Video Retrieval with the Informedia Digital Video Library System

Alexander Hauptmann[1], Rong Jin[1], Norman Papernick[1], Dorbin Ng[1], Yanjun Qi[1], Ricky Houghton[2], and Sue Thornton[2]

[1]School of Computer Science,
Carnegie Mellon University
Pittsburgh, PA

[2]Sonic Foundry - MediaSite Systems
Pittsburgh, PA

Background: The Informedia Digital Video Library System.

The Informedia Digital Video Library [1] was the only NSF DLI project focusing specifically on information extraction from video and audio content. Over a terabyte of online data was collected, with automatically generated metadata and indices for retrieving videos from this library. The architecture for the project was based on the premise that real-time constraints on library and associated metadata creation could be relaxed in order to realize increased automation and deeper parsing and indexing for identifying the library contents and breaking it into segments. Library creation was an offline activity, with library exploration by users occurring online and making use of the generated metadata and segmentation.

The goal of the Informedia interface was to enable quick access to relevant information in a digital video library, leveraging from derived metadata and the partitioning of the video into small segments. Figure 1 shows the IDVLS interface following a query. In this figure, a set of results is displayed at the bottom. The display includes a window containing a headline, and a pictorial menu of video segments each represented with a thumbnail image at approximately ¼ resolution of the video in the horizontal and vertical dimensions. The headline window automatically pops up whenever the mouse is positioned over a result item; the headline window for the first result is shown.

IDVLS also supports other ways of navigating and browsing the digital video library. These interface features were essential to deal with the ambiguity of the derived data generated by speech recognition, image processing, and natural language processing. Consider the filmstrip and video playback IDVLS window shown in Figure 2. For this actual video in the IDVLS library, the segmentation process failed, resulting in a thirty-minute segment. This long segment was one of the returned results for the query "Mir collision." The filmstrip in Figure 2 shows that the segment is more than just a story on the Russian space station, but rather begins with a commercial, then the weather, and then coverage of Hong Kong before addressing Mir. By overlaying the filmstrip and video playback windows with match location information, the user can quickly see that matches don't occur until later in the segment, after these other stories that were irrelevant to the query. The match bars are optionally color-coded to specific query words; in Figure 2 "Mir" matches are in red and "collision" matches in purple. When the user moved the mouse over the match bars in the filmstrip, a text window displayed the actual matching word from the transcript or Video OCR metadata for that particular match; "Mir" is shown in one such text window in Figure 2.

By investigating the distribution of match locations on the filmstrip, the user can determine the relevance of the returned result and the location of interest within the segment. The user can click on a match bar to jump directly to that point in the video segment. Hence, clicking the mouse as shown in Figure 2 would start playing the video at this mention of "Mir" with the overhead shot of people at desks. Similarly, IDVLS provided "seek to next match" and "seek to previous match" buttons in the video player allowing the user to quickly jump from one match to the next. In the example of Figure 2, these interface features allowed the user to bypass problems in segmentation and jump directly to the "Mir" story without having to first watch the opening video on other topics.

Figure 1. Text Query and Result Set in the Informedia System.

From the 11 hours of video, we extracted about 8000 shots, where a shotbreak was defined as an edited camera cut, fade or dissolve using standard color histogram measures. Instead of documents, the Video TREC track had defined shots as the unit of retrieval. We aggregated the MPEG I-frames for each shot to be alternative images for each shot. Whenever something matched to an image within a shot, the complete shot was returned as relevant. In total, there were about 80,000 images to be searched.

IDVLS Processing Components:

a. IMAGE PROCESSING

SHOT BREAKS: Color histogram analysis is applied to the MPEG-encoded video. This enables the software to identify editing effects such as cuts that mark shot changes. A single representative frame from each shot is chosen for use in poster frames or in the filmstrip view.

VIDEO OCR: The majority of traditional image processing techniques like optical character recognition (OCR) assume they work with a single image, but image processing for video works with image sequences where each image in the sequence often changes only slightly from the previous image. An overview of the Informedia Project's Video OCR (VOCR) process illustrates these points; Sato et. al discuss VOCR elsewhere in detail [14].

The goal of VOCR was to generate an accurate text representation for text superimposed on video frames. The VOCR process is as follows:

- Identify video frames containing probable text regions, in part through horizontal differential filters with binary thresholding.
- Filter the probable text region across the multiple video frames where that region is identified as containing approximately the same data. This time-based filter improves the quality of the image used as input for OCR processing.
- Use commercial OCR software to process the final filtered image of alphanumeric symbols into text. Optionally improve the text further through the use of dictionaries and thesauri.

The text detection phase can be used in key frame selection heuristics, just as face detection is used. The resulting text from VOCR processing has been used as additional metadata to document the contents of a video segment; this text can be searched just like transcript text.

OCR technology has been commercially available for many years. However, reading the text present in the video stream requires a number of processing steps in addition to the actual character recognition. First the text must be detected. The it must be extracted from the image, and finally converted into a binary black and white representation, since the commercially available OCR engines do not recognize colored text on a variably colored background. Since the extraction and binarization steps are quite noisy and do not produce perfect results, we decided to run the OCR engine on every 3rd frame where text was detected. Thus we obtained over 100 OCR results for a single occurrence of text on the screen that might last for just over 10 seconds. Frequently many of the results would be slightly different from each other, with a very high error rate. On this video collection, the word accuracy for detected text was estimated to be 27%.

FACE DETECTION AND MATCHING: The Informedia system implements the detection of faces in images as described in [2] and face matching through 'eigenfaces'. While we experimented with face recognition using a commercial system [22] as well as an implementation of Eigenfaces [15], the accuracy of face recognition in this type of video collection was so poor, that it proved useless. Therefore, we only used a face detector that reported the presence of faces in each key frame.

IMAGE MATCHING: Color histograms have been widely adopted by many image retrieval systems [5, 6,11] and, they served as the initial image query technique available to IDVLS users. While color histograms were applicable to the broad range of images accessible in the IDVLS library, their use in image indexing and retrieval revealed a number of problems. The histograms did not include any spatial information and hence were prone to false positives. Finally, they were unsuited for retrieving images in finer granularities, e.g., particular colors or regions. Referring to Figure 3, a user looking for a shot of grasslands could instead have retrieved these assorted images of predominantly blue and green colors.

b. Audio Processing

SPEECH RECOGNITION:

The audio processing component of our video retrieval system splits the audio track from the MPEG-1 encoded video file, and decodes the audio and downsamples it to 16kHz, 16bit samples. These samples are then passed to a speech recognizer. The speech recognition system we used for these experiments is a state-of-the-art large vocabulary, speaker independent speech recognizer [18]. For the purposes of this evaluation, a 64000-word language model derived from a large corpus of broadcast news transcripts was used. Previous experiments had shown the word error rate on this type of mixed documentary-style data with frequent overlap of music and speech to be just over 30%.

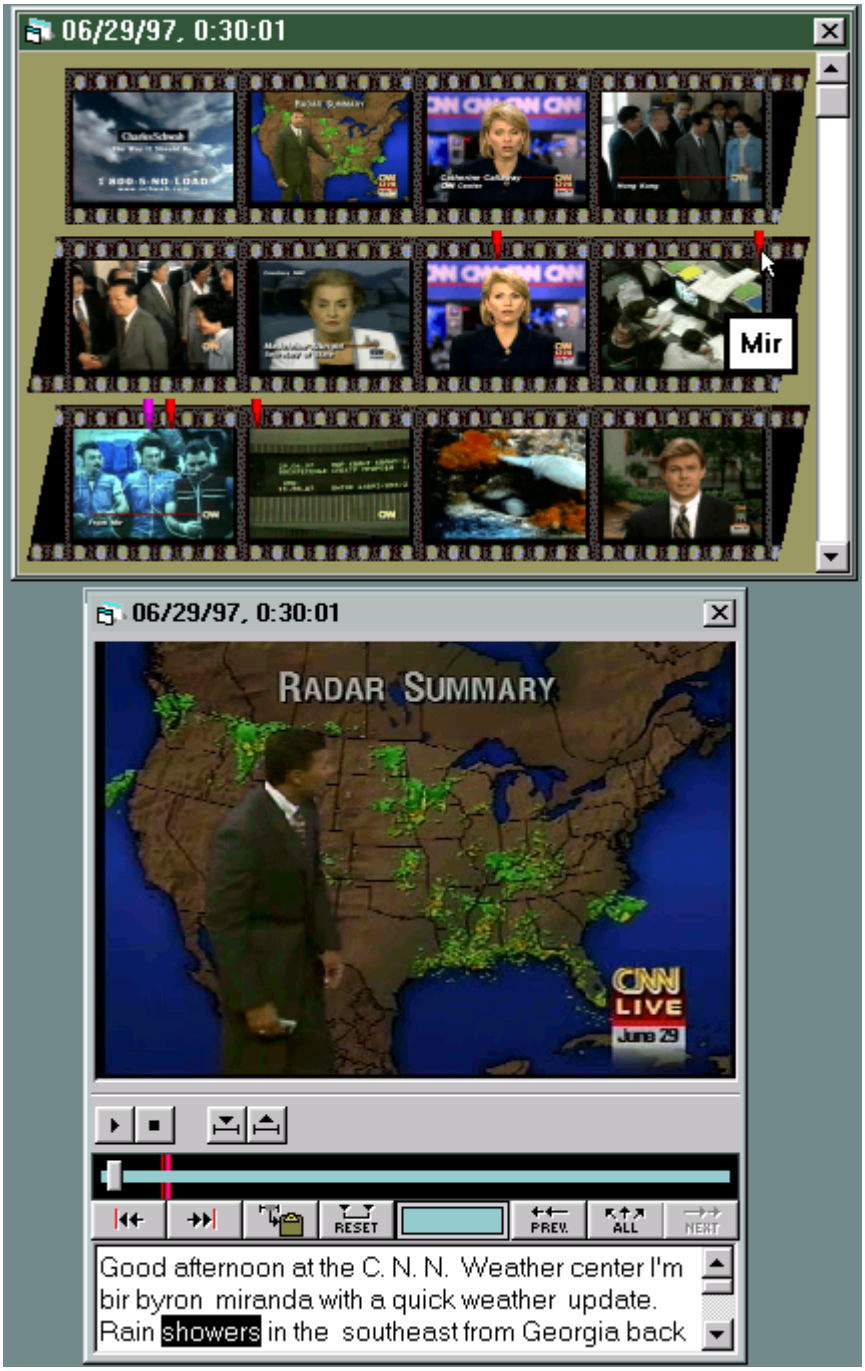

Figure 2. Marking the filmstrip of a query with word location in the transcript and OCR.

SPEAKER IDENTIFICATION

Our speaker identification technology is based on standard Gaussian mixture models as defined by [9]. We use the segmented method with multiple training samples derived from chunks of audio that are 30 seconds in duration. We use weighted rank scoring as defined by Markov and Nakagawa [10].

c. **Text Analysis**

Titles: Segments are scanned for words that have a high inverse document frequency, and that are strongly distinguishing segments. All text is indexed and searchable. All retrieval of textual material was done using the OKAPI formula [13]. The exact formula for the Okapi method is shown in Equation (1)

$$Sim(Q,D) = \sum_{qw \in Q} \left\{ \frac{tf(qw,D)\log(\frac{N - df(qw) + 0.5}{df(qw) + 0.5})}{0.5 + 1.5\frac{|D|}{avg\_dl} + tf(qw,D)} \right\} \qquad (1)$$

where *tf*(*qw*,*D*) is the term frequency of word *qw* in document *D*, *df*(*qw*) is the document frequency for the word *qw* and *avg_dl* is the average document length for all the documents in the collection.

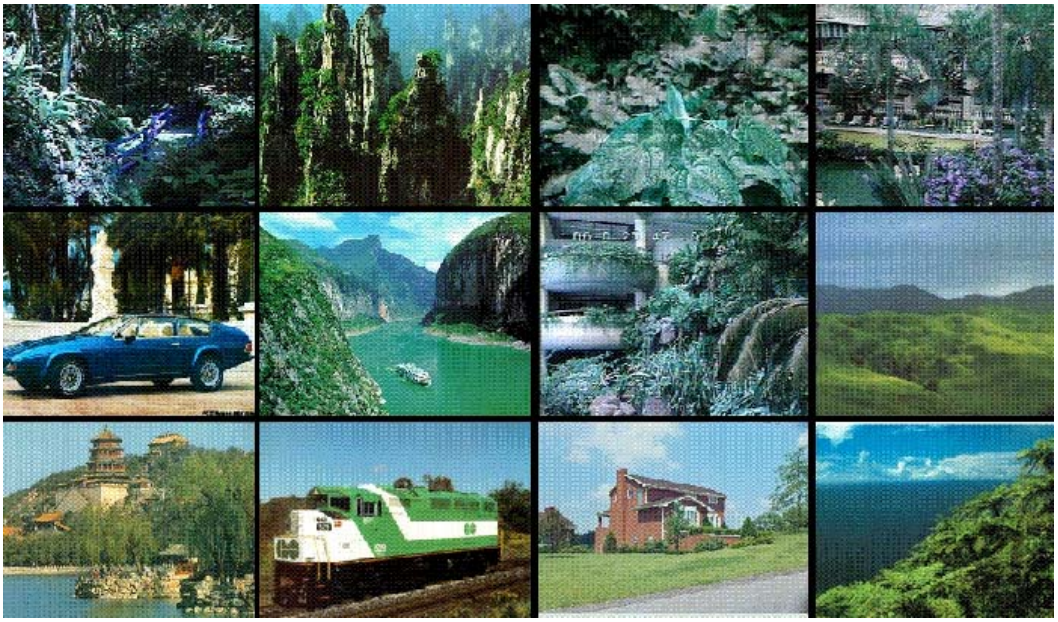

Figure 3. Result of a color histogram search

Approach to the TREC Video Track

Our approach to the Video Track in Trec was to use the Informedia system with only minor changes and see how well it would work. We treated general information queries the same as known item queries. Specific modifications are discussed in the sections for the interactive and automatic system. For simplicity, we always assumed that the unit of retrieval was a single shot.

The Interactive Retrieval System

Since Informedia only uses static images for image matching, we decided make up for this shortcoming by utilizing multiple image search engines:

- Histo144v50 Image Search

Histo144v50 is based on a simple color histogram of the target image. First, the image is converted to the Munsell color space. We are using the Munsell Color space as described in [8]. The hue is isolated. Miyahara and Yoshida describe using Godlove's formula to represent the perceptual distance between some colors in the HVC space. We are using Euclidian distance to approximate Godlove's formula. The image is broken into 9 equally sized regions. A 16-bin histogram is taken of each region. The histograms are appended to each other to form a 144 dimensional vector. The vector is then reduced in dimensionality to 50 by multiplying with a previously computed singular value decomposition. Each vector is then placed in a tree data structure that allows K-nearest-neighbors searches.

- MCPv50 Image Search

MCPv50 computes the color and texture of the target image. The image is broken into 9 equally sized regions. A 15-bin histogram is taken for the Red, Green, and Blue. Then, six texture histograms of 15 bins each are taken. All of these vectors are append to make a 1215 dim vector. This vector is reduced to 50 dim by multiplying with a previously computed singular value decomposition. Each vector is than placed in a tree data structure that allows K-nearest-neighbors searches.

- Cuebik Image Search

Cuebik is based on one of the behaviors of the IBM QBIC image search engine. A palette of 255 colors is chosen for a database by marking the strongest colors found in a large sample of images. The target image is reduced to 256 equally sized regions. Each region is mapped to one of the palette colors, and recorded. A search is done by choosing a set of regions and finding all images that have the same color in the same region.

In addition to the above image search engines, we also used a downloadable version of the original IBM QBIC system as well as a search engine provided by James Wang from the University of Pennsylvania.

The search process foe each interactive query was as follows:

1. Determine key words in the text description of the query and use Video OCR text search to find them.
2. Use the supplied query images to initiate a search for relevant segments.
3. If a segment key frame or title looks related to the answer, open up its filmstrip and view details.
4. If the segment filmstrip looks related to the topic, but does not provide an answer, look one segment forward and back. If the topic in the adjacent segment is the same, scan the filmstrip of an additional segment forward or back.
5. If a frame answers the query, use that frame for relevance feedback with each of the image search engines to find more like it.
6. If a frame seems to be related, but does not answer the question, use that frame with each of the search engines to find more like it.
7. Repeat all steps as needed.

Automatic Retrieval

In the following we will elaborate only on the known item query set, because comprehensive relevance judgments were available for this set allowing automatic estimation of precision and recall for variations of our video retrieval system. The 34 known item queries are distinguished from the remaining 'general search' queries in that the information need tends to be more focused and all instances of query-relevant items in the corpus are known. This allows an experimental comparison of systems without the need for further human evaluations.

Since the evaluation could be done automatically, the top 100 search results were scored for all systems. The general unit of retrieval was a 'shot', in other words a time range between two shot changes, for

```
<videoTopic num="005" interactive="N-I" automatic="Y-A" knownItems="Y-K">
  <textDescription text="Scenes that show water skiing"/>
  <videoExample src="BOR17.MPG"  start="0h01m08s" stop="0h01m18s"/>
</videoTopic>
```

Figure 3. A sample known-item query in the automatic condition.

assuming that if two images are similar, their underlying generation models should also be similar, we can compute the similarity of image I_1 to image I_2 as $P(I_1 \mid M_2)$, i.e. the probability of generating image I_1 from the statistical model M_2. Preliminary experiments had shown that this model is more effective for image retrieval from the Video TREC collection than some of the traditional vector methods working on extracted features like e.g. QBIC [6,11].

Automatically Combining Metadata

When the various sources of data were combined for information retrieval, we used a linear interpolation with very high weights on the binary features such as face detection or speaker identification. This allowed these features to function as almost binary filters instead of being considered more or less equal to OCR, speech transcripts or image retrieval.

Experimental Results for the Automatic System

Evaluation Metrics

There are two aspects involved in any retrieval evaluation:

- **Recall**. A good retrieval system should retrieve as many relevant items as possible.
- **Precision**. A good retrieval system should only retrieve relevant items.

Many evaluation metrics have been used in information retrieval [21] to balance these two aspects. In the video retrieval track at TREC, a simple measure of precision at 100 items retrieved was used for scoring the systems. However, since there were only an average of 5.5 items relevant for each query, a perfect retrieval system that returned all relevant items at the top and filled the rest of the top 100 result slots with irrelevant items would only achieve a precision of 5.5 %.

Because our collection contains only small numbers of relevant items, we adopted the average reciprocal rank (ARR) [23] as our evaluation metric, similar the TREC Question Answering Track. ARR is defined as follows:

For a given query, there are a total of N_r items in the collection that are relevant to this query. Assume that the system only retrieves k relevant items and they are ranked as $r_1, r_2, \ldots, r_k$. Then, the average reciprocal rank is computed as

$$ARR = \left\{ \sum_{i=1}^{k} i / r_i \right\} / N_r \qquad (1)$$

As shown in Equation (1), there are two interesting aspects of the metric: first, it rewards the systems that put the relevant items near the top of the retrieval list and punish those that add relevant items near the bottom of the list. Secondly, the score is divided by the total number of relevant items for a given query. Since queries with more answer items are much easier than those with only a few answer items, this factor will balance the difficulty of queries and avoid the predominance of easy queries.

Table 1. Results of video retrieval for each type of extracted data and combinations.

Retrieval using:	**Average Reciprocal Rank**	**Recall**
Speech Recognition Transcripts only	1.84 %	13.2 %
Raw Video OCR only	5.21 %	6.10 %
Raw Video OCR + Speech Transcripts	6.36 %	19.30 %
Enhanced VOCR with dictionary post-processing	5.93 %%	7.52 %
Speech Transcripts + Enhanced Video OCR	7.07 %	20.74 %
Image Retrieval only using a probabilistic Model	14.99 %	24.45 %
Image Retrieval + Speech Transcripts	14.99 %	24.45 %
Image Retrieval + Face Detection	15.04 %	25.08 %
Image Retrieval + Raw VOCR	17.34 %	26.95 %
Image Retrieval + Enhanced VOCR	18.90 %	28.52 %
Image Retrieval + Face Detection + Enhanced VOCR	18.90 %	28.52 %
Image Retrieval + Speech Transcripts + Enhanced VOCR	18.90 %	28.52 %
Image Retrieval + Face Detection + Speech Transcripts +Enhanced VOCR	18.90 %	28.52 %

Results for Individual Types of Metadata

The results are shown in Table 1. The average reciprocal rank (ARR) and recall for retrieval using only the speech recognition transcripts was 1.84% with a recall of 13.2%. Since the queries were designed for video documents, it is perhaps not too surprising that information retrieval using only the OCR transcripts show much higher retrieval effectiveness to an ARR of 5.21% (6.10% recall). The effects of post-processing on the OCR data were beneficial, the dictionary-based OCR post-processing gave a more than 10% boost to 5.93 % ARR and 7.52 % recall. Again, perhaps not too surprisingly, the image retrieval component obtained the best individual result with an ARR of 14.99 % and recall of 24.45 %. Since the face detection could only provide a binary score in the results, we only evaluated its effect in combination with other metadata.

Results When Combining Metadata

Combining the OCR and the speech transcripts gave an increase in ARR and recall at 6.36 % and 19.30 % respectively. Again post-processing of the OCR improved performance to 7.07 % ARR and 20.74 % recall. Combining speech transcripts and image retrieval showed no gain over video retrieval with just images (14.88 % ARR, 24.45 % recall). However, when face detection was combined with image retrieval, a slight improvement was observed (15.04 % ARR, 25.08 % recall).

Combining OCR and image retrieval yielded the biggest jump in accuracy to an ARR of 17.34 % and recall of 26.95 % for raw VOCR and to an ARR of 18.90 % and recall of 28.52 % for enhanced VOCR. Further combinations of image retrieval and enhanced OCR with faces, and speech transcripts yielded no additional improvement. The probably cause for this lack of improvement is the redundancy to the other extracted metadata.

Discussion

What have learned from this first evaluation of video information retrieval? Perhaps it is not too surprising that the results indicate that image retrieval was the single biggest factor in video retrieval for this evaluation. Good image retrieval was the key to good performance in this evaluation, which is consistent with the intuition that video retrieval depends on finding good video images when given queries that include images or video.

One somewhat surprising finding was that the speech recognition transcripts played a relatively minimal role in video retrieval for the known-item queries in our task. This may be explained by the fact that discussions among the track organizers and participants prior to the evaluation emphasized the importance of a video retrieval task as opposed to ‘spoken document retrieval with pictures’.

There was a strong contribution of the OCR data to the final results. The results also underscore the fact that video contains information not available in the audio track. As a previous study noted, only about 50% of the words that appear as written text in the video are also spoken in the audio track [14], so the information contained in the text of the pictures is not redundant to the spoken words in the transcripts. Overall, the queries presented a very challenging task for an automatic system. While the overall ARR and recall numbers seem small it should be noted that about one third of the queries were unanswerable by any of the automatic systems participating in the Video Retrieval Track. Thus for these queries nothing relevant was returned by any method or system.

We would like to caution that the known-item queries do not represent a complete sample of video queries. Video retrieval on general search queries, with less specific information needs, might result in a somewhat different conclusion about the combination of information sources. A preliminary analysis showed that ‘general search’ queries in the video track tended to be much more ‘speech oriented’, which is why the best performing system on that set of queries was entirely based on speech recognition transcripts.

Clearly, we can think of a number of improvements to the speech recognition component, using a parallel corpus for document and query expansion, and relevance feedback. However, the same techniques could be used to improve the OCR transcriptions as well.

References

[1] Wactlar, H.D., Christel, M.G., Gong, Y., and Hauptmann, A.G. “Lessons Learned from the Creation and Deployment of a Terabyte Digital Video Library”, *IEEE Computer* **32**(2): 66-73.

[2] Rowley, H., Baluja, S., and Kanade, T. "Human face detection in visual scenes", CMU, 1995. Technical Report CMU-CS-95-158.

[3] *Informedia Digital Video Library Project Web Site*. Carnegie Mellon University, Pittsburgh, PA, USA. URL http://www.informedia.cs.cmu.edu

[4] Hauptmann, A.G., Witbrock, M.J. and Christel, M.G. Artificial Intelligence Techniques in the Interface to a Digital Video Library, *Extended Abstracts of the ACM CHI'97 Conference on Human Factors in Computing Systems,* (New Orleans LA, March 1997), 2-3.

[5] Gong, Y. *Intelligent Image Databases: Toward Advanced Image Retrieval*. Kluwer Academic Publishers: Hingham, MA.

[6] Faloutsos, C., Barber, R., Flickner, M., Hafner, J., Niblack, W., Petkovic, D., Equitz, W. Efficient and Effective Querying by Image Content. *Journal of Intelligent Information Systems* 3(3/4), 231-262.

[7] Christel, M., Winkler, D., and Taylor, R. Multimedia Abstractions for a Digital Video Library. ACM Digital Libraries '97 (Philadelphia, PA, July 1997).

[8] M. Miyahara and Y. Yoshida, "Mathmatical transform of (R,G,B) color data to Munsell (H,V,C) color data", SPIE Vol 1001 "Visual Communications and Image Processing '88", pp 650-7. 1988.

[9] H. Gish and M. Schmidt, Text-Independent Speaker Identification, IEEE Signal Processing Magazine, October 1994, pages 18 – 32.

[10] K. Markov and S. Nakagawa, Frame Level Likelihood Normalization For Text-Independent Speaker Identification using Gaussian Mixture Models, ICSLP 96 1764-1767, 1996.

[11] Hafner, J. Sawhney, H.S. Equitz, W. Flickner, M. and Niblack, W. "Efficient Color Histogram Indexing for Quadratic Form Distance," IEEE Trans. Pattern Analysis and Machine Intelligence, 17(7), pp. 729-736, July, 1995.

[12] Kantor, P. and Voorhees E.M, Report on the Confusion Track, in Voorhees E.M, Harman, D.K., (eds.) "The Fifth Text Retrieval Conference, (TREC-5) 1997.

[13] Robertson S.E., et al.. Okapi at TREC-4. In The Fourth Text Retrieval Conference (TREC-4). 1993.

[14] Sato, T., Kanade, T., Hughes, E., and Smith, M. Video OCR for Digital News Archive. In *Proc. Workshop on Content-Based Access of Image and Video Databases.* (Los Alamitos, CA, Jan 1998), 52-60.

[15] Satoh, S., and Kanade, T. NAME-IT: Association of Face and Name in Video. *IEEE CVPR97,* Puerto Rico, 1997.

[16] Schmidt, M., Golden, J., and Gish, H. "GMM sample statistic log-likelihoods for text-independent speaker recognition," *Eurospeech-9*, Rhodes, Greece, September 1997, pp.855 - 858.

[17] Schneiderman, H. and Kanade, T. Probabilistic Modeling of Local Appearance and Spatial Relationships of Object Recognition, *IEEE CVPR*, Santa Barbara, 1998

[18] Singh, R., Seltzer, M.L., Raj, B., and Stern, R.M. "Speech in Noisy Environments: Robust Automatic Segmentation, Feature Extraction, and Hypothesis Combination," *IEEE Conference on Acoustics, Speech and Signal Processing*, Salt Lake City, UT, May, 2001.

[19] A.W.M. Smeulders, M. Worring, S. Santini, A. Gupta and R. Jain, "Content-Based Image Retrieval at the End of the Early Years," IEEE Trans. Pattern Analysis and Machine Intelligence, 22(12), pp. 1349-1380, December, 2000.

[20] Swain M.J. and Ballard, B.H. "Color Indexing," Int'l J. Computer Vision, vol. 7, no. 1, pp. 11-32, 1991.

[21] Tague-Sutcliffe, J.M., "The Pragmatics of Information Retrieval Experimentation, revised," Information Processing and Management, 28, 467-490, 1992.

[22] Visionics Corporate Web Site, FaceIt Developer Kit Software, http://www.visionics.com, 2002.

[23] Voorhees E.M, and Tice, D.M., "The TREC-8 Question Answering Track Report," The Eighth Text Retrieval Conference (TREC-8), 2000