International Conference on Acoustics, Speech, and Signal Processing (ICASSP'04), Montreal, Canada, May 17-21, 2004

SEARCHING FOR A SPECIFIC PERSON IN BROADCAST NEWS VIDEO

Ming-yu Chen, Alexander Hauptmann

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
{mychen, alex}@cs.cmu.edu

ABSTRACT

People as news subjects play an important role in broadcast news and finding a specific person is a major challenge for multimedia retrieval. Beyond mere content-based general retrieval, this task requires exploitation of the structure, time sequence and meaning of news content. We introduce a comprehensive approach to discovering clues for finding a specific person in broadcast news video. Various information aspects are investigated, including text information, timing information, scene detection, and face recognition. Experimental results on the TREC 2003 video search task show that our approach can achieve surprisingly high performance by exploiting broadly diverse information to find specific named people.

1. INTRODUCTION

Multimedia information has become dramatically more pervasive in recent years. Due to its abundance and variance, an essential research issue is to investigate accurate access to multimedia content. Lately, content-based retrieval [1] has attracted much attention and many research efforts are underway to achieve this goal. They include image retrieval by color similarity [2], by texture similarity [3], video segmentation [4], and video retrieval [5]. These efforts all provide different views on retrieving multimedia content and augment our fundamental understanding of how to analyze multimedia data.

This paper targets very specific content in broadcast news, namely finding a specific person. Since news events are strongly related to people in the news, finding *'Person X'* is a frequent challenge in broadcast news retrieval. [6] has shown that text information and face recognition technology can be useful to search for a person. This paper proposes a comprehensive approach, which utilizes text information, timing information, news anchor scenes and face information to achieve accurate person retrieval. Our evaluation is based on the subset of six topics from the TREC 2003 Video Track (TRECVID) [8], which required finding a specific person in broadcast news video.

2. TEXT AND TIMING INFORMATION

In broadcast news, content information in text form includes closed captions, speech transcription and video optical character recognition (VOCR). We consider all text words that occur within the same camera shot as part of one document representing the text content of this shot. A shot is defined as unbroken sequence of frames taken from one camera and is the basic retrieval unit for the TRECVID evaluations.

2.1. Text search

Text search is the most intuitive way to retrieve shots of a person, given the name as text. We employ the OKAPI text retrieval formula [9], whose basic idea is to rank the relationship between a query and a document using term frequency and document frequency:

$$T(Name,S)=\sum_{qw\in Name}\left\{\frac{tf(qw,S)\log\left(\frac{N-df(qw)+0.5}{df(qw)+0.5}\right)}{0.5+1.5\frac{|S|}{avg\_dl}+tf(qw,S)}\right\} \quad (1)$$

where $tf(qw,S)$ is the term frequency of word qw in the shot document S, $df(qw)$ is the document frequency for the word qw, |S| denotes the length of the document S, N denotes the number of documents in the collection and avg_dl is the average document length in words for all the documents in the collection.

Text retrieval finds shots where the person's name is mentioned. However, not every mention of a person's name corresponds to an image of that person, due to the reporting structure of broadcast news. In typical news segments, the anchorperson briefly describes the news at the beginning, followed by shots of the news event or interview. The person often occurs during the news event shots or at times together with the anchor, e.g. in split screen interviews. Text retrieval gives an approximate location in the video where the person could be found, but not the exact shots containing that person.

2.2. Propagation of Text Information

To overcome this problem, we perform text retrieval with a propagated window. Our assumption is that the person's name and image will appear within the same news story. Using a fixed window size, we propagate the text retrieval result out from the shot which mentions the name to capture the person's image occurrence. The propagation is

$$T_p(S) = \sum_{|S-S_i| \in w} \alpha^{|S-S_i|} T(Name, S_i) \quad (2)$$

where *w* denotes the window size, we assigned 12, and α is the parameter to decide the rate of propagation. We assigned 0.8 to α in our experiments.

The propagation approach only estimates the occurrence in nearby shots because we assume the person's image will be shown near where his/her name is mentioned. While this propagation enhances the chance to find the person's image, it does not provide the exact information of where the person is and ignores the local structure of news reports.

2.3. Weighting nearby shots with prior distributions

Shot weighting information can be derived from the temporal structure of broadcast news reports, which is quite predictable. Normally, an anchorperson introduces a news story and then the scene changes to the actual news event. The person's name is usually first mentioned by the anchor and later the person actually appears in the news story. This timing information provides a way to estimate the time difference between the mention of the person's name and their appearance in video.

Using "Madeleine Albright" in the TRECVID 2003 data development collection as an example to investigate the time relationship between the mentioning of the name and the appearance of the person, figure 1 plots the frequency of the appearance of her image and the mention of her name based on the elapsed time between them. The bold line shows where the name is. The columns denote the frequency of her appearance in video at given time distances. The negative time distance means the person's image appears after his/her name is mentioned.

This plot supports our intuition about the temporal structure. A prior distribution can now be constructed from the frequency to estimate the appearance of the person based on time information. This prior distribution formula is:

$$T_{prior}(S) = \sum_{|S-S_i| \in w} \delta((S-S_i)) T(Name, S_i) \quad (3)$$

where δ (x) function denotes the distribution from developing collection. *w* denotes the window size. To improve our estimates based on limited training data, we smooth the distribution with a small probability.

Figure 2 shows the top 5 results of searching for "Madeleine Albright" in the TRECVID-2003 test collection. The top (left) keyframe in the first row does not contain her image. This confirms our suspicions that even though the shot is strongly related to the text name, it does not contain video of the person. The propagation approach in the second row removes the false shot in the top row, but instead contains the anchor shot. The reason is that the propagation scheme only guesses at the occurrence of the person without any specific clue which shot contains the person's image. In the bottom row, the top 5 results obtained with the prior distribution look perfect, indicating that the prior distribution indeed adds more knowledge of where the person is.

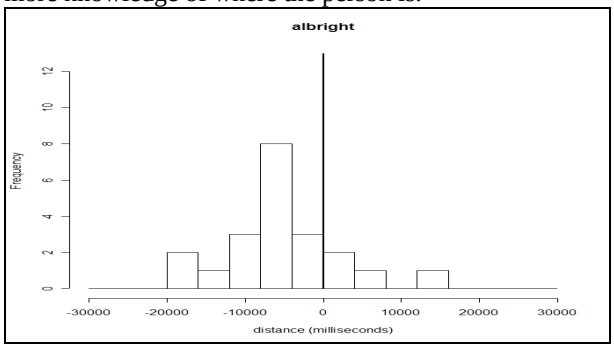

Figure 1: Relationship between the name of the person and time.

Figure 2: The first row depicts the top 5 shots from text search. The second row shows the top 5 shots with a flat propagation scheme. The last row uses a prior distribution.

3. ANCHOR INFORMATION

The prior distribution of the name in text and the corresponding face appearance gives a global estimate of where the person may be, but does not consider the individual image content of each shot. In our experiments, shots with anchorpersons frequently appeared in the top results. Because an anchorperson often introduces a news story and key people, there is a high likelihood that the anchor will mention the person's name. Therefore, we apply an anchor shot detector to remove anchor shots as false alarms based on local scene information.

The approach to building an anchor detector relies on multimodal classification. There are three components to detecting an anchor: the color histogram from image data, speaker id from audio data and face information from face detection [7]. Fisher's Linear Discriminant (FLD) is applied to select distinguishable features from each sub-model. Selected features are synthesized into a new

feature vector, which represents the content of the shot. The classification is performed on those representative vectors.

The prior distribution of text information estimates the shots which may contain the person. The anchor detector discovers the shots with anchorpersons. A linear combination constructs the revised distribution

$$P(S) = \alpha T_{prior}(S) + \beta\, Anchor(S) \quad (4)$$

where *Anchor*(S) denotes anchor detector output for shot S. α and β are the parameters to combine the prior distribution and the anchor detector. These two parameters are trained from TRECVID 2003 development collection.

Figure 3 shows the top 5 retrieval results for Madeleine Albright after combining anchor detection and text information. Compared to the third row of figure 3, only the ranks change. However, in experimental evaluations, the mean average precision (MAP) improves after filtering the anchor shots. These experimental results are reported in figure 6.

Figure 3: Top 5 retrieval results from combination of anchor information and text information

4. FACE RECOGNITION

Another important piece of local information when searching for images of a person is the person's face. Unlike text information and anchor shot removal, face recognition could provide an exact match of the person's image.

We applied the well-known Eigenface algorithm [10] for face recognition. The algorithm tries to encode the most variant parts of faces and ignore the similar parts. Faces are collected from a face detection system, converted to gray levels and normalized to a standard size. Principal component analysis (PCA) is performed to construct an eigenspace, which captures the variance of the faces. The eigenspace is constructed from the first N eigenvectors, which correspond to descending eigenvalues. Faces are projected into this eigenspace and therefore called eigenfaces. Eigenfaces have several advantages: first, they encode the most distinguishing parts of faces. Second, they decrease the dimensionality of the face information, which benefits computation. The eigenface representation of faces has been show to be a fairly robust approach to face recognition.

However, there are drawbacks to the eigenface algorithm. The most important one concerns pose. If a face is slightly tilted, it has a high chance to be similar to other tilted faces. Since only distinguishing parts of the face information are captured by eigenfaces, and pose constitutes a significant difference between faces, pose tends to dominate the recognition results. Lighting conditions present another serious problem for the eigenface algorithm. In broadcast news, due to large variations in news footage, both pose and lighting condition of faces varies widely, resulting in unreliable face recognition.

Our approach tries to simplify the face recognition problem. We consider text information most trustworthy but face recognition information can give additional clues to locally correct the text retrieval result. The main problem with face recognition for searching for a person in broadcast news is the wide variance of conditions. Therefore, we use varied external images (faces) as patterns to retrieve relevant faces. Although, the retrieval task is based on shots, there are many slightly different faces within the frames comprising one shot. We therefore extract all the faces in i-frames to enhance the match. Let the external eigenfaces be denoted as $\{F_1, F_2, F_3 \ldots\ldots, F_n\}$ and the eigenfaces in the broadcast news denoted as $\{f_1, f_2, f_3 \ldots\ldots, f_m\}$. For every external eigenface, there is a rank list, which is the similarity of the external eigenface to every eigenface in the news. The external eigenfaces provide the variability in pose and lighting, giving us a more robust prediction when we combine all the results using the following formula

$$R(f_i) = \frac{1}{n}\sum_{j=1}^{n}\frac{1}{r_j(f_i)} \quad (5)$$

where $r_j(f_i)$ denotes the rank of eigenface f_i in terms of similarity to external eigenface F_j, and $R(f_i)$ denotes the more robust rank obtained by combining all face recognition results.

Since shots are the basic units for the search task; we combine all the per i-frame face recognition results from one shot with the combination schema:

$$F(S) = \frac{1}{k}\sum_{j=1}^{m} R(f_j \subset S) \quad (6)$$

where k denotes the number of faces in shot S, and *F*(S) denotes the face recognition result for shot S.

A linear combination again constructs the final prediction of the appearance of the target person by combining prior text information, anchor detector and face recognition

$$P(S) = \alpha T_{prior}(S) + \beta\, Anchor(S) + \gamma F(S) \quad (7)$$

where α, β and γ are parameters trained on the TRECVID 2003 development collection.

Figure 5 shows the top 5 results of face recognition, external images, and the results after combining text information, anchor information and face information. Figure 6 shows the evaluation of each approach described here to search for Madeleine Albright in the TRECVID 2003 test collection.

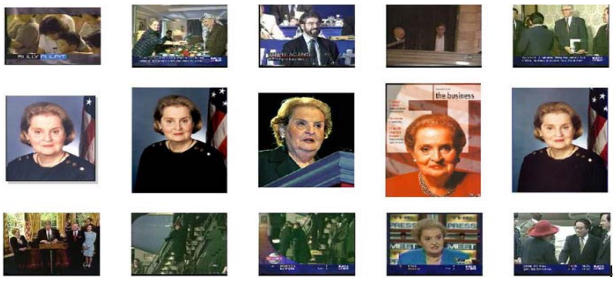

Figure 5: The first row shows the top 5 results based purely on face recognition. The second row shows external images from a Google image search. The third row shows the final combination of text, anchor and face results.

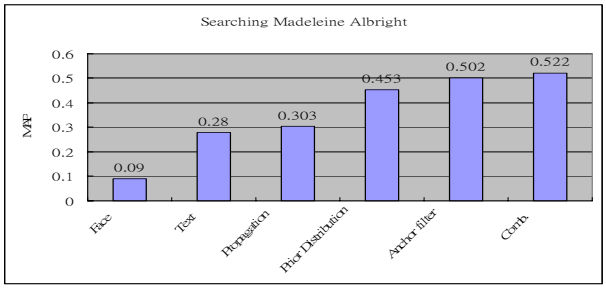

Figure 6: Searching for Madeleine Albright in the TRECVID 2003 test collection. The face result is based on matched Eigenfaces. The propagation and the prior distribution columns are extensions from initial text retrieval by equation (2) and (3); the anchor filter column shows the combination of anchor filtering with the prior distribution. The last column shows the final combined results.

5. EXPERIMENT RESULTS

We confirm the generality of our approach on the TRECVID-2003 Search Task, containing approximately 65 hours of broadcast news video. There were five topics in the task related to finding a specific person, specifically "Yasser Arafat", "Osama Bin Laden", "Morgan Freeman", "Mark Souder" and "Pope John Paul II". The results in Table 1 show that our approach not only performs well in searching for "Madeleine Albright", but also works for searching other people.

Our evaluation metric is mean average precision (MAP). The average shows the improvement of our approach over all 5 topics.

6. CONCLUSION

In this paper, we address the task of finding a specific person using clues including text, time, scene content and face. We present a framework to utilize those clues given minimal training data and external resources to achieve accurate person retrieval. Ultimately, text information proves most trustworthy, but face and scene detectors provide local refinement to improve performance.

ACKNOWLEDGMENTS

This research is partially supported by the Advanced Research and Development Activity (ARDA) under contract numbers MDA908-00-C-0037 and MDA904-02-C-0451.

	Face	Text	Propagation	Prior Dist	Anchor	Comb.
Arafat	0.125	0.200	0.252	0.268	0.278	0.387
Bin Laden	0.007	0.143	0.561	0.511	0.465	0.432
Souder	0.113	0.667	0.641	0.432	0.432	0.461
Freeman	0.587	0.517	0.148	0.445	0.445	0.551
John Paul	0.005	0.368	0.311	0.269	0.315	0.269
Average	0.167	0.379	0.383	0.385	0.387	**0.420**

Table 1: Retrieval results using different methods. Face recognition (face), text retrieval (Text), propagation schema (propagation), prior distribution (prior dist.), text with anchor filter (anchor), and combination of text, anchor and face (comb.) are reported.

7. REFERENCES

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

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

[3] R.W. Conners and C.A. Harlov, "A theoretical comparison of texture algorithms," *IEEE Trains. on Pattern Analysis and Machine Intelligence*, 2(3), pp. 204-222, 1980.

[4] H.J. Zhang, A. Kankanhalli and S. Smoliar, "Automatic partitioning of video," *Multimedia Systems*,1(1), pp.10-28, 1993.

[5] A. Hauptmann, T.D. Ng, R. Baron, W. Lin, M. Chen, M. Derthick, M. Christel, R. Jin, and R. Yan, "Video Classification and Retrieval with the Informedia Digital Video Library System," *Proc. of the TREC-11*, Gaithersburg, MD, 2002.

[6] S. Satoh and T. Kanade, "NAME-IT: Association of Face and Name in Video," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Juan, Puerto Rico, pp. 775-781, 1997.

[7] H. Schneiderman and T. Kanade, "Object Detection Using the Statistics of Parts," *International Journal of Computer Vision*, (to appear).

[8] TRECVID, The NIST TREC Video Retrieval Evaluation, homepage: http://www-nlpir.nist.gov/projects/t01v/

[9] S. E. Robertson, S. Walker, M. M. Hancock-Beaulieu, M. Gatford, and A. Payne. Okapi at TREC-4. In Fourth Text REtrieval Conference (TREC-4), pages 73-96, Gaithersburg, MD, 1996. National Institute of Standards and Technology,

[10] Pentland A., Moghaddam B., and Starner T. View-Based and Modular Eigenspaces for Face Recognition IEEE Conference on Computer Vision & Pattern Recognition, Seattle, WA, July 1994.