Video Retrieval using Speech and Image Information

Alexander G. Hauptmann, Rong Jin, and Tobun D. Ng
School of Computer Science,
Carnegie Mellon University
Pittsburgh, PA

ABSTRACT

Video contains multiple types of audio and visual information, which are difficult to extract, combine or trade-off in general video information retrieval. This paper provides an evaluation on the effects of different types of information used for video retrieval from a video collection. A number of different sources of information are present in most typical broadcast video collections and can be exploited for information retrieval. We will discuss the contributions of automatically recognized speech transcripts, image similarity matching, face detection and video OCR in the contexts of experiments performed as part of 2001 TREC Video Retrieval Track evaluation performed by the National Institute of Standards and Technology. For the queries used in this evaluation, image matching and video OCR proved to be the deciding aspects of video information retrieval.

Keywords: Video Search and Retrieval, Video Indexing, Multimedia Information Retrieval

1. INTRODUCTION: INFORMATION RETRIEVAL FROM VIDEO CONTENT

Video is a rich source of information, with aspects of content available both visually and acoustically. Until now, there has never been a large-scale, standardized evaluation of video information retrieval. This paper tries to carefully analyze and contrastively evaluate and compare different types of video and audio information as used in a video information retrieval task. While there have been no serious studies of automatic video information retrieval to date, some components of video information have been examined in the context of information retrieval, most notably spoken document retrieval, image retrieval and OCR.

Spoken Document Retrieval: A textual representation of the audio content from a video can be obtained through automatic speech recognition. Information retrieval from speech recognition transcripts has received quite a bit of attention in recent years in the spoken document retrieval track at TREC7, TREC 8 and TREC 9. The current 'consensus' from a number of published experiments in this area is that as long as speech recognition has a word error rate better than 35% word error, then information retrieval from the transcripts of spoken documents is only 3-10% worse than information retrieval on perfect text transcriptions of the same documents.

Image Similarity Matching. Example-based image retrieval task has been studied for many years. The task requires the image search engine to find the set of images from a given image collection that is similar to the given query image. Traditional methods for content-based image retrieval are based on a vector model [1, 11]. These methods represent an image as a set of features and the difference between two images is measured through a (usually Euclidean) distance between their feature vectors. While there have been no large-scale, standardized evaluations of image retrieval systems, most image retrieval systems are based on features such as color, texture, and shape that are extracted from the image pixels [10].

OCR document retrieval: A different, textual, representation is derived by reading the text that present in the video images using optical character recognition (OCR). At TREC 5, experiments have shown that information retrieval on documents recognized through OCR with a character error rate of 5% and 20% degrades IR effectiveness by 10 % to 50 % depending on the metric, when compared to perfect text retrieval [2].

In contrast, video information retrieval much more complex and combines elements of spoken documents, OCR documents, image similarity as well as other audio and image features. In this paper we will examine the effects of multi-modal information retrieval from video documents. There are only area of audio analysis that we examined was automatic speech recognition. While analyzing the video imagery, we considered the color similarity of images, and the presence of faces and text that was readable on the screen. We explored these dimensions of audio analysis and image analysis separately and in combination in our video retrieval experiments. We will present experiments with each different types of extracted metadata performed separately and also combined together in the context of the TREC Video Retrieval evaluation performed by the National Institute of Standards and Technology.

The remainder of the paper is structured as follows: Section 2 describes the video retrieval evaluation task in more detail and section 3 introduces the Informedia Digital Video Library System and its methods to extract and retrieve metadata, namely speech transcripts, video OCR, as well as image-based metadata extraction and retrieval used for face detection and image similarity matching. The results are presented in section 4 for individual and combined metadata. Finally, section 5 concludes with an analysis of the implications of these results.

2. THE TREC VIDEO RETRIEVAL EVALUATION

The Text REtrieval Conference (TREC) has sponsored contrastive evaluations of information retrieval systems for the last 10 years. While most of the evaluations were concerned with text retrieval, there have also been evaluations of document collections with OCR errors and spoken document collections that include speech recognition errors. In 2001 the first video information retrieval evaluation was performed. The 2001 TREC Video Retrieval evaluation made a corpus of 11 hours of MPEG-1 encoded broadcast video available to all participants. The data consisted of NIST project and promotional videos, documentary material from NASA and the Bureau of Reclamation and Land Management, a series of lectures, as well as BBC stock footage. While both an interactive and an automatic version of the evaluation was performed, we will only report experiments with a fully automatic system, since the user influence in the interactive systems could not easily be factored out.

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. We used 34 known item queries that 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. An automatic known-item query had, on average, just over 2 relevant video clips as answers, with the largest answer set containing 10 relevant items. Three queries contained only text descriptions (suitable for speech or OCR analysis), 19 known-item queries had example still images, and 3 queries listed audio as a specific source of information. 16 of the known-item queries included at least 1 example

```
<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 1. A sample Video TREC query asking for a general scene containing water skiers

video clip. Many queries provided a combination of video examples, still images and/or audio.

The TREC video queries might be classified into the following types, based on what the query was looking for:

- Specific scenes or objects. Some queries searched for specific objects or scenes, such as the statue of liberty, a space shuttle or a rocket launch, a lunar vehicle, corn on the cob, etc.
- Less specific scenes. These include queries that look for a pink flower, a waterskiier, people on a beach, a yellow boat, traffic scenes, water, etc.
- Shots of specific people: Queries of this type would look for a person by name e.g. Harry Hertz, Lynn Bondurant, Ronald Reagan, or a specific unnamed person by example, such as a query looking for other instances of the person provided in the sample video clip or still image)
- Camera operation: Queries would specify a camera operation (pan, tilt and/or zoom), often in addition to a scene such as a zoom-in of canyons, or a pan over grasslands.

Queries sometimes combined elements of several query types, e.g. looking for a person standing in front of the X-29 experimental plane. Queries could also be classified by type of example data or by the feature class, which would be able to find the query:

- The text spoken during the clip ('space shuttle')
- The OCR on the screen during the clip ("Harry Hertz")
- By similarity to the non-text audio characteristics to the samples given in video or audio ('male speaker', water sounds, specific speaker)
- By similarity to the whole example still image
- By similarity to whole still images extracted from the example video clip
- By similarity to the motion in the sample video clip
- By similarity to specific objects somewhere in the example video or still images (lunar rover, corn on the cob).

While the latter classification seems most desirable, it is frequently not possible prior to retrieval to know which aspects of the query might be appropriate for retrieval. Sometimes the text description gave a hint to the human reader, but parsing the query accurately proved to be beyond the scope of our system.

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 example editing cuts and fades. Systems had to determine shot changes automatically. An item was considered relevant if at least 33% of the length of the returned item overlapped with the target item in the list of shots relevant to this query and less than 33% of the time range for the returned item was outside the target range. This requirement ensured a reasonable overlap of the returned shot with the target shot [13]. An example of a typical query is shown in Figure 1. This query is to be used for automatic systems, but not for interactive evaluations. It is a known item query, indicating that all results are known inside the video collection. According to the text description, the query is looking for video scenes of water skiing, and gives an example of the type of video that is desired. A few frames from the example video are extracted and also shown in Figure 1.

```
<videoTopic num="022"  interactive="Y-I" automatic="Y-A" knownItems="Y-K">
  <textDescription text="Find pictures of Harry Hertz, Director of the National Quality Program, NIST" />
  <!-- imageExample src="http://www.quality.nist.gov/nqpstaff/harry.jpg"-->
</videoTopic>
```


Figure 2: A sample Video TREC query with a still image as example data asking for a specific person

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.

3. THE INFORMEDIA DIGITAL VIDEO LIBRARY SYSTEM.

The Informedia Digital Video Library [19] was the only NSF DLI-1 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.

The Informedia research challenge was how much can the video and audio be analyzed automatically and then made to be useful to a user. Broadly speaking, the Informedia project wants to enable search and discovery in the video medium, similar to what is widely available for text. One prerequisite for achieving this goal is the automated information extraction and metadata creation from digitized video. Once the metadata has been extracted, the system enables full-content search and retrieval from spoken language and visual documents. The approach that was ultimately successful was the integration of speech, image and natural language understanding for library creation and exploration. While much of the Informedia project has focused on interactive tools and techniques [18] for finding relevant video clips in a large digital video collection, this paper will discuss the automated processing and retrieval techniques implemented in Informedia.

Table 1 A summary of different TREC Video Queries for both Known-Item and General Search Queries

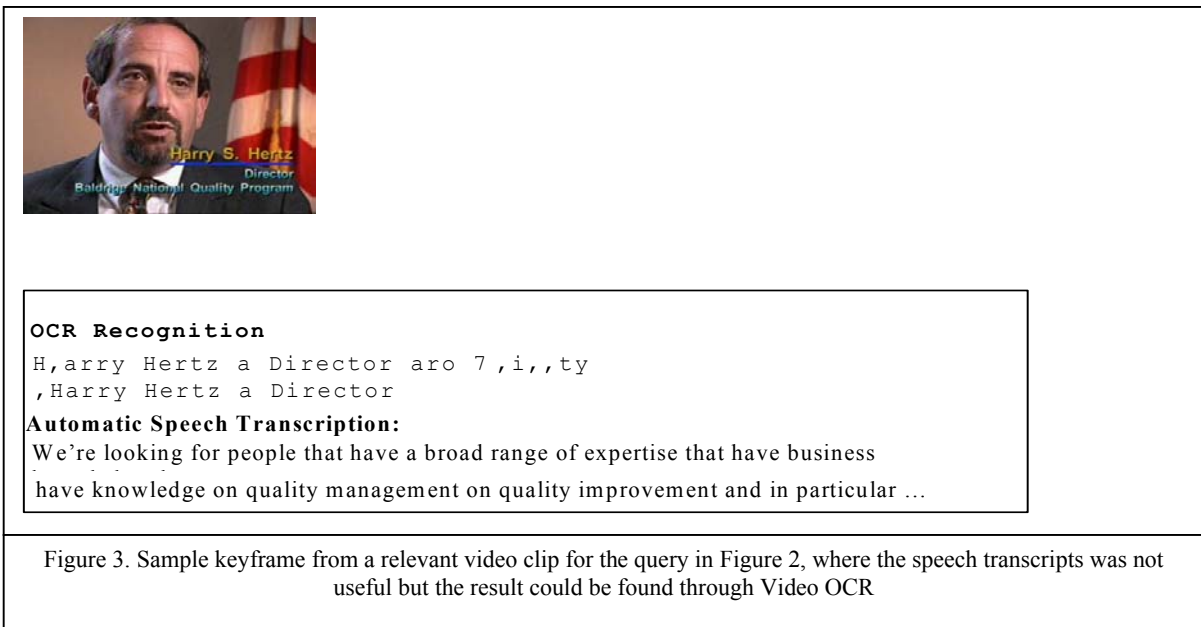|	
Topic 3, 14 & 19: Lunar rover on moon	Topic 51: Splashing water
Topic 4: Mountains as prominent scenery	Topic 52: Space shuttle on launch pad
Topic 5, 9 & 31: Water skiing	Topic 57 & 60: Explosions and blasting
Topic 6: Yellow boat	Topic 59: Space Shuttle "Discovery"
Topic 7: Pink flower	Topic 64: Male interviewees
Topic 8: Planet Jupiter	Topic 67: Research aircraft X-29
Topic 10: Swimming pools	Topic 69: Logo of Northwest Airlines
Topic 11: People on beach	Topic 70: Who is the producer of the video
Topic 12: Surface of planet Mars	Topic 71: Street Traffic
Topic 13: Speaker in front of U.S. flag	
Topic 15: Corn on the cob	
Topic 16: Deer with antlers	
Topic 17: Airliner landing	
Topic 20: Pictures of Ron Vaughn	
Topic 21: Ronald Reagan speaking	
Topic 22: Harry Hertz	
Topic 23: Lou Gossett, Jr.	
Topic 24 & 58: Pictures of R. Lynn Bondurant	
Topic 25: R2D2 and 3CPO from Star Wars	
Topic 26: Victim location system	
Topic 27: A biplane over a field	
Topic 32: Helicopter landing	
Topic 35 & 36: Where else does this person appear	
Topic 2, 37, 56, 72: Rocket and shuttle launches	
Topic 41: Shots with crowds of 8+ people	
Topic 42: Scenes with David J. Nash	
Topic 44: Pan and tilt camera action	
Topic 50: Natural outdoor scenes with birds	

3.2   Methods for Extracting Textual Metadata

3.2.4   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 [9]. 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%.

3.2.5   Video OCR

A different, textual, representation is derived by reading the text that present in the video images using optical character recognition (OCR). 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. Our video optical character recognition system [5] uses the following approach to identify and recognize captioned text that appears on the video. Given the number of frames contained in typical broadcast news, it is not computationally feasible to process each and every video frame for text. For this reason a rough or quick text region detection is performed first. Then the text must be extracted from the image, and converted into a binary black and white representation, since the commercially available OCR engines do not recognize colored text on a variably colored background. Unlike text printed on white paper, the background of the image tends to be complex, with the character hue and brightness very near the background values.

OCR Recognition
`H,arry Hertz a Director aro 7 ,i,,ty`
`,Harry Hertz a Director`
Automatic Speech Transcription:
We're looking for people that have a broad range of expertise that have business
have knowledge on quality management on quality improvement and in particular …

Figure 3. Sample keyframe from a relevant video clip for the query in Figure 2, where the speech transcripts was not useful but the result could be found through Video OCR

Text area detection is often referred to as text segmentation. This is done in our system by detecting vertical edges and smoothing them. The regions where edges were detected are grouped into clusters and bounding boxes are applied. A number of heuristics then identify text boxes based on their aspect ratio, absolute size and the fill factor of the bounding boxes.

Once a text box is detected, enhancement takes place. Individual text areas are combined into region. Multi-frame integration looks at the potential bounding boxes over several frames and finds the minimal (white) pixel values across that range. This uses an assumption that the text is stable in the image, i.e. overlaid, while the background may be moving inside the image. Only text that is on the screen for at least 1 second is readable by humans.

The next step before OCR is character detection through filter integration. Different filters looking for horizontal, vertical, left diagonal, and right diagonal lines are combined and blurred to a gray scale. Adaptive thresholding on the gray-scale histogram is then used to create binarized black on white text. Character segmentation occurs at the troughs on the histogram.

The OCR is further complicated by the fact that the text has very low resolution, frequently only about 10 pixels of height per character. This resolution is due to the NTSC television standard of 325x248 pixels per image. To overcome the resolution problem, the detected text is magnified and sub-pixel interpolation performed to increase resolution without incurring jagged edges as artifacts of the magnification.

The potential text region is then extracted as a tiff image and submitted to a commercial optical character recognition package for the final stage of recognizing the text. Since the extraction and binarization steps are quite noisy and do not produce perfect results, our system runs the OCR engine on every 3rd frame where text was detected. Thus we obtain 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 only slightly different from each other. On this video collection, the word accuracy for detected text was estimated to be 27%.

3.2.5.1 Correcting OCR errors

We explored two different methods for correcting errors in the OCR transcriptions, both applied only to unmatched query words. The first method generates a new set of n-gram strings to match the unedited the OCR transcriptions. These n-gram strings include strings with an edit distance of 1 character (1 deletion, insertion or substitution) and all possible n-gram substrings with at least 3 characters.

Our second method for OCR correction involved the dictionary spelling correction method provided in MS Word. Through an application program interface to the features of MS Word 2000, an OCR recognized string was expanded into its possible "corrected" spellings. We proceeded in a very conservative fashion, only expanding words that MS Word had flagged as incorrectly spelled. This dramatically reduced the number of spurious word candidates and avoided false matches.

3.2.6   Information Retrieval with Text Material

All retrieval of textual material was done using the OKAPI formula [3]. 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.

3.2 IMAGE-BASED INFORMATION EXTRACTION AND RETRIEVAL

3.2.1 Face Detection

Face detection is one of a class of object detection tasks that are useful for image and video analysis. At the core is a simple recognition problem: does this image or region contain a face or not. There are more subtle aspects to faces, such as the facial activities, expressions and emotions, which are currently beyond the scope of automatic analysis on general broadcast video. What we really want to know is whose face is in the picture, and knowing that a face is there is only a partial step towards the true goal. This holds true for many other object recognition tasks, however as humans we are predisposed to finding faces as a general class more interesting than trees, rocket launches clouds, or buildings.

Extensive work in face detection has been done at CMU by Rowley [4]. This approach modeled the statistics of appearance implicitly using an artificial neural network. The neural network was trained on multiple 'face' windows templates, each 20x20 pixels. Images that might contain larger faces were subsamples to reduce their size. Training was

done on a large set of rotated, scaled, translated and mirrored faces. The training also incorporated negative examples from false alarms in training. To increase confidence, overlapping detected faces were merge. rbitration between multiple neural networks that were trained from different initializations.

Currently we use Schneiderman's approach [8], which applies statistical modeling to capture the variation in facial appearance. This approach tries to learn the statistics of both object appearance and "non-object" appearance using a product of histograms. Each histogram represents the joint statistics of a subset of wavelet coefficients and their position on the object. Schneiderman's approach is to use many such histograms representing a wide variety of visual attributes. The detector then applies a set of models that describe the statistical behavior of a group of wavelet coefficients.

The logical next step after face detection is to recognize or match similar faces. Eigenfaces treat a face image as a two-dimensional N by N array of intensity values. From a set of training images, a set of eigenvectors can be derived that constitute the Eigenfaces. Every unknown new face is mapped into this eigenvector subspace and we can calculate the distance between faces through corresponding points within the subspace [20]. While we experimented with face recognition using a commercial system [15] as well as an implementation of Eigenfaces [6], the accuracy of face recognition in this type of video collection was so poor, that it proved useless. Therefore, we only present results using a face detector that reported the presence of faces in each key frame.

3.2.2 Image Retrieval for Video Clips

To obtain the similarity between the query image I_Q and any image I' in the collection, our model computes the probability of generating the image I' given the observation of the query image I_Q. The model assumes that images are generated through some stochastic process. Given the observation of an image I, we can find the underlying probabilistic model M that generated this image. The optimal probabilistic model for an image I should maximize the generation probability $P(I|M)$. By 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 $P(I_1 | M_2)$, i.e. the probability of generating image I_1 from the statistical model M_2.

To accomplish the video retrieval task using still image retrieval methods, we need to compute the similarity between video shots by using the similarity between images. Let V_Q be a query video example and be represented as a set of I-frame images, i.e. $V_Q = \{V_Q{}^1, V_Q{}^2, ..., V_Q{}^n\}$. Let V_S be a video shot from the collection and be represented as another set of I-frame images, i.e. $V_S = \{V_S{}^1, V_S{}^2, ..., V_S{}^m\}$. The similarity of video shot V_S with respect to query video example V_Q is defined as

$$Sim(V_S, V_Q) \equiv \arg\max_{i \in [1..m]} \left\{ \sum_{j=1}^{n} Sim(V_S{}^i, V_Q{}^j) \right\} \tag{10}$$

where $Sim(V_S{}^i, V_Q{}^j)$ is the similarity of image $V_S{}^i$ with respect to image $V_Q{}^j$.

Using the TREC video collection and the automatic known-item queries, we compared our probabilistic image retrieval model against two other vector-based image retrieval algorithms, namely the well-known QBIC image search engine [1] and a Munsell-color histogram based image retrieval algorithm [21]. Both of these two algorithms represent an image as a vector of features and compute the similarity between images based on the Euclidean distance between their representation vectors.

3.3 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.

4. EXPERIMENTAL RESULTS

4.1 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 [12] 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) [15] 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 2. 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 %
VOCR with n-gram post-processing	5.89 %	11.81 %
Enhanced VOCR with dictionary post-processing	**5.93 %%**	**7.52 %**
Speech Transcripts + VOCR with n-gram post-processing	5.11 %	16.07 %
Speech Transcripts + dictionary enhanced Video OCR	7.07 %	20.74 %
Image Retrieval using QBIC 'Histogram' mode	6.65 %	12.31 %
Image Retrieval using QBIC 'Draw' mode	10.12 %	17.62 %
Image Retrieval using Munsell Color Space histograms	8.60 %	13.56 %
Image Retrieval only using a probabilistic Model	**14.99 %**	**24.45 %**
Probabilistic Image Retrieval + Speech Transcripts	14.99 %	24.45 %
Probabilistic Image Retrieval + Face Detection	15.04 %	25.08 %
Probabilistic Image Retrieval + Raw VOCR	17.34 %	26.95 %
Probabilistic Image Retrieval + dictionary enhanced VOCR	**18.90 %**	**28.52 %**
Probabilistic Image Retrieval + Face Detection + dictionary enhanced VOCR	18.90 %	28.52 %
Probabilistic Image Retrieval + Speech Transcripts + dictionary enhanced VOCR	18.90 %	28.52 %
Probabilistic Image Retrieval + Face Detection + Speech Transcripts +Enhanced VOCR	18.90 %	28.52 %

4.2 RESULTS FOR INDIVIDUAL TYPES OF METADATA

The results are shown in Table 2. 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 n-gram post-processing improved the OCR output to 5.89% ARR (11.81% recall). The effects of dictionary 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 probabilistic 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. The main findings from the results on individual features are:

- Probabilistic image retrieval provided the best result for any single metadata type.
- Speech recognition was surprisingly ineffective, especially when compared to OCR.

4.3 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).

Interestingly enough, combining the n-gram post-processed OCR with the speech transcripts (ARR of 5.11% and recall of 16.07%) did not improve the retrieval effectiveness. But the dictionary-based post-processing method, which on its own had about the same precision and 40% *lower* recall than the n-gram method, provided a more effective combination with the speech transcripts at 7.07% ARR and 20.74% recall. This is about a 10% increase over the previous best combination. N-gram OCR correction initially appeared as good as the dictionary method, but much worse in combination with speech transcripts, possibly due to over-generation of word candidates. Combining OCR and image retrieval thus 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.

To understand the low success rate of the speech transcription, we looked at the distribution of the query keywords with respect to relevant shots. Figure 4 shows that in this collection, the query words found in the speech transcript did not correspond to relevant shots, nor were there many relevant shots near the query words. For VOCR, the hit rate was somewhat better, as expected from the higher ARR and precision scores.
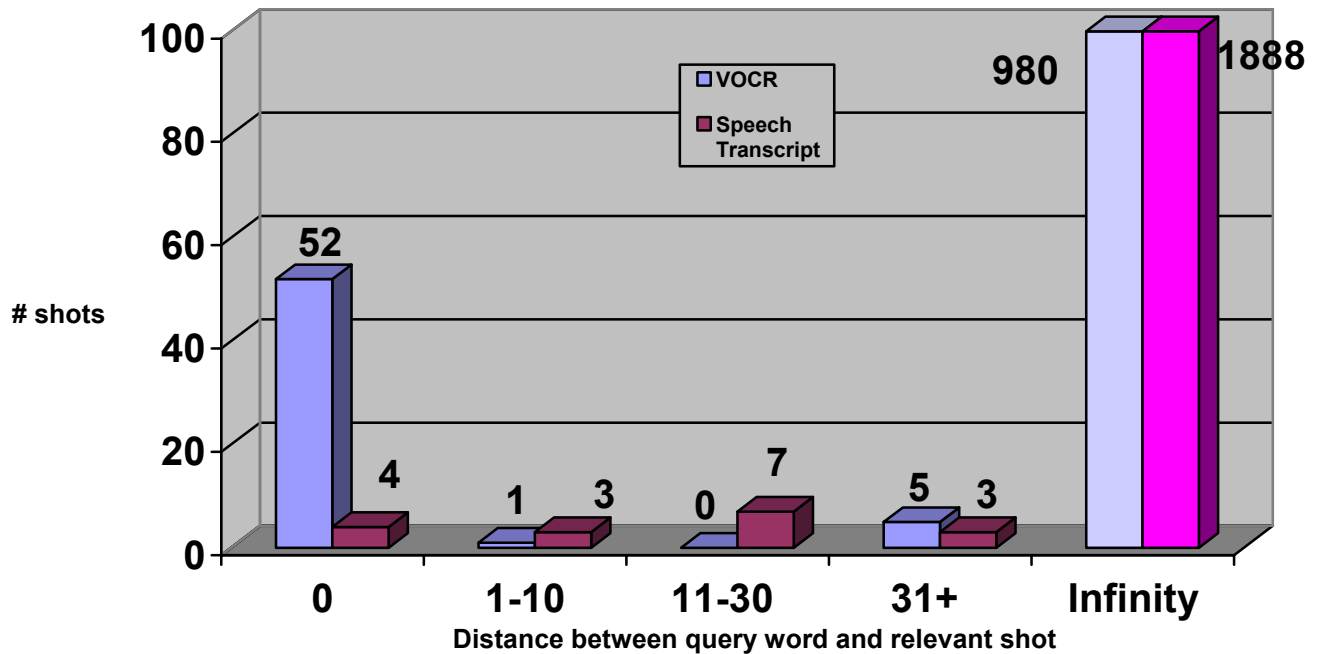
Figure 4. Number of relevant shots at a given distance to a query keyword for words from Video OCR and speech transcription. Infinite denotes the case where no relevant shot was found within the same video file which contained the matched query word.

The main findings for combined metadata retrieval are:
- OCR and Image retrieval provide good complementary information
- Query words found in the speech transcript do not correlate well to relevant shots
- All other combinations have negligible impact on retrieval results.

4. DISCUSSION

What have we 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 [5], so the information contained in the text of the pictures is not redundant to the spoken words in the transcripts. Our most surprising finding is the dominating importance of OCR over speech recognition in this video retrieval task. This surprise was perhaps due to queries that were designed for video documents and not merely text transcripts. Another possible explanation is that OCR text appears directly inside a relevant image, while relevant words can be spoken in the vicinity near the relevant video clip, but not directly during the target shot.

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. In the future we also plan to evaluate speaker identification [7] and sound classification as an additional source of extracted data for retrieval.

5. ACKNOWLEDGEMENTS

REFERENCES

1. 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.
2. 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.
3. Robertson S.E., et al.. Okapi at TREC-4. In The Fourth Text Retrieval Conference (TREC-4). 1993.
4. Rowley, H., Baluja, S. and Kanade, T. Human Face Detection in Visual Scenes. Carnegie Mellon University, *Technical Report CMU-CS-95-158*, Pittsburgh, PA.
5. 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.
6. Satoh, S., and Kanade, T. NAME-IT: Association of Face and Name in Video. *IEEE CVPR97,* Puerto Rico, 1997.
7. 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.
8. Schneiderman, H. and Kanade, T. Probabilistic Modeling of Local Appearance and Spatial Relationships of Object Recognition, *IEEE CVPR*, Santa Barbara, 1998
9. 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.
10. 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.
11. Swain M.J. and Ballard, B.H. "Color Indexing," Int'l J. Computer Vision, vol. 7, no. 1, pp. 11-32, 1991.
12. Tague-Sutcliffe, J.M., "The Pragmatics of Information Retrieval Experimentation, revised," Information Processing and Management, 28, 467-490, 1992.
13. TREC 2001 National Institute of Standards and Technology, Text REtrieval Conference web page, http://www.trec.nist.gov/, 2001.
14. The TREC Video Retrieval Track Home Page, http://www-nlpir.nist.gov/projects/trecvid/
15. Visionics Corporate Web Site, FaceIt Developer Kit Software, http://www.visionics.com, 2002.
16. Voorhees, E. and Harman, D., The Eighth Text Retrieval Conference(TREC-8), Gaithersburg, MD, 2000 http://trec.nist.gov/pubs/trec8/t8_proceedings.html
17. Voorhees E.M, and Tice, D.M., "The TREC-8 Question Answering Track Report," The Eighth Text Retrieval Conference (TREC-8), 2000

18. 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.

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

20. M. Turk and A. Pentland. Eigenfaces for recognition. Journal of Cognitive Neuroscience, 3(1):71-86, 1991.

21. A. Del Bimbo " Visual Information Retrieval", Morgan Kaufmann Ed., San Francisco, USA, 1999.

22. A. Mojsilovic, J. Kovacevic, J. Hu, R.J. Safranek, and S.K. Ganapathy, "Matching and Retrieval Based on the Vocabulary and Grammar of Color Patterns," IEEE Trans. Image Processing, 9(1), pp. 38-54, 2000

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