Indexing for topics in videos using foils

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

Abstract

A long-standing goal of distance learning has been to provide a quality of learning comparable to the face-to-face envir onment of a traditional classr oom for teaching or training. One of the fundamental problems in achieving this go al is providing effe ctive ways of high-level semantic querying such as for the retrieval of relevant learning material relating to a topic of discussion. In this paper we present a method of identifying video segments relating to a topic of discussion by indexing vide os using the image and text ontent of foils. Specifically, we present a novel method of locating and recognizing foil images in vide o using the color and spatial layout geometry of their regions. We then search the audio asso ciate d with vide o base d on the text content of the foil to identify relate d vide o segments in which concepts represente d on a foil ar e heard. Finally, we combine the results of foil image and text search of vide o exploiting their time co-occurrenc e. The resulting identification of topics is evaluated in the domain of classr oom letur es and talks.

1. In troduction

Despite the progress made in image and video conten t retriev al, making high-lev el semantic queries, such as looking for specific even ts, has still remained a far reac hing goal. Y et, most practical applications em bedding conten t-based retriev al require precisely a way to handle suc h queries. One suc h application is the domain of distributed or distance learning whose long-standing goal has been to provide a quality of learning comparable to the face-to-face en vironment of a traditional classroom for teac hing or training. Effective preparation of online multimedia courses or training material for users is currently ridden with problems of high cost of manual indexing, slow turnaround, and inconsistencies from human interpretation. Automatic methods of cross-linking and indexing multimedia information are very desirable in suc h applications, as they can provide an ability to respond to higher lev el semantic queries, such as for the retriev al of learning material relating to a topic of discussion. Automatic cross-linking multimedia information, how ever, is a non-trivial problem as it requires the detection and identification of even ts whose common threads appear in multiple information modalities. An example of such an ev entare points in a video where a topic was discussed. F rom a surv ey 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 conv ey the topic being discussed at that point in time. Figure 3c,f,i shows examples of such slides. When such lectures are video taped, at least one of the cameras used captures the displa yed 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. How ever,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 remov ed.

In this paper, therefore, we focus on identifying video segments relating to a topic of discussion by indexing videos using both the image and text content of foils. Specifically, w e presen t a method of locating and recognizing foil images in video using the color and spatial lay out geometry of their regions. The detection of slide con taining regions in video frames is made possible by an illumination-invariant description of the background color of foils. The recognition of foils in the selected video regions is then performed using the technique of region hashing. Region hashing is an extension of the principle of geometric hashing. Specifically, region hashing models the spatial lay out of regions through affine intervals. Recognizing or localizing objects then in v olv es finding matching region pairs through hashing of multiple query affine intervals against a database of affine interv als using an efficien t index structure, called the interv alhash tree. The technique of region hashing for 2d and 3d object recognition is being described in a submission to another conference. In this paper,

age recognition and indexing. We also report on the search of associated audio using text content of foils. Since foil text and image-based retrieval can point to different locations in video as pertaining to a topic, we combine the results of these searches using their time co-occurrence. The resulting indexing method is part of a distributed learning system that was recently delivered to customers for indexing and browsing of teaching and training videos.

The detection of topics using foils represents a novel application of content-based retrieval of videos. While very few researchers in computer vision and video retrieval communities have focused on topic-based retrieval of videos based on visual content, there is considerable work, in the text and spoken document retrieval community on the automatic detection of topics based on textual content[6, 3, 5, 9]. Some work has also been done in the multimedia authoring community in terms of synchronization of foils with video using a structured note-taking environment such as Zenpads[1], or using off-line synchronization using image content for constrained camera geometries[8]. However, to our knowledge, no work has yet been done on the detection of topic events using a combination of visual and audio search of foils.

2. Detection of foil images in video

Detecting the presence of foils in a video stream 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 3a,d,g show examples of different slide appearances possible in videos. Since the boundary of the slide need not always be visible in the video frame, simple methods of foil detection such as those attempted in earlier approaches[8] that rely on the rectangular structure of the slide will not be sufficient in unstructured environments. Our approach to foil detection is based on detecting the background color of foils. While some slide backgrounds are textured or shaded, most slides are made with uniform color backgrounds. There is considerable color variation, however, in the appearance of the slide from its original electronic or hardcopy form (see Figure 3). To enable a robust detection of slide using the background color, we develop an illumination-invariant descriptor for colored surfaces, and use it for modeling the background colors of slides.

surfaces

From the image irradiance equation, we can relate the light falling on the image $I(\lambda, r)$ to the physical properties of the scene being imaged as

$$I(\lambda, r) = \rho(\lambda, \mathbf{r})F(\mathbf{r})E(\lambda) \quad (1)$$

where ρ is the surface reflectance function, $F(\mathbf{r})$ is the component that depends on surface geometry, and $E(\lambda)$ is the intensity of the ambient illumination. Here we consider a simpler approximation to the image irradiance equation in which the spectral distribution of the illuminant is assumed to be spatially invariant over the surface.

The surface reflectance and hence the resulting appearance of a surface is determined by the composition as well as the concentration of the pigments of the material constituting the surface. For most surfaces, the composition of the pigments can be considered independent of their concentration so that spectral reflectance $\rho(\lambda, \mathbf{r})$ of the surface can be written as a product of two terms $\rho_1(\lambda)$ and $\rho_2(\mathbf{r})$. Again, since the background regions of slides are uniform color regions, this approximation holds. The resulting image irradiance equation becomes:

$$I(\lambda, r) = \rho_1(\lambda)\rho_2(\mathbf{r})F(\mathbf{r})E(\lambda) = H(r)L(\lambda) \quad (2)$$

Since the spectral reflectance function $\rho_1(\lambda)$ is independent of illumination and spatial distribution of reflectance, it can serve as an illumination-invariant and pose-invariant descriptor. However, this descriptor cannot be directly recovered from the image irradiance equation alone for unknown illumination. An equivalent representation of the spectral reflectance component can be obtained by projection in a suitable color space. In particular, we can filter the image intensity along three color channels, namely, the red, green and blue channels, specified using a triple $(I_j(r), j = 1, 2, 3)$ as[1]

$$I_j(r) = \int_0^\infty I(\lambda, r)h_j(\lambda)d\lambda = H(r)\int_0^\infty L(\lambda)h_j(\lambda)d\lambda \quad (3)$$

where $h_1(\lambda), h_2(\lambda), h_3(\lambda)$ are the transfer functions of the channels. We actually selected a luminance-chrominance color space called the YES space[11] which can be derived from RGB space as $(I_Y(r), I_E(r), I_S(r))$ where $I_Y(r) = \sum_i K_{Yi}I_i(r),$, $I_E(r) = \sum_i K_{Ei}I_i(r)$, and $I_S(r) = \sum_i K_{Si}I_i(r)$, where $(K_{Y1}, K_{Y2}, K_{Y3}) = (0.253, 0.684, 0.063)$, $(K_{E1}, K_{E2}, K_{E3}) = (0.5, -0.5, 0.0)$,

[1] This assumes that the camera has been calibrated so that the filtered values correspond to calorimetric-RGB.

of the color variations across a surface are due to intensity, a certain amount of tolerance to changes in surface geometry due to pose changes, can therefore, be achieved by factoring out the luminance component to form a 2d color space. For lambertian surfaces, the image irradiance clusters in the resulting 2d color space show more or less an elliptic shape and a strong directional component[2]. Such clusters can then be completely specified through their location, spread(size), and orientation by their mean, eigen values and eigen vectorsrespectively. In particular, the direction of the cluster is a ratio of spectral responses, and is an illumination and pose-invariant color descriptor. To show this, we can model the cluster of image irradiances in the 2d color space by a 2D Gaussian. Assuming ergodicity, the covariance matrix of this Gaussian cluster has a single eigen vectorgiven by: $\Gamma = \begin{pmatrix} 1 \\ \frac{[\sum_j k_{Sj} \int_0^\infty L(\lambda) h_j(\lambda) d\lambda]}{[\sum_j k_{Ej} \int_0^\infty L(\lambda) h_j(\lambda) d\lambda]} \end{pmatrix}$. Re-expanding the terms $L(\lambda)$ and making the subsitutions $S_1(\lambda) = \sum_j k_{Ej} h_j(\lambda)$, and $S_2(\lambda) = \sum_j k_{Sj} h_j(\lambda)$, we have

$$\Gamma = \begin{pmatrix} 1 \\ \frac{[\int_0^\infty \rho_2(\lambda) E(\lambda) S_1(\lambda) d\lambda]}{[\int_0^\infty \rho_2(\lambda) E(\lambda) S_2(\lambda) d\lambda]} \end{pmatrix}. \quad (4)$$

Under the coefficient model of sensor response, and assuming sensors of narrow-band sensitivity[4], the following approximation holds:

$$\frac{[\int_0^\infty \rho_2(\lambda) E(\lambda) S_1(\lambda) d\lambda]}{[\int_0^\infty \rho_2(\lambda) E(\lambda) S_2(\lambda) d\lambda]} \approx \frac{[\int_0^\infty \rho_2(\lambda) S_1(\lambda) d\lambda]}{[\int_0^\infty \rho_2(\lambda) S_2(\lambda) d\lambda]} \quad (5)$$

making the direction of the cluster Γ, an illumination and pose-invariant color descriptor.

While the direction of the cluster is independent of object pose and ambient illumination, it can be shown that its location and spread is a function of both pose and illumination. This means that under pose and illumination changes the clusters from different instances of the surface undergo translation and shear but not a change in orientation. This result is still consistent with the observations made by Slater and Healey in [10] under the linear combination of surface reflectance model that indicate that changes of illumination and surface geometry correspond to an affine transformation of the color distribution. **Detecting foils in video**

Using the above result, we generate models of background color by taking sample patches from different

[2]The directionality of the clusters can also be inferred from other physics-based models such as the dichromatic reflection model.

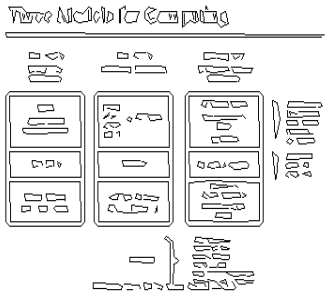


Figure 1: Regions used in foil recognition

.

Video	# foils	video frames	total key-frames	frames		Matches	
				found	Actual	correct	false
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: Performance of foil indexing.

Video	#	Topic in top 10			Topic in top 3		
	slides	image	text	both	image	text	both
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 2: Illustration of precision and recall of topic indexing of videos using foils.

such surface color class is described by a collection of multiple clusters possessing same orientation. Given a new query slide image and a video, we first determine the background color class of the query. For this, we project image pixels into the color space in which the clusters of surface color classes are represented, and assign each pixel to the nearest cluster based on Mahalanobis distance. The class label of each cluster is assigned to the pixels, and connected components of pixels are formed to get the initial regions. These regions are re-projected in the color space, and the orientation of the resulting cluster is verified for a match with the orientation of the surface class whose label was assigned to the region. The first step thus ensures a match of the regions to some appearance of a surface, while the second step ensures a spectral match of the regions. Using this algorithm, multiple assignments of pixels to surface clusters are possible, but these are often eliminated in a later smoothing operation that removes small holes within regions. From the resulting region, the region that encloses all other regions is retained as the background region, and the corresponding surface color class label is then used as the query color class to detect slide containing regions in video. The process of detecting slide containing regions within a video frame is identical to the one described for slide query except that only the pixels belonging to the query color class are retained and processed. 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. For edited and non-edited videos involving zooming and panning from speaker to audience, to projected slides on a screen, the keyframes containing slides usually alternate with other scene keyframes, so that it is sufficient to detect which of the keyframes contain the slides rather than looking at the entire video. To handle videos with fixed camera settings that generate very few shots, we also allow a regular sampling of video (for eg., once per sec) to ensure at least twice as many keyframes as the number of slides used in the talk. The background color detection algorithm, however, runs at frame rate, so that it could potentially be applied to all video frames. Figure 3b,e,h shows background color detection in sample video frames shown in Figure 3a,d,g using the query slides of Figure 3c,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 in the next section.

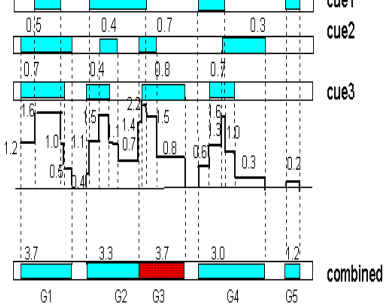


Figure 2: Illustration of method of combining cues by grouping co-occurrence intervals.

2.2 Recognition of foils in video

Background color-based selection can point, in some cases, to multiple regions in a video frame as candidate slide containing regions. Even if one slide region is indicated, it cannot be used to identify which of the foils is depicted in the region, as all foils of a set tend to have the same (slide master) background. Because differences between successive slides can be small (eg. when a topic is continued), 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. 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 distortions. For constrained camera geometries, perspective effects can also be modeled as described in [8]. To model the spatial layout of foil regions, 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[7]. Using this, the relative location of a pair of foil regions can be specified precisely and in a pose-invariant fashion by listing the affine coordinates of features of one region computed with respect to triplets of features from the other region serving as a basis frame. A simpler yet effective way of describing their relative location is through affine intervals, i.e. the interval in which affine coordinate values lie. Since such intervals bound the affine coordinates, they are also affine-invariant for 2d

ally match a pair of 2d object regions will have identical affine intervals. In practice, occlusions and missing features cause the affine intervals to not exactly register (this is particularly common in the case of foils appearing in video). Since occlusions remove the contribution to the affine interval from the lost features, but leave the contribution from the features that are visible, unaffected, the overall effect is to shrink the image affine intervals to become a subset of the corresponding region-pair's affine intervals. Similar observation holds for region segmentation errors. Thus a matching region pair may be conveyed by the presence of large overlap with a region-pair's affine interval. Of course, if the basis features used to compute affine intervals themselves are lost due to occlusions, then the entire affine interval will be found missing. To account for this, we have to compute affine intervals w.r.t multiple basis triples. As observed in geometric hashing, this could cause a large number of affine intervals to be generated. To keep the complexity low, and yet not affect localization accuracy, we choose consecutive or adjacent features to form basis triples. The uniqueness of the affine intervals, however, is not guaranteed, since two different distributions of affine coordinates could be bound by the same interval. The chance of this can be minimized though if we accumulate evidence from multiple object region pairs for common affine intervals. This is similar to the principle of geometric hashing, and is termed region hashing.

Recognizing foils using region hashing

To use region hashing, all the foils 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. Figure 1 shows the detected regions on the slide of Figure 3i. Corners on curves are then used as features. Specifically, we choose consecutive features along curves 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

$$\text{Object layout} = \{(R_i, C(R_i))(R_j, C(R_j)), \{Int_{ij}, B_{ik}\})\} 1 \leq i, j, \leq N \quad (6)$$

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 $(\alpha_{jmin}, \beta_{jmin}), (\alpha_{jmax}, \beta_{jmax}) >$ of features F_j of region R_j computed with respect to kth 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 false matches. The affine interval information is consolidated and represented in an index structure called the interval hash tree. The interval hash tree is a two-way interval tree, i.e., an interval tree on the alpha coordinate is in turn organized as an interval tree on the beta coordinate. The details of the interval hash tree construction and search are reported in the companion submission mentioned earlier and are skipped here, except to point out that it is a balanced binary search tree with the optimal search property that only relevant database intervals are searched in response to a set of query affine intervals, making the search for all query intervals $O(log^2 n + K)$ where n is the number of database intervals, and K is the actual number of them that overlap with query intervals.

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 $F_O = (R_{Oi}, C(R_{Oi}), R_{Oj}, C(R_{Oj}), < Int_{Oij}, B_{Oijm} >)$ after such indexing be denoted by $\{R_k, C(R_k), R_l, C(R_l), I_p, < Int_{kl}, B_{kln} >)\}$. We first discard region pairs if the corresponding region identities do not match i.e., $C(R_k) \neq C(R_{Oi})$ and $C(R_l) \neq C(R_{Oj})$, ovr their overlap is less than a certain threshold. The score of the basis retrieved B_{kln} is then incremented by the extent of interval overlap $\frac{2Int_{kl} \cap Int_{Oij}}{Int_{kl} \cup Int_{Oij}}$. 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.

Results

We now illustrate recognition of foils through some examples. Figure 3a,d,g shows keyframes from sample videos. The recognized slide in each of video frames at the located slide region shown in Figure 3b,e,h are shown in Figure 3c,f,i respectively.

We have tested the slide matching technique on a total of 20 classroom videos collected from multiple

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. 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.

3. Indexing using foil text

Any errors in foil recognition, can either cause portions of video discussing a topic to be entirely missed, or worse still, can point to the wrong point of time. In such cases, and also in cases, where the slide was not displayed, or displayed but not captured by the camera, an audio search of the topic can be useful. Topic search in audio is an area of intense exploration in the spoken document retrieval community [9 5]. In this work, we restrict ourselves to a simple indication of topic by a search of the transcribed audio using keywords listed on a slide. In fact, we used an existing speech recognition engine, and a spoken document retrieval system developed in our lab[2] for our experiments. In particular we extract text content from foils (for Powerpoint or Freelance Graphics foils, we use an OLE code developed in our lab), and use all of the text as query to the audio retrieval system. The audio retrieval system uses an audio index that records the time at which non stopwords were heard in the audio track and weighs them by their inverse document frequency. The audio index is generated using a large vocabulary speech recognition system (65,000 word vocabulary using IBM's Via Voice engine). A further analysis of the transcript is done to impose some sentence structure through tokenization and part-of-speech tagging, and stop words are removed to prevent excess false positives[2]. When this audio retrieval system was applied to the task of of a given query slide, it indicated several points in the video as matches where one or more of the keywords shown on the slide were heard, and these are ranked using the relevancy scores of the corresponding decoded words. The speech recognition errors account for some of the mismatches so that the words on slides need not be actually heard in some of the time intervals indicated.

4. Indexing of topics based on foil image and text

Since both foil indexing and text-based retrieval can have false negatives and positives, they can result in either a video segment being incorrectly weighted for relevancy to a topic 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. The times of co-occurrence of different cues, must be combined though, if accurate detection of beginning and ending times are desired. Our approach to combining search results from multiple cues is based on forming groups of time co-occurrence intervals of individual matches. Each such group is then weighted by the individual scores of relevance. Specifically, consider the matches returned using image and audio search of the video denoted by $\{(L_j(C_i), H_j(C_i), F_j(C_i)\}$, where $L_j(C_i), H_j(C_i)$ are the the lower and upper end points of the time interval of the jth match (in the video tract) using the ith modal cue (image or audio match), and $F_j(C_i$ is the normalized relevance score of the jth match in the ith cue. By forming functions of time from the relevance scores as

$$S_{ij}(t) = \begin{cases} F_j(C_i) & L_j \leq t \leq H_j \\ 0 & otherwise \end{cases} \quad (7)$$

We get a cumulative distribution of relevance scores as:

$$T(t) = \sum_i \sum_j S_{ij}(t) \quad (8)$$

This distribution is multimodal as shown in Figure 2 indicating that the overlap of time intervals of 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-cooccurrence intervals (the local minima are those where there is a sign change in the derivative). Figure 2 shows the groups formed by this operation for time intervals from sample searches. If we denote the interval spanned by G_k as $I_{G_k} = [L_k(G_k), H_k(G_k)]$, then the overlap between the jth match of ith cue, can be given by $O_{jk}(C_i) = I_{G_k} \cap I_{ij} / I_{G_k} \cup I_{ij}$, where $I_{ij} = [L_j(C_i), H_j(C_i)]$. Each

vance score as

$$S(G_k) = \sum_i [F_j(C_i + Ojk(C_i)] \text{ where } I_{ij} \text{ overlaps } I_{G_k} \quad (9)$$

The results of indexing for topics based on combined used of foil text and image content are best illustrated through video examples, in which the topic can be heard in the audio track while a visual of the slide scene appears in the video track. The format of the proceedings, however, does not allow us to represent these results.

5. Evaluation of the foil-based topic detection system

The indexing of topics 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.

Precision in localization of topic

To evaluate the precision in localization of the topic, 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. For audio search, the ending time is based on the size of the audio document chunk, namely, 100 sec. The result is shown in Table 3 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 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 taking 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 difference between the automatic and manually detected topic location and duration.

Precision and recall in topic indexing

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 2 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.

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] A. Amir, S. Srinivasan, D. Ponceleon, and D. Petkovic. Automating indexing of cuevideo for searching and browsing. In *Proc. Special Interest Group on Information Retrieval (SIGIR)*, pages 326–327, 1999.

[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] B.V. Funt and G.D. Finlason. Color constant color indexing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(5):522–529, 1995.

[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] 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.

[8] S. Mukhopadhyay and B. Smith. Passive capturing and structuring of lectures. In *Proc. ACM Multimedia*, pages 477–488, 1999.

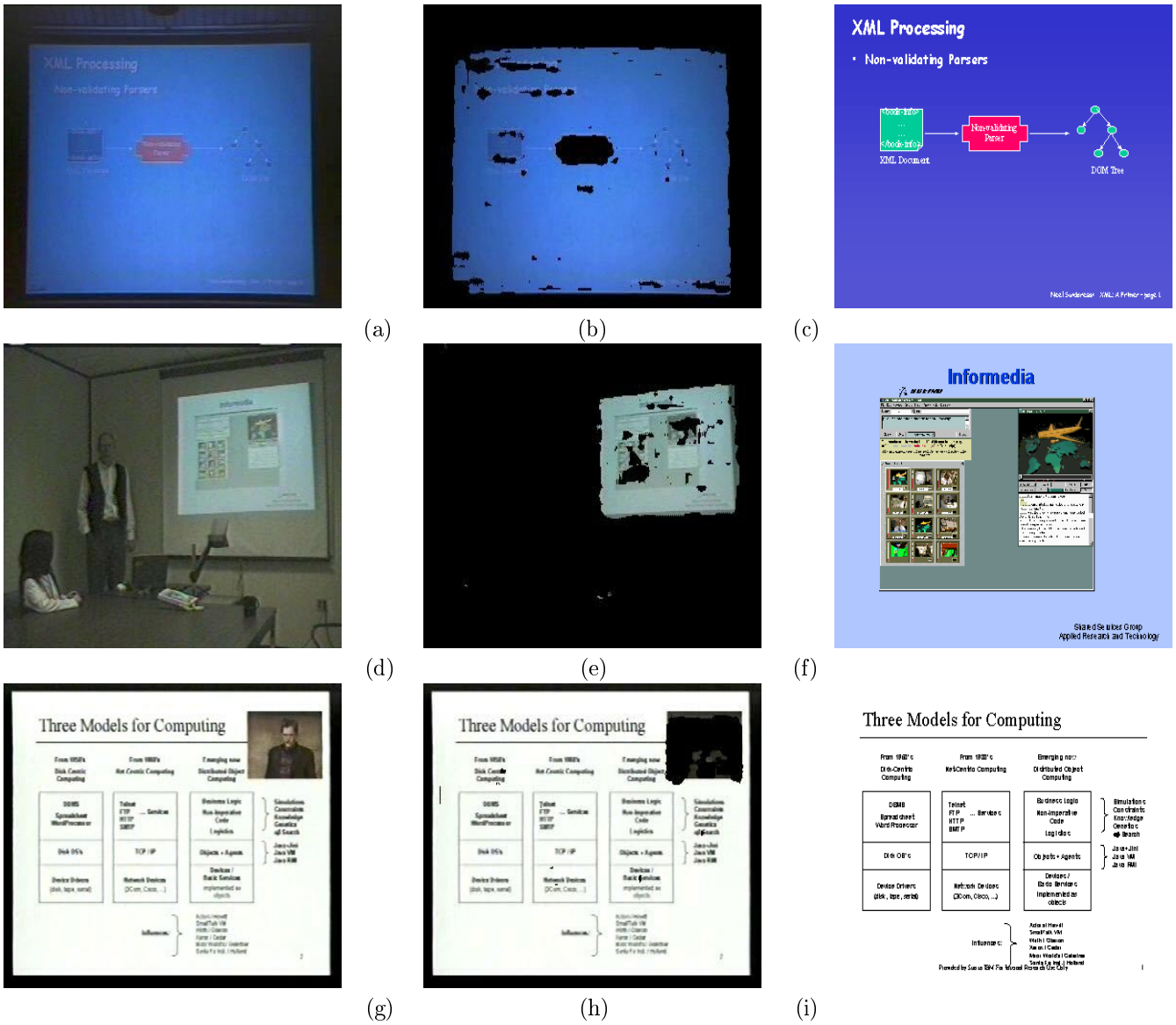
[9] R. Schwartzet al. A maximum likelihood model for topic classification in broadcast news. In *Proc. European Conf. on Speech Communication and Technology*, 1997.

[10] D. Slater and G. Healey. Combining color and geometric information for the illumination-invariant recognition of 3d objects. In *Proceedings of the International Conference on Computer Vision*, pages 563–568, 1995.

[11] Xerox color encoding standard, March 1989.

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

T able 3: P erformance of foil indexing.



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure 3: Illustration of foil detection and recognition. The first column shows the video frames, the second column sho ws theslide-containing regions detected based on bac kground color,and the third column shows the corresponding recognized slide in the detected region.