algorithm is good, an individual word may be spoken in multiple contexts so that the matches to individual words of a phrase tend to span wide sections of the audio track. The phrase-based retrieval, therefore, consolidates these matches such that the order of words in the query phrase is preserved in the matching spoken phrases found. That is, given a query phrase sequence $S_Q = (q_1, q_2, \dots, q_n)$, we retrieve matches to individual words q_i by recording the set $\{t_{qij}, p_{qij}\}$ where t_{qij} is the time at which there is a j th match to the i th query word q_i based on the word index or the phonetic index (or both), and p_{qij} is probability of relevance of a match. Here we use a simple linear combination of matching word and phone index term probabilities to arrive at p_{qij} . The resulting sets $\{t_{qij}, p_{qij}\}$ for all query phrase words are then arranged in time-sorted order to form a long match sequence

$$S_M = (s_1, s_2, \dots, s_m) \quad (2)$$

where the i th match $s_i = (q_j, t_{qjk} = t_i, p_{qjk})$ in the combined sequence corresponds to a k th match for some query word q_j and m is the total number of matches to all the query words in the phrase. The best match to the overall query phrase that preserves the order of occurrence of

words is then found by enumerating all common contiguous subsequences $W_q = (w_1, w_2, \dots, w_l)$ of S_M and S_Q . The sequence W_q is a contiguous subsequence of S_M if there exists a strictly increasing sequence (i_1, i_2, \dots, i_l) of indices of S_M such that $w_j = s_{i_j}$ for $j = 1, 2, \dots, l$ and $i_j - i_{j-1} < \tau$. The threshold τ represents average time between two words in a spoken phrase. When the words are consecutive, this is typically of the order of 1 second for most speakers. The probabilities of relevance of each such subsequence is then computed simply as the average of the relevance score for each of its element matches, as the matches to the individual words can be assumed to be mutually exclusive. All those with probabilities of relevance above a chosen threshold are retained as matches to a query phrase. Thus the method of indexing for a phrase exploits the probabilities of relevance of individual matching words in the phrase, as well as their spoken order taking into account the average separation in time between spoken words in a phrase.

topical audio event detection

The above query phrase-based audio retrieval can be repeated for all query phrases on a foil to obtain places in the video where one or more of the query phrases were heard. Due to errors in speech recognition, and due to partial phrase matches, the match positions are again widely distributed potentially spanning the entire video. Figure 2b shows the phrasal match distribution in the audio to the phrases on slide depicted in Figure 2a. Similarly, Figure 3b shows the distribution for the phrases on the query slide shown in Figure 3a. While the individual matches to phrases can be widely distributed, we notice that there are points in time where a number of these matches either co-occur or occur within a short span of time. If such matches could be grouped based on inter-phrasal match distance, then it is likely that at least one such group corresponds to the place in the audio where the topic conveyed by the foil was discussed. *This is the central observation exploited in detecting the topical audio event.* To arrive at a suitable threshold for inter-phrasal match distance, we recorded the match distributions for phrases on over 350 slides and a collection of over 20 videos depicting one or more of the slides covering multiple speakers and course content. We then noted the inter-phrasal match distance difference during the duration over which the topic conveyed by the foil was actually discussed (as noted manually). The resulting distribution of the difference indicated a peak in the distribution between 1 and 20 seconds indicating that for most speakers and most topics, the predominant separation between utterances of phrases tended to be between 1 to 20 seconds apart. We, therefore, used a distance threshold of 20 seconds to group consecutive phrasal matches into time groups using a simple connected component algorithm. During grouping, we allowed multiple occurrences of a match to a phrase within a group to handle cases when a phrase emphasizing a point of discussion was uttered frequently. The resulting time intervals form the basic localization units of the topical event using the audio cue.

Not all such interval groups, however, are relevant to the topical audio event. That is, while it is common for multiple matches to occur for individual topical phrases that look equally good, a discussion containing all the topical

phrases on a given foil are seldom repeated. To compute the probabilities of relevance of the derived interval groups to the topical audio event, we combine the probabilities of individual phrasal matches within the group. Let the topical audio event be denoted by E_a , and let the probability that a time interval group $G_j = (L_j(E_a), H_j(E_a))$ contains E_a be denoted by $P(G_j; E_a)$. Here $L_j(E_a), H_j(E_a)$ are the lower and upper end points of the time interval of the j th match for the topical audio event E_a . Let the time and probability of matches to query phrase qp_i be denoted as $\{(T_{qpij}, P_{qpij})\}$. Since the individual phrase matches within G_j occupy distinct time intervals, the mutual exclusiveness assumption holds, so that $P(G_j; E_a)$ can be assembled as

$$P(G_j; E_a) = \frac{\sum P_{qpr_s}}{\sum_{\text{all } i} \sum_{\text{all } j} P_{qpij}} \quad (3)$$

where the intervals $T_{qpr_s} \in G_j$. The intervals ranked highest using the above probability of relevance then represents the best match to the topical event based on audio information.

Examples

We now illustrate indexing of topical audio events based on foil phrases. Figure 2a and 3a show two topical foils with text phrases representative of the topic of discussion. Figure 2b and 3b show the matches for the all the phrases on each of the slides (here phrase 1 match is indicated in red, (2-green) (3-blue) (4-cyan) (5-yellow) (6-magenta) (7-black), and so on). The result of grouping using the inter-phrase distance threshold of 20 seconds is shown in Figure 2c and 3c respectively, thereby identifying candidate durations for the topical audio event. The topical relevance scores are also listed in these figures. The results on the correctness of topic localization using the most probable match is discussed under Section 5.

4. TOPICAL EVENT DETECTION BY MULTI-MODAL FUSION

We now discuss the detection of the overall topical event using the evidence for the event obtained by visual and audio clues. Since both foil indexing and phrase-based retrieval can have false negatives and positives, they can result in either a video segment being incorrectly weighted for relevancy to a topical event or indicating the same topic at a wrong location. The time co-occurrence of these matches, however, can be a strong clue to the correctness of the detected location for the topic. Simplistic combining methods for multi-modal fusion such as "AND" or "OR" of the intervals do not yield satisfactory solutions. That is, a simple AND of the durations can result in too small a duration to be detected for the overall topic, while an "OR" of the results can potentially span the entire video segment, particularly, when the audio and video matches are spread over the length of the video. Other combination methods such as winner-take-all[7] used in past approaches are also not appropriate here since the probabilities of relevance of durations for events given by neither the audio nor the video matches are particularly salient for clear selection. Finally, weighted linear combination methods are also not appropriate as they do not exploit time co-occurrence.

Number of Videos in Test Collection	6
Total Duration of Videos	4.5 hours
High fidelity Recording with Professional Speaker (35% WER)	1 hour
Low fidelity Recording with Amateur Speaker (65% WER)	3.5 hours
Average Number of In-Vocabulary Queries Per Video	8
Average Number of Matches Per Query	9 (ranges 2 - 43)
Average Query Length	1 word

Table 2: Test data statistics for word spotting experiments.

Retrieval System	Average Precision	Average Recall
Word-Based	0.98	0.59
Phonetic	0.87	0.73
Combined	0.89	0.75

Table 3: Word spotting performance data.

Video	Number of foils	Avg # of phrases	Avg. # of phrase matches	Number of times Correct	Actual Present
1.	13	5.9	4.3	9	13
2.	24	6.8	7.2	18	24
3.	6	7.2	3.8	5	6
4.	11	4.9	4.3	8	11
5.	10	8.3	6.8	7	10
6.	10	8.3	7.9	6	10

Table 4: Precision in audio event detection.

Video	Slide #	Topic duration detected							
		Foil image		foil text		Combined		Ground truth	
		start	end	start	end	start	end	start	end
1.	3	2:02	3:46	2:28	4:08	2:02	3:47	2:25	3:32
2.	6	6:33	7:36	6:29	8:09	6:33	7:36	6:33	7:29
3.	5	12:56	17:44	14:23	15:23	12:56	17:44	12:40	16: 02
4.	3	3:31	5:48	4:20	5:10	3:31	5:48	2:54	6:01
5.	4	16:04	16:49	12:22	14:24	16:04	16:49	15:29	17:01
6.	6	7:45	10:43	6:23	8:24	6:23	10:43	7:01	10:51

Table 5: Accuracy of topical event duration detection.

Video	# slides	Topic Occurrences in top 10 matches			Topic Occurrences in top 3		
		foil image	foil text	combined	foil image	foil text	combined
1.	10	9	10	10	8	6	9
2.	27	22	25	25	24	16	26
3.	32	28	30	31	26	20	28
4.	16	13	12	14	10	8	12
5.	23	16	19	20	14	14	19
6.	18	15	16	16	12	10	14

Table 6: Illustration of precision and recall of topical event indexing of videos using foils

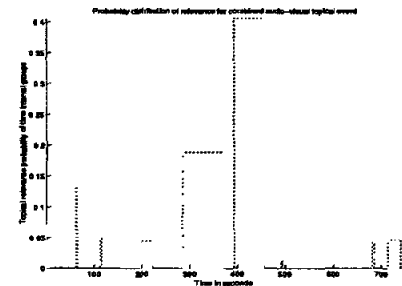
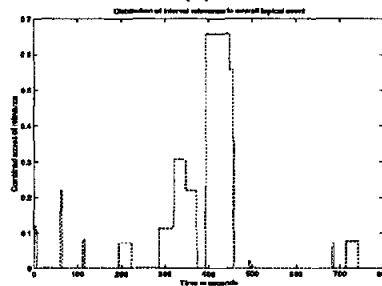
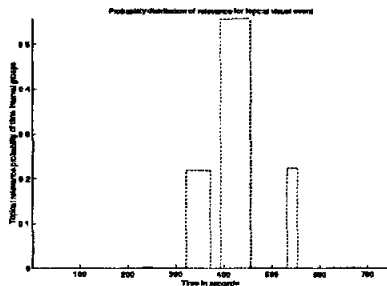
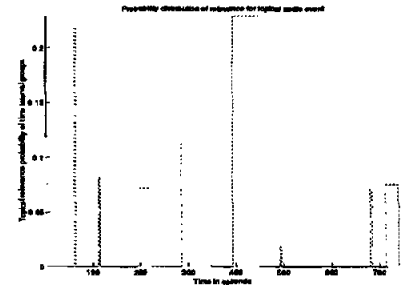
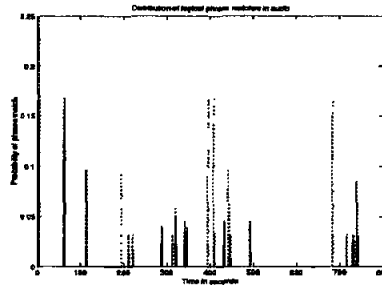
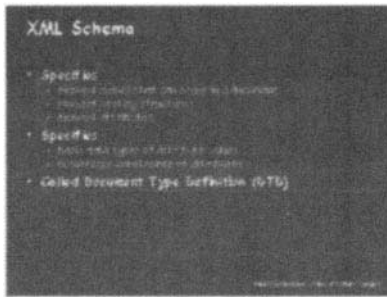


Figure 2: Illustration of topical event detection. (a) A query slide. (b) Phrasal match distribution in the audio track. The colors represent matches for the individual phrases. (c) Candidate audio events and their relevance probabilities obtained by grouping phrasal matches. (d) Topical visual event indicating appearance of foil in video track. (e) Cumulative distribution of combined topical audio and video events. (f) Segmentation of cumulative distribution to indicate combined topical event and their relevance probabilities.

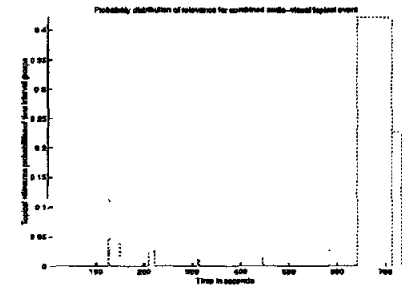
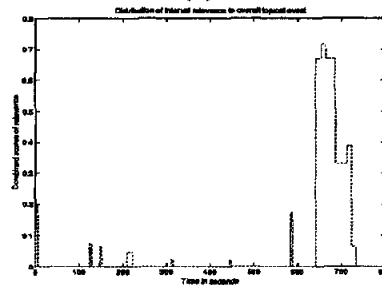
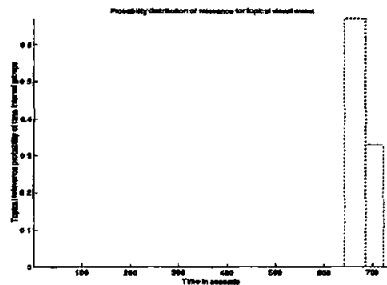
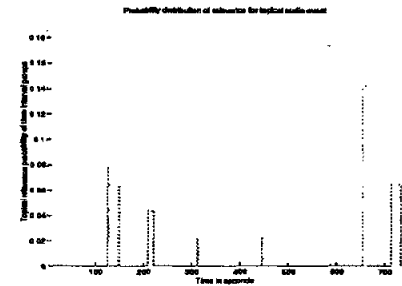
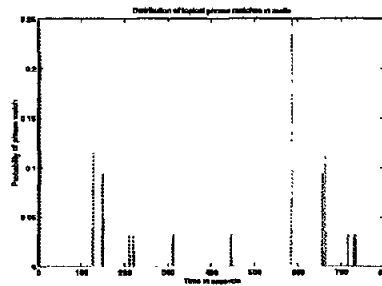
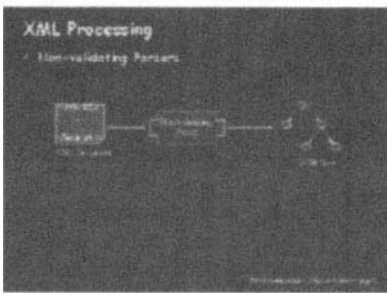


Figure 3: Illustration of topical event detection – Another example. (a) A query slide. (b) Phrasal match distribution in the audio track. The colors represent matches for the individual phrases. (c) Candidate audio events and their relevance probabilities obtained by grouping phrasal matches. (d) Topical visual event indicating appearance of foil in video track. (e) Cumulative distribution of combined topical audio and video events. (f) Segmentation of cumulative distribution to indicate combined topical event and their relevance probabilities.

Our approach to multi-modal fusion is based on the following guiding rationale. (a) The combination method should exploit the time co-occurrence of individual cue-based event detections. (b) The selected duration for the overall topical event must show graceful begin and end to match the natural perception of such events. (c) The combination should exploit the underlying probabilities of relevance of a duration to event given by individual modal matches.

The durations likely to contain the individual cue-based events can be denoted by $\{G_j(E_i), F_j(E_i)\}$, where $G_j(E_i) = (L_j(E_i), H_j(E_i))$ are the the lower and upper end points of the time interval of the j th match in time using the i th modal topical event E_i as defined earlier. In our case, E comprises of $E_1 =$ topical video event where a visual match to a slide appeared in the video stream and $E_2 =$ topical audio event where the text phrases on slide were heard in the audio track² Here $F_j(E_i)$ is the probability of relevance of interval G_j to the event $E_i = P(G_j(E_i)/E_i)$. In general, the probability that the time duration of the entire video contains the overall topical event E is given by

$$P(G; E) = P(G; E_1) + P(G; E_2) - P(G; E_1) * P(G; E_2). \quad (4)$$

To localize the duration most likely to contain the topical event E better, we can split the interval G based on the beginning and end times indicated by the individual modal events. That is, from the intervals $L_j(E_i), H_j(E_i)$, we get the end point sequence e_1, e_2, \dots, e_{2M} where M is the total number of matches to all cues. The sequence of time intervals is then $\delta t_1, \dots, \delta t_{2M-1}$ where $\delta t_i = e_i - e_{i-1}$. The probability that each such interval is likely to contain the event E can now be given by

$$P(\delta t_i; E) = P(\delta t_i; E_1) + P(\delta t_i; E_2) - P(\delta t_i; E_1) * P(\delta t_i; E_2) \quad (5)$$

The individual values $P(\delta t_i; E_k)$ are then given as

$$P(\delta t_i; E_k) = \begin{cases} P(G_j; E_k) & e_i = H_j(E_k) \\ 0 & e_i = L_j(E_k) \end{cases} \quad (6)$$

The distribution of $P(\delta t_i; E)$ is often multimodal (unless all event matches co-occur), as shown in Figure 2e and 3e indicating that the overlap of time intervals of multi-modal matches grows and shrinks in cycles. By noting the local maxima of this distribution, and adjacent local minima around them, we form groups of time co-occurrence intervals (the local minima are those where there is a sign change in the derivative). The minima of the distribution correspond to situations where evidence from a cue disappears to be replaced next by evidence from another cue. These breakpoints are often the places where there is a graceful ending of the topic of discussion, a fact which has also been later verified through our experiments. Each such group δT_i is then taken as an indication of the localization of a match to the queried topical event. The probability of relevance $P(\delta T_i; E)$ is then taken to be simply $\max\{P(\delta t_j; E), \delta t_j \in \delta T_i\}$. This is based on the rationale that once the individual time intervals t_j have been grouped as one time unit for the event, it is sufficient to consider the best evidence for it within the specified time interval.

²Although we do not show here, the method of multi-modal fusion described here can be easily generalized to more than two cues.

5. RESULTS

The indexing of topical events using foils was attempted as part of new distributed learning system that supports search and browse of multimedia documents based on text, image and audio content. The distributed learning system was delivered to a customer, and the following studies resulted during the evaluation phase prior to the delivery of the system.

The best way to view the results is by playing the corresponding indexed video segments for each query topical event indicated by sample foils, and noting the differences in the indicated time location using each of the cues. Since this is not possible in the paper version of the proceedings, we restrict to illustrating the results through the following studies.

Precision in visual event detection

We now report on our results of testing the precision of a match to a topical visual event. We have tested the slide matching technique on a total of 20 classroom videos collected from multiple university and training sources. The number of slides associated with each course varied, with the minimum being 10 and a maximum of 37, giving rise to a database of about 600 slide objects (generated from Powerpoint slides, Freelance graphics slides, and hardcopy slides or lecture notes respectively). For each of the videos, we evaluated the accuracy of foil detection using color, as well as accuracy of recognition in the frames detected to contain foils. The resulting performance is indicated in Table 1. Here we report the results for the duration most likely to contain the visual event as indicated by the verification scores of foil recognition. Note from the table that, for some videos, when a foil is projected for long time, multiple keyframes can indicate the same foil. Also, errors in keyframe extraction can miss the depiction of a foil. The color-based detection method is conservative as can be seen by Column 4 where the number of early detections are always more than the actual number of foils found. It can be concluded from Column 4, 6, and 7 that both detection and recognition of foils in videos is reliable. In practice, though verification errors can leave more than one choice for a matching slide in a video frame, and can also cause some misses, particularly for badly occluded slides, or where zooming and panning effects leave only a small portion of a slide visible.

Precision in audio event detection

The test set to evaluate the precision in audio event detection remained the same as above. Here, however, we used the text phrases on each of the slides for querying the audio content. We then manually recorded the number of times the duration indicated to be most likely to contain the audio event, actually contained such an event. The result is indicated in Table 4. It can be seen from the table that while the most probable duration contained the audio event in most cases, the performance is not as good as for the slide image-based detection.

Precision in duration detection of topical events

Next, we evaluated the precision in the detection of the dura-

tion of the topical event using each of the cues. For this, we chose a set of 40 slide queries and 10 sample videos showing one or more of the slides. We indexed the video using slide image, slide text, and their combination, and in each case, noted the beginning and ending times. The result is shown in Table 5 for a sample of 10 slide queries in two sets of videos. Here rows 1- 5 are edited videos, in which the camera panned to the slide more or less around the time it was starting to be discussed. The second set of videos consisted of unedited videos, videos with single fixed camera and amateur videos taken with a handycam. Here we also record the ground truth beginning and ending times, as obtained by manual verification. From this table, we note that, the duration of the topic was spanned best by combining the two types of searches. The foil image-based indexing was accurate in identifying the beginning of a topic for edited videos, while the duration indicated showed a mismatch for unedited videos. This is understandable since the duration of the topic event indicated by foil image match is the time between two different consecutive slide appearances, which assumes that the camera pans to the slide as soon as it is displayed. Lastly, note that even with combined search, there is still a difference between the automatic and manually detected topic location and duration.

Precision and recall in topical event indexing

To test the precision and recall in topic indexing, we recorded the number of times a match to the topic was indicated in the top 10 results, and the number of times the correct match appeared in the top 3 results for a set of slides per video. The result is indicated in Table 6 for a topic search of a sample of slide queries in the corresponding videos in which they appear. The number of slides used for each video is shown in Column 2. As can be seen, the foil image-based search has fewer false positives, while foil text-based search has fewer false negatives in topic identification. The combined use of both cues, has fewer false positives and negatives.

6. CONCLUSIONS

In this paper we have tried to expose the challenges facing event detection in digital videos by focusing on topical events. In doing so, we have demonstrated a novel application of content-based retrieval of videos for the indexing of topics of discussion in classroom lecture/talk environments.

7. REFERENCES

- [1] G. Abowd et al. Teaching and learning as multimedia authoring: The classroom 2000 project. In *Proc. ACM Multimedia*, pages 104–111, 1996.
- [2] D.E. Appelt et al. Maestro: Conductor of multimedia analysis technologies. *cacm*, 43:57–63, February 2000.
- [3] K. Bharat and M. Henzinger. Improved algorithms for topic distillation in a hyperlinked environment. In *Proc. 22nd Annual SIGIR Conference*, pages 326–327, 1999.
- [4] J.C. Clark and N. Ferrier. Modal control of an attentive vision system. In *Proceedings of the International Conference on Computer Vision*, pages 514–523. 1988.
- [5] G. Hauptmann, D. Lee, and P.E. Kennedy. Topic labeling of multilingual broadcast news in the informedia digital video library. In *Proc. ACM Digital Libraries/SIGIR MIDAS Workshop*, 1999.
- [6] S. Jones and G. Paynter. Topic-based browsing within a digital library using keyphrases. In *Proc. 4th ACM Conference on Digital Libraries*, pages 114–121, 1999.
- [7] C. Koch and S. Ullman. Selecting one among the many: A simple network implementing shifts in selective visual attention. Technical report, Artificial Intelligence Lab, M.I.T., AI-Memo-770, January 1984.
- [8] Y. Lamdan and H.J. Wolfson. Geometric hashing: A general and efficient model-based recognition scheme. In *Proceedings of the International Conference on Computer Vision*, pages 218–249, 1988.
- [9] S. Mukhopadhyay and B. Smith. Passive capturing and structuring of lectures. In *Proc. ACM Multimedia*, pages 477–488, 1999.
- [10] W. Niblack. Slidefinder: A tool for browsing presentation graphics using content-based retrieval. In *Proc. IEEE Workshop on Content-based Access of Image and Video Libraries*, pages 114–118, 1999.
- [11] R. Schwartz et al. A maximum likelihood model for topic classification in broadcast news. In *Proc. European Conf. on Speech Communication and Technology*, 1997.
- [12] J.M. Siskind and Q. Morris. A maximum likelihood approach to visual event classification. In *European Conf. Computer Vision*, pages 347–362, 1996.
- [13] S. Srinivasan et al. Query expansion for imperfect speech: Applications in distributed learning. In *Proc. IEEE Workshop on Content-based Access of Image and Video Libraries (CBAIVL-2000)*, 2000.
- [14] S. Srinivasan and D. Petkovic. Phonetic confusion matrix-based spoken document retrieval. In *Proc. Special Interest Group on Information Retrieval (SIGIR) 2000*, 2000.
- [15] S. Srinivasan, D. Petkovic, and D. Poncelion. Towards robust features for classifying audio in the cuevideo system. In *Proc. ACM Multimedia*, pages 393–400, 1999.
- [16] T. Syeda-Mahmood. Indexing of topics using foils. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 2000.
- [17] T. Syeda-Mahmood, P. Raghavan, and N. Megiddo. Interval hash trees: An efficient index structure for searching object queries in large image databases. In *IEEE Workshop on Content-based Access of Image and Video Libraries*, 2000.
- [18] T.F. Syeda-Mahmood and Y.-Q. Cheng. Indexing colored surfaces in images. In *Proceedings Int. Conf. on Pattern Recognition*, 1996.