Video Grammar for Locating Named People

Jun Yang
School of Computer Science
Carnegie Mellon University, Pittsburgh, PA 15213
(412) 268-9747
juny@cs.cmu.edu

Alexander Hauptmann
School of Computer Science
Carnegie Mellon University, Pittsburgh, PA 15213
(412) 268-1448
alex@cs.cmu.edu

ABSTRACT
Finding a named person in broadcast news video is important in video retrieval. Relying on the text information such as video transcript and OCR text, this task suffers from the temporal mismatch between a person's visual appearance and the occurrence of his/her name in the text. By exploring video grammar regarding the concurrence pattern between faces and names, we propose an extended text-based IR method to overcome this problem, which yield superior performance.

1. INTRODUCTION
Large volumes of digital videos make effective and efficient access to video content an imperative task for the digital library community. One important research issue is to find the visual appearances (or faces) of a named person in the video [1], especially in broadcast news videos which are mainly about people. This can be facilitated by video transcript and video OCR text (VOCR for short), but their usefulness is severely limited since people do not always temporally co-appear with their names. We attack this problem by exploring an aspect of video grammars, namely the concurrence pattern between people's faces and their names, to reformulate the text-based IR model, which yields better performance in people finding. Variations of our approach are discussed and their performance compared with experiments conducted on the TREC03 Video Track data [2]. These data are divided into a training set (FSD) and a testing set (FST), each consisting of over 100 hours of ABC and CNN news video.

2. CONCURRENCE PATTERN
A clue for finding a specific person in news video is the occurrence of his/her name in the accompanying text (transcript or VOCR). This indicates that a person is likely to show up in close proximity. The search is conducted at the shot level, i.e., finding all video shots that contain the visual appearances of the intended person, where a shot is defined as an unbroken sequence of video frames taken by one camera. Text from transcript and VOCR are temporally aligned with the video, and thus each shot is associated with the text that falls within its boundary. To find a specific person, we can treat each shot as a text document and use his/her name as a query to find the shots having that name based on text-based retrieval, typically, vector model with TFIDF weighting [3].

However, this straightforward method has a severe problem: a person does not necessarily appear when the name is mentioned in the text. Based on the statistics we 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. This mismatch seriously compromises the performance of text-based shot retrieval.

The timing between a person's faces and his/her names is related to the *video grammar* of broadcast news, namely the style and structure typical of a certain genre of video. For example, a news story starts with an anchorperson briefing the story, followed by several shots showing the news event. The name of a news subject is normally mentioned by the anchor, but his/her face is often not shown at that time. In the following shots, this person may appear several times in the video, interleaved with occurrences of the name in the text. But sometimes a person not mentioned by the anchor later appears. This grammar can be different depending on text type (transcript vs. VOCR), channel (ABC vs. CNN), target person, etc. Generally, although a person may not appear with the name in the same shot, he/she usually appears in close proximity.

The video grammar results in what we called *concurrence pattern*, which models the probability that a person appears in a shot at a certain distance from the occurrence of his/her name. To illustrate this pattern, we labeled all the faces of "Bill Gates" in FSD, and plot in Fig.1 the frequency of his face appearing at each quantized distance (seconds) from the closest name occurrence. "0" on the distance axis is where the name is mentioned, and positive distance means he appears after his name.

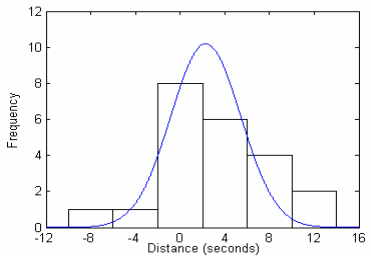

Figure 1: The frequency of Bill Gates' faces w.r.t to his name

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 can estimate a Gaussian distribution from the distances from each of his face in FSD to the *closest* name in transcript or VOCR using maximum likelihood estimation. We superimpose the curves of the estimated Gaussian distribution for "Bill Gates" in Fig.1, which nicely captures the shape of the bins showing the frequencies.

3. THE PEOPLE-FINDING METHOD
Given the concurrence pattern, the mismatch of the text-based retrieval can be overcome by propagating similarity scores from the shots containing the intended name to the neighboring shots. Thus, the probability that a person named X appears in shot S is:

$$P(X,S) = \sum_{|S-S_i|<w} f(S,S_i)R(X,S_i) \quad \text{and} \quad f(S,S_i) = \int_{start(S)}^{end(S)} N(u,\sigma^2)$$

where S_i denotes a shot that has the intended name X and is within a window of size $2w$ around S, and $R(X, S_i)$ is the similarity score

between S_i and name X computed from the text-based retrieval method. $f(S, S_i)$ is a weighting function calculated from a Gaussian distribution $N(u,\sigma^2)$ that models the concurrence pattern.

The concurrence pattern depends on many factors, such as:

- *Transcript vs. VOCR*: It is reasonable to assume that the concurrence pattern for transcript is different from that for VOCR. When a name appears in VOCR (i.e., visible in the current frame), most likely the person is also being shown, which is not necessarily true for transcript. Thus, instead of using one distribution, we should train separate distributions for transcript and for VOCR to handle different types of text.
- *CNN vs. ABC*: Editors of the two TV channels may have distinct styles in editing the news video, resulting in different concurrence patterns. If the discrepancy is large, using separate distributions for the two channels will be beneficial.
- *Local or global distribution*: We can train local distributions on a per-person basis and use each of them for a specific person, or we can train a global distribution using all the training data. The choice depends on whether each person has a unique distribution, and whether there is sufficient training data for everyone. We prefer local distribution if each person has a unique distribution and there is enough training data.

4. EXPERIMENT AND DISCUSSION

20 persons are selected for study, varying from frequent ones like "Michael Jordan" to rare ones like "Alan Greenspan". Fig.2 shows the mean and standard deviation of the Gaussian distributions of each person as well as of some global distributions. Each distribution is labeled with a name and the number of training data used. The first 4 items on the left are the global distribution for video transcript (*Trans*), transcript in ABC (*Trans_ABC*) and in CNN (*Trans_CNN*), and VOCR (*VOCR*). On average, a person appears 1.8 seconds after his/her name is mentioned. But the length of the delay is different for ABC (2.2 sec) and CNN (1.5 sec), showing a difference in the editing style. VOCR has a much smaller mean than that of transcript, which implies that names occurring in VOCR are stronger indicators of the appearance of people. The distributions for each person are trained based on transcripts and ordered in descending number of training data used. The distributions for the first 9 people (each with 20+ training data) are relatively similar to each other and have small variances. In comparison, the distributions for the other people (with fewer training data) differ significantly. But it is premature to say that each "infrequent" person has a unique distribution, since our observation is biased by the limited training data.

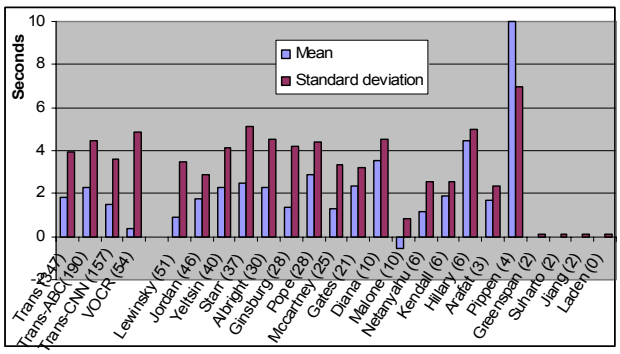

Figure 2: Gaussian distributions estimated from FSD

In the experiment, each variation of the proposed method is used to find the 20 people in FST and evaluated using average MAP (mean average precision) [3]. We divide the experiments into 4 groups, each examining a specific design option while the others are fixed, as shown by different colors in Fig.3.

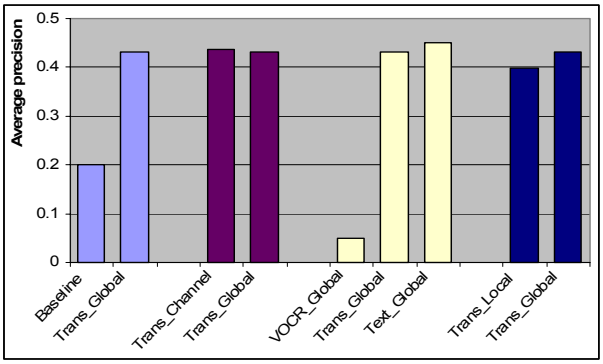

Figure 3: Performance comparison

- In Group 1, both using transcript, our method based on global distribution (*Trans_Global*) achieves over twice the performance of the baseline method (*Baseline*), which does not consider the concurrence pattern. This suggests that the concurrence pattern is very helpful in people-finding.
- Group 2 examines whether using two specific distribution for ABC and CNN (*Trans_Channel*) is superior to using a uniform one. As we can see, although the improvement is small (about 1%), using separate distributions does help.
- Group 3 shows that VOCR (*VOCR_Global*) is not as useful as transcripts, because many names seldom occur in VOCR. Based on the statistics, the name occurrences in VOCR are only 10% of that in transcripts, echoing the difference in performance. Combing their results using logistic regression (*Text_Global*), we improve the transcript-based search by 2%, implying that these two types of text are complementary.
- Group 4 investigates the use of local distributions (*Trans_Local*) estimated on a per-person basis, which is not as good as the global distribution. By looking into the MAP for each query (person), we find that the performance suffers on infrequently appearing people, while for frequent people the choice of distribution almost makes no difference.

5. CONCLUSION

This paper addresses finding named persons in broadcast news video based on transcript and video OCR. As an aspect of video grammars, the concurrence pattern between people's faces and names is studied and modeled using Gaussian distribution. An extended text-based IR method is proposed for people-finding, and experiments have validated the performance of this method.

6. REFERENCES

[1] Satoh, S. and Kanade, T.: NAME-IT: Association of Face and Name in Video. IEEE Conf. on CVPR, 1997, 775-781.

[2] The NIST TREC Video Retrieval Evaluation, http://www-nlpir.nist.gov/projects/trecvid/.

[3] Baeza-Yates, R. and Ribeiro-Neto, N.: Modern Information Retrieval. Addison Wesley, Essex, England, 1999.