International Conference on Image and Video Retrieval (CIVR'04), Dublin City University, Ireland, July 21-23, 2004

Finding Person X: Correlating Names with Visual Appearances

Jun Yang, Ming-yu Chen, and Alex Hauptmann

School of Computer Science, Carnegie Mellon University
Pittsburgh, PA 15213, USA
{juny, mychen, alex}@cs.cmu.edu
http://www.informedia.cs.cmu.edu

Abstract. People as news subjects carry rich semantics in broadcast news video and therefore finding a named person in the video is a major challenge for video retrieval. This task can be achieved by exploiting the multi-modal information in videos, including transcript, video structure, and visual features. We propose a comprehensive approach for finding specific persons in broadcast news videos by exploring various clues such as names occurred in the transcript, face information, anchor scenes, and most importantly, the timing pattern between names and people. Experiments on the TRECVID 2003 dataset show that our approach achieves high performance.

1 Introduction

The dramatic increase of digital videos demands more efficient and accurate access to video content. Content-based analysis and retrieval has been extensively used for video segmentation [2], video retrieval [3], and image retrieval [1]. As discussed in [4], finding a specific person in videos is essential to understand and retrieve videos. Although solving this problem might be difficult for general videos, in this paper we target at very specific content namely broadcast news video. Since news videos are strongly related to human subjects, finding "*person X*" is an important and frequent challenge. Taking advantage of the multimodal content in videos, we propose a people-finding approach which exploits name occurrence in transcript, video structure, and visual information such as faces and news anchor scenes. Specifically, this approach utilizes a timing model to overcome the temporal offset between names and persons, which will otherwise compromise performance. Our approach was developed and evaluated using the dataset from TREC 2003 Video Track (VIDTREC) [5], which is divided into a training set (FSD) and a testing set (FST), each consisting of over 100 hours of ABC, CNN, and C-SPAN news video.

2 Transcript search with timing-based score propagation

An essential clue for finding a person in the broadcast news video is the mention of his/her name in the transcript, acquired either from a speech recognizer or from closed

captions. This clue indicates that this person is likely to appear visually. We do not address the rare cases where a person appears without his/her name being mentioned. In this section, we discuss using transcript to find and rank video shots that contain specific persons. Here a video shot is defined as an unbroken sequence of frames taken by one camera and it serves as a basic structural unit in our video retrieval.

2.1 Basic transcript-based search

Since the transcript is temporally aligned with the video, each shot is associated with a portion of the transcript that falls within its boundary. Therefore, an intuitive way to finding a specific person in video is to use text-based retrieval techniques to find the shots which contain the name. Specifically, we employ the TFIDF retrieval method [6], which gives the similarity between a shot S and a person named X as:

$$R(X,S) = \sum_{t_i \in X} tf_i \cdot \log \frac{N}{n_i} \Bigg/ \sqrt{\sum_{t_i \in X} \left(tf_i \cdot \log \frac{N}{n_i} \right)^2} \qquad \textbf{(1)}$$

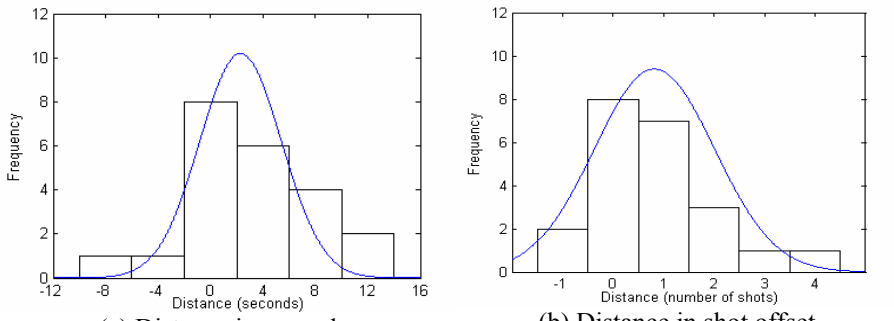
where tf_i is the frequency of term t_i (as a part of the name X) in the transcript of shot S, N is the total number of shots, and n_i is the number of shots whose transcript has t_i.

2.2 Modeling timing between names and persons

The method above is subject to a severe problem: it is not necessarily the case that a person appears in the video concurrently with the name mentioned in the transcript. Based on the statistics we have collected, in more than half the cases, a person does not show up in the shot where the name is mentioned, but before or after that shot. Undoubtedly, this mismatch seriously compromises the performance of text-based shot retrieval, which explores only the shots containing the person's name.

The timing between visual appearances (i.e., face) and occurrences of a name is related to the "video grammar" of broadcast news. In a typical news story, an anchorperson briefs the news at the beginning, followed by several shots showing the news event and sometimes interviews and reporters. The name of a human subject in the news is normally first mentioned by the anchorperson, while his/her face is not always shown at that time. In the following shots, this person may appear several times in the video, roughly interleaved with occurrences of the name in the transcript. However, there are also cases where a person not mentioned by the anchorperson later appears in the shots, with or without his name mentioned in close proximity.

Generally, no simple pattern is able to capture the possibility of such timing, but it is still true that a person is more likely to appear in the (temporal) proximity where his name is mentioned. Loosely speaking, the closer is the shot to name occurrence, the more likely it contains the person's visual appearance. As an example, we collected all the visual appearances of "Bill Gates" in FSD, and plot in Fig.1 the frequency of these appearances at each quantized distance from their closest occurrence of his name. The distance is measured in terms of time or shot offset (number of shots between). The "0" point on the distance axis is where the name is mentioned, and positive distance means that a person appears visually after the name is mentioned.



(a) Distance in seconds (b) Distance in shot offset

Fig. 1. The frequency of Bill Gates' visual appearances associates with name occurrences, and the Gaussian curves capturing the frequency distribution.

Based on Fig.1, it is intuitive to model the frequency of a person's visual appearance w.r.t his name occurrence using a Gaussian model. For a specific person, we estimate a Gaussian distribution from the distances from each of his visual appearances in FSD to the *closest* name occurrence, both of which are manually labeled, using maximum likelihood estimation. Again, the distance is measured in terms of time or shot offset. In Fig.1, we superimpose the curves of the estimated Gaussian distributions for "Bill Gates", which nicely capture the shape of the bins showing the frequencies.

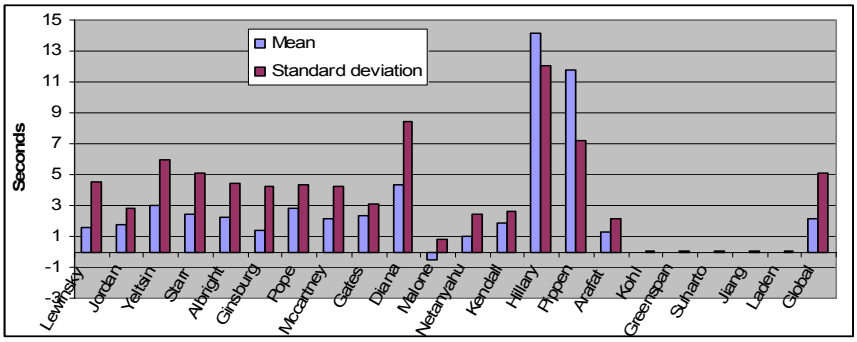
Totally 20 persons are selected for study, varying from frequently appearing ones like "Michael Jordan" to rare ones like "Alan Greenspan". Table 1 shows the number of visual appearances of each person in FSD and FST respectively. The mean and standard deviation of the Gaussian distribution of each person estimated on FSD is ploted in Fig.2 (a) for time-based distance and in Fig.3 (b) for shot-based distance. People are ordered from left to right in descending frequency of their visual appearance in FSD. A global distribution computed from a pool of the training data from all the people is shown alongside.

Table 1. The 20 people studied and the number of their visual apperances in FST and FSD.

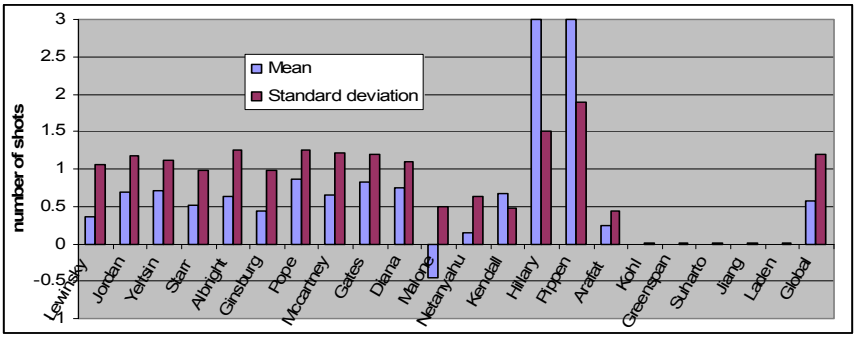
Name	Lewinsky	Jordan	Yeltsin	Starr	Albright	Ginsburg	Pope	Mccartney	Gates	Diana
FSD/FST	53 / 44	47 / 75	40 / 10	37 / 35	30 / 40	28 / 22	29 / 45	26 / 10	22 / 19	12 / 7
Name	Malone	Netanyahu	Kendall	Hillary	Arafat	Kohl	Greenspan	Suharto	Jiang	Laden
FSD/FST	11 / 19	7 / 42	6 / 3	6 / 12	3 / 33	3 / 6	2 / 6	2 / 20	2 / 19	0 / 26

As shown in Fig.2 (a), for the first 9 people on the left, each of who appears 20+ times in FSD, the estimated distributions have similar mean values (1-3 sec.) and moderate standard deviations (3-6 sec.). This suggests that the Gaussian assumption is reasonably good for these people, and their distributions are similar to each other. Therefore, on average a person appears about 2 seconds after his name is mentioned in the "grammar" of news video. For the people with less than 20 appearances in FSD, however, the estimated distributions differ significantly: the mean varies from -2 to 14 seconds, and the standard deviation can be as large as 12 seconds. But it is not fair to say that each infrequent name has a unique distribution, since our observation is

biased by the insufficient training data in FSD used to estimate their distributions. We will explore this question further in our experiments. The same trend is observed in the shot-based distributions in Fig.2 (b).



(a) time-based distance



(b) shot-based distance

Fig. 2. The mean and standard deviation of the Gaussian distributions for each person

2.3 Search methods with score propagation

Given the timing information, it is obvious that the basic transcript-based search can be improved by propagating the similarity scores from the shots containing the intended person's name to the neighboring shots in a window. The propagation is carried out as:

$$R_p(X,S) = \sum_{|S-S_i|<w} f(S,S_i)R(X,S_i) \quad (2)$$

where w is the size of the window measured either by time or by shot offset, and $f(S, S_i)$ is a weighting function with output within (0, 1), which decides the score being propagated to neighboring shots. The summation traverses all the shots S_i that are in the neighborhood of S and have the intended name in the transcript.

The weighting function $f(S, S_i)$ can take many forms, depending on the design decisions made along the following dimensions:

- *Flat window* or *weighted (Gaussian) window*: In a flat window, $f(S, S_i)$ is a constant and all the shots in the window are propagated with the same score. In a weighted window, however, the score propagated to each shot is determined by its probability of containing the person's visual appearance, which is calculated from the density function of a Gaussian distribution. In this case, $f(S, S_i)$ is

$$f(S, S_i) = \int_{start}^{end} N(u, \sigma^2) \quad (3)$$

  where *start* and *end* are the starting and ending position of S in relation to S_i (which has the intended name), and $N(u, \sigma^2)$ is the density function of the Gaussian distribution.

- *Time-based* or *shot-based distance measure*: This decides whether to use a time-based Gaussian model $N_X^t(u, \sigma^2)$ or a shot-based one $N_X^s(u, \sigma^2)$. This makes a difference since the shot length differs a lot, and it is unclear which measure is more desirable as to revealing the relationship between a person's visual appearance and the name occurrence.
- *Local, global* or *combined Gaussian distribution*: To search for a person, we can use the local Gaussian distribution trained particularly for this person $N_X(u, \sigma^2)$, the global distribution trained on all the people $N_G(u, \sigma^2)$, or a combination of them $N_C(u, \sigma^2)$. Intuitively, if each person has a unique distribution and there is enough training data, the local (people-specific) model is more desirable; otherwise the global one is better. The combined model uses a distribution integrated from both the local distribution and the global one. Inspired by the smoothing techniques used to overcome the sparse training data problem in information retrieval [8], this model "smoothes" a person's local distribution estimated from insufficient data with the global distribution. Specifically, the probability density function of the combined distribution is a linear combination of that of the local and the global distribution, where the weight is determined by the amount of training data associated with the person. It is formulated as:

$$N_C = \alpha N_X + (1 - \alpha) N_G \quad \text{and} \quad \alpha = sigmoid(\frac{T_X}{\beta} - \gamma) \quad (4)$$

  where α is the weight computed from the number of training data T_X for person X, and β and γ are constants, which are set to 10 and 1 as determined by our informal experiments. According to the property of sigmoid function, α approaches 1 when T_X increases, and vice versa (e.g., α = 0.5 when T_X=10, and α = 0.88 when T_X=30). Therefore, the more training data we have observed, the more the combined distribution is determined by the local distribution.

3 Face searching and Anchor filtering

Visual information provides valuable clues for finding a person in news video. Unlike text information which roughly estimates where a person is, visual information can tell the exact position and time of the person's appearance. Face recognition

technology can match a person's face visually and predict its identity, though its performance is significantly affected by pose and illumination variances. Another important visual clue comes from the anchor detection, since people as news subjects seldom occur during the anchor shot.

We apply the well-known Eigenface algorithm [9] for face recognition. Faces are collected using a face detection system [10], converted to gray levels and normalized to a standard size. Principal component analysis (PCA) is performed to construct Eigenfaces, which encode the most distinguishing parts of faces while ignore similar parts. The Eigenface representation has been shown to be a fairly robust approach to face recognition. However, it also has several drawbacks and the most serious one is pose variations, as non-frontal faces usually have much poorer recognition results than frontal ones. Lighting conditions present another serious problem. In broadcast news, due to the large variations in news footage, both the pose and lighting condition of faces vary largely, resulting in unreliable face recognition.

To avoid the face recognition difficulties, we first use the trustworthy text information to find some shots as initial results, and apply face recognition on them to obtain additional clues for refining the initial results. In this way, the number of faces to be recognized is largely reduced and the accuracy can be improved. To address the wide variance on pose and lighting conditions, we find external images that contain the target face with varied conditions and use them as examples to recognize relevant faces. The (internal) faces to be recognized are extracted from the i-frame of the shots to be examined. Let the external Eigenfaces be denoted as {*F1, F2, F3 ..., Fn*} and the internal Eigenfaces be denoted as {*f1, f2, f3 ..., fm*}. By matching every internal face with a specific external face F_j based on Eigenface, we obtain a ranking of all internal faces ordered by descending similarity to F_j. The final rank of an internal face is combined from its ranks with all the external faces, given as:

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

where $R_j(f_i)$ denotes the similarity rank of internal face f_i with external face F_i, and $R(f_i)$ denotes the final rank of f_i. Since the external faces provide variances in pose and lighting condition, the final rank gives us a more robust prediction. Since a shot may has more than one i-frames, we average the rank of the face on every i-frame of the shot to get the score indicating how likely the shot contains the target face. More details of our face recognition method can be found in [11].

The inclusion of anchor detection assumes that anchors seldom co-occur with a news subject person. We have built an anchor detector [3] based on multimodal classification that combines three information sources: the color histogram from image data, speaker ID from audio data, and face information from face detection. Face information contains the position, size and detection confidence of faces. Fisher's Linear Discriminant (FLD) is applied to select distinguishing features for each source of information. Selected features are synthesized into a new feature vector of each shot, and the classification is performed on these feature vectors.

The final prediction of the appearance of the target person is made by linearly combing the results of text-based search, anchor detection and face recognition:

$$P(S) = \alpha\, T_{prior}(S) + \beta\, Anchor(S) + \gamma F(S) \quad (6)$$

where α, β and γ are weights for the three predictions, which are trained on a held-out set from FST (as FSD has been used to train to distribution).

4 Experiment results

Experiments in finding the 20 selected persons in the TRECVID 2003 collection are conducted to determine the best people-finding method among those proposed in Sect. 2.3. Firstly, we compare the performance of the basic transcript-based search method without score propagation (denoted as *Baseline*), the method with flat-window propagation (*Flat_Win*), the one with shot-based Gaussian propagation using the local distribution estimated from FSD (*Shot_Gauss_Local*), and its time-based counterpart (*Time_Gauss_Local*). For each person, we use each method to find the shots in FST that contain his/her visual appearance and compute the *mean average precision* (MAP) [7] of the results. Note that the propagation window sizes in each method have been fine-tuned based on the FSD data.

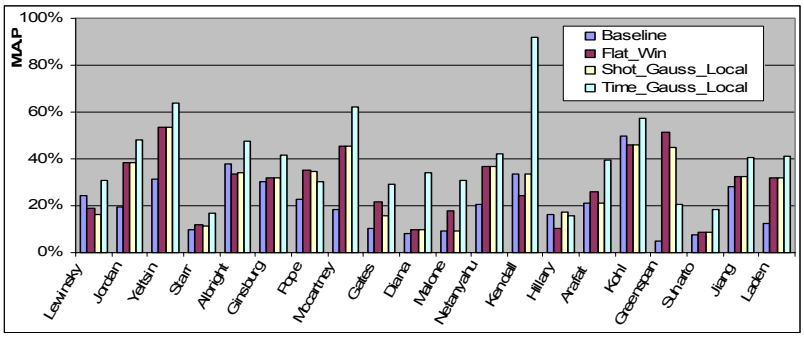


Fig. 3. Performance comparison of three propagation methods with baseline method

As shown in Fig.3, in all the 20 queries at least one propagation approach outperforms the baseline, and for 15 queries among them, all the three propagation approaches outperform the baseline. This suggests that score propagation based on timing information can greatly help the task of people-finding. Moreover, in 17 out of the 20 queries, the time-based Gaussian approach is the best performer, whose average MAP (0.40) is much higher than that of the flat-window approach (0.29) and shot-based Gaussian approach (0.28). Thus, time-based Gaussian is a better propagation strategy than the other two, implying that time is a better distance measure than shot offset w.r.t. revealing the timing between names and people.

Fig.4 shows the average MAP (over 20 queries) of the time-based Gaussian method using local, global, and combined distribution respectively, in comparison to that of baseline and flat-window approach. As shown, the approach with combined distribution outperforms the global one by 2%, which beats the local one by another 2%, and all are about twice the performance of the baseline approach.

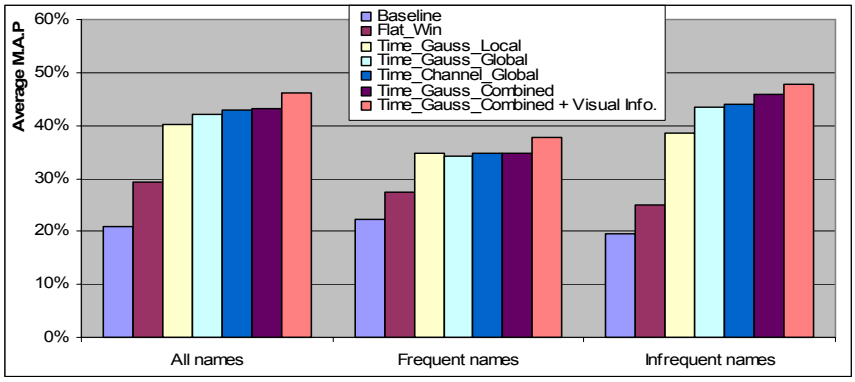


Fig. 4. Performance comparison of local, global, combined distribution with visual information

The three types of distribution cause more interesting discrepancy on the performance of finding frequently occurring people versus that of finding infrequent ones. Here frequent people are those who appear visually 20+ times in both FSD and FST (cf. Table 1), while infrequent ones are those appearing 20- times in both FSD and FST. By this standard, there are 7 frequent and 8 infrequent people among the 20 people, while 5 people cannot be clearly classified due to their unbalanced appearances in FST and FSD. As we can see, for frequent names the choice of distributions does not have any significant influence on the performance, while for infrequent names the difference is substantial. Specifically, for all the 7 frequent people, the MAP of global distribution never differs from that of local distribution by over 10%, while for 5 out of the 8 infrequent ones, global distribution enhances the MAP by over 20%. This echoes our observation in Sect.2.1 that the distribution of frequent names is similar to each other and thus to the global one, which is dominated by the dense training data of frequent people. Therefore, the performance of finding such people is almost unaffected by the choice of distribution. For infrequent people, since their local distribution is poorly estimated using their insufficient training data, the performance can benefit from using the more stable global distribution. It is interesting to see that the combined distribution is better than the global one, which implies that each name has a unique "true" distribution that lies between the global and the local one. However, this conclusion can be challenged due to insufficient queries (8 infrequent names) and the small improvement (about 4%).

Since our data consist both ABC and CNN news, it is interesting to know if these two channels have different styles that lead to different distributions. Thus, we train two channel-specific global distributions on FSD and test them on FST. As shown in Fig.4, this approach (*Time_Channel_Global)* improves MAP over the uniform global distribution by only 1%, suggesting that ABC and CNN have similar editing styles.

Finally, we combine transcript search with time-based smoothed distribution and vision information. The combination weights we trained from the held-out set are 1.0 for transcript information, -0.812 for anchor filtering and 0.087 for face recognition. These weights reflect the fact that face recognition is very unreliable, while the anchor detection has the ability to remove false positives. As shown in Fig.4, combining transcript with visual information gave another 3% improvement, which is

mainly derived from anchor detection. Among the 20 people, the visual information enhances the MAP on 4 people substantially (over 20%), and we find that they all appear with frontal faces in the video. 10 people have minor improvement (1%-20%) on their MAP with visual information, while the rest 6 people do not improve at all.

5 Conclusion

In this paper, we address the task of finding a person using clues including transcript, video structure, and vision information. Gaussian distribution has been proved experimentally an effective model to describe the timing pattern between a person's visual appearances and the occurrences of his/her names. Specifically, a "smoothed" Gaussian distribution estimated using both the local and global training data produces the best performance, especially for infrequently appearing people. Finally, combing visual information such as face recognition and anchor detection with transcript information brings additional benefit to the person-finding task.

6 Acknowledgement

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

Reference

1. Smeulders, et al.: Content-Based Image Retrieval at the End of the Early Years. IEEE Trans. Pattern Analysis and Machine Intelligence, Vol 22, No 12 (2000) 1349-1379.
2. Zhang, H.J, Kankanhalli, A., Smoliar, S.W. “Automatic partitioning of full-motion video”, ACM Multimedia Systems, 1(1), 1993.
3. Hauptmann, A., et al. Informedia at TRECVID 2003: Analyzing and Searching Broadcast News Video, Proceedings of TREC 2003, (2003).
4. Satoh, S. and Kanade, K.: NAME-IT: Association of Face and Name in Video. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition (1997) 775-781.
5. The NIST TREC Video Retrieval Evaluation, http://www-nlpir.nist.gov/projects/trecvid/.
6. Salton, G., McGill, M.: Introduction to Modern Information Retrieval. McGraw-Hill, New-York (1983).
7. Baeza-Yates, R. and Ribeiro-Neto, N.: Modern Information Retrieval. Addison Wesley, Essex, England (1999).
8. Zhai, C. and Lafferty, J.: A Study of Smoothing Methods for Language Models Applied to Ad Hoc Information Retrieval. Proc. 24th Int'l ACM SIGIR Conf. (2001): pp. 334-342.
9. Pentland, A., Moghaddam, B., and Starne,r T.: View-Based and Modular Eigenspaces for Face Recognition IEEE Conference on Computer Vision & Pattern Recognition (1994).
10. Schneiderman, H. and Kanade T., “Object Detection Using the Statistics of Parts,” International Journal of Computer Vision 2003.
11. Chen, M.Y., Hauptmann, A., "Searching for a Specific Person in Broadcast News Video," Int'l Conf. on Acoustics, Speech, and Signal Processing, May, 2004 (to appear).