# *TO AUTHORS:*

***1. MS Word and LaTex are acceptable formats.***

# *2. Include in text only B/W FIGURES (high quality and clear)*

***3. Submit separate files for all figures and tables. You can also***

***submit color figures (for the Web version of the Handbook)***

***3. FOLLOW THE FORMAT OF THIS SAMPLE  
4. FOLLOW THE FORMAT OF REFERENCES***

***5. Include index terms at the end of the chapter***

# Chapter 1

# ANALYZING PERSON INFORMATION IN NEWS VIDEO

###### Shin’ichi Satoh and Masao Takamashi

Department of Informatics

*National Institute of Informatics, Tokyo, Japan*

## 1. Introduction

Person information analysis for news videos, including face detection and recognition, face-name association, etc., has attracted many researchers in the video indexing field. One reason for this is the importance of person information. In our social interactions, we use face as symbolic information to identify each other. This strengthens the importance of face among many types of visual information, and thus face image processing has been intensively studied for decades by image processing and computer vision researchers. As an outcome, robust face detection and recognition techniques have been proposed. Therefore, face information in news videos is rather more easily accessible compared to the other types of visual information.

In addition, especially in news, person information is the most important; for instance, “*who* said this?”, “*who* went there?”, “*who* did this?”, etc., could be the major information which news provides. Among all such types of person information, “*who* is this?” information, i.e., face-name association, is the most basic as well as the most important information. Despite its basic nature, face-name association is not an easy task for computers; in some cases, it requires in-depth semantic analysis of videos, which is never achieved yet even by the most advanced technologies. This is another reason why face-name association still attracts many researchers: face-name association is a good touchstone of video analysis technologies.

This article describes about face-name association in news videos. In doing this, we take one of the earliest attempts as an example: Name-It. We briefly describe its mechanism. Then we compare it with corpus-based natural language processing and information retrieval techniques, and show the effectiveness of corpus-based video analysis.

## 2. Face-Name Association: Name-It Approach

Typical processing of face-name association is as follows:

* Extracts faces from images (videos)
* Extracts names from speech (closed-caption (CC) text)
* Associates faces and names

This looks very simple. Let’s assume that we have a segment of news video as shown in Figure 1. We don’t feel any difficulty in associating the face and name when we watch this news video segment, i.e., the face corresponds to “Bill Clinton” even though we don’t know the person beforehand. Video information is composed mainly of two streams: visual stream and speech (or CC) stream. Usually each one of these is not direct explanation of another. For instance, if visual information is shown as Figure 1, the corresponding speech will not be: “The person shown here is Mr. Clinton. He is making speech on...,” which is the direct explanation of the visual information. If so the news video could be too redundant and tedious to viewers. Instead they are complementary each other, and thus concise and easy to understand for people. However, it is very hard for computers to analyze news video segments. In order to associate the face and name computers need to understand visual stream so that a person shown is making speech, and to understand text stream that the news is about a speech by Mr. Clinton, and thus to realize the person corresponds to Mr. Clinton. This correspondence is shown only implicitly, which makes the analysis difficult for computers. This requires image/video understanding as well as speech/text understanding, which themselves are still very difficult tasks.

Name-It [] is one of the earliest systems tackling the problem of face-name association in news videos. Name-It assumes that image stream processing, i.e., face extraction, as well as text stream processing, i.e., name extraction, are not necessarily perfect. Thus the proper face-name association cannot be realized only from each segment. For example, from the segment shown in Figure 1, it is possible for computers that the face shown here can be associated with “Clinton” or “Chirac,” but the ambiguity between these cannot be resolved. To handle this situation, Name-It takes a corpus-based video analysis approach to obtain sufficiently reliable face-name association from imperfect image/text stream understanding results.

***2.1 Architecture***

The architecture of Name-It is shown in Figure 1. Since closed-captioned CNN Headline News is used as news video corpus, given news videos are composed of a video portion along with a transcript (closed-caption text) portion. From video images, the system extracts faces of persons who might be mentioned in transcripts. Meanwhile, from transcripts, the system extracts words corresponding to persons who might appear in videos. Since names and faces are both extracted from videos, they furnish additional timing information, i.e., at what time in videos they appear. The association of names and faces is evaluated with a “co-occurrence” factor using their timing information. Co-occurrence of a name and a face expresses how often and how well the name coincides with the face in given news video archives. In addition, the system also extracts video captions from video images.

***2.2 Example***

Extracted video captions are recognized to obtain text information, and then used to enhance the quality of face-name association. By the co-occurrence, the system collects ambiguous face-name association cues, each of which is obtained from each news video segment, over the entire news video corpus, to obtain sufficiently reliable face-name association results. Figure 3 shows the results of face-name association by using five hours of CNN Headline News videos as corpus.

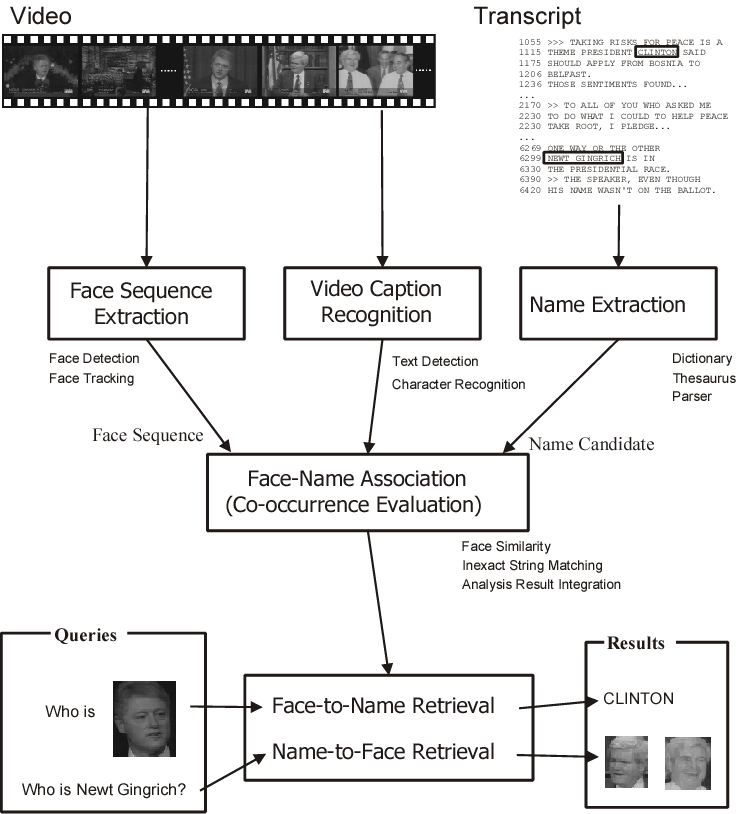


Figure 1. The architecture of Name-It.

A key idea of Name-It is to evaluate co-occurrence between a face and name by comparing the occurrence patterns of the face and name in news video corpus. To do so, it is obviously required to locate a face and name in video corpus. It is rather straight forward to locate names in closed-captioned video corpus, since closed-caption text is symbol information. In order to locate faces, a face matching technique is used. In other words, by face matching, face information in news video corpus is symbolized. This enables co-occurrence evaluation between faces and names. Similar techniques can be found in the natural language processing and information retrieval fields. For instance, the vector space model [] regards that documents are similar when they share similar terms, i.e., have similar occurrence patterns of terms. In Latent Semantic Indexing [], terms having similar occurrence patterns in documents within corpus compose a latent concept. Similar to these, Name-It finds face-name pairs having similar occurrence patterns in news video corpus as associated face-name pairs. Figure 2 shows occurrence patterns of faces and names. Co-occurrence of a face and name is realized by correlation between occurrence patterns of the face and name. In this example, “MILLER” and F1, “CLINTON” and F2, respectively, will be associated because corresponding occurrence patters are similar.



Figure 2. Face and name association results.

**3. Conclusions and Future Directions**

This article describes about face-name association in videos, especially Name-It, in order to demonstrate the effectiveness of corpus-based video analysis. There are potential directions to enhance and extend corpus-based face-name association. One possible direction is to elaborate component technologies such as name extraction, face extraction, and face matching. Recent advanced information extraction and natural language processing techniques enable almost perfect name extraction from text. In addition, they can provide further information such as roles of names in sentences and documents, which surely enhances the face-name association performance.

Advanced image processing or computer vision techniques will enhance the quality of symbolization of faces in video corpus. Robust face detection and tracking in videos is still challenging task (such as []. In [] a comprehensive survey of face detection is presented). Robust and accurate face matching will rectify the occurrence patterns of faces (Figure 3), which enhances face-name association. Many research efforts have been made in face recognition, especially for surveillance and biometrics. Face recognition for videos could be the next frontier. In [] a comprehensive survey for face recognition is presented. In addition to face detection and recognition, behavior analysis is also helpful, especially to associate the behavior with person’s activity described in text.

.

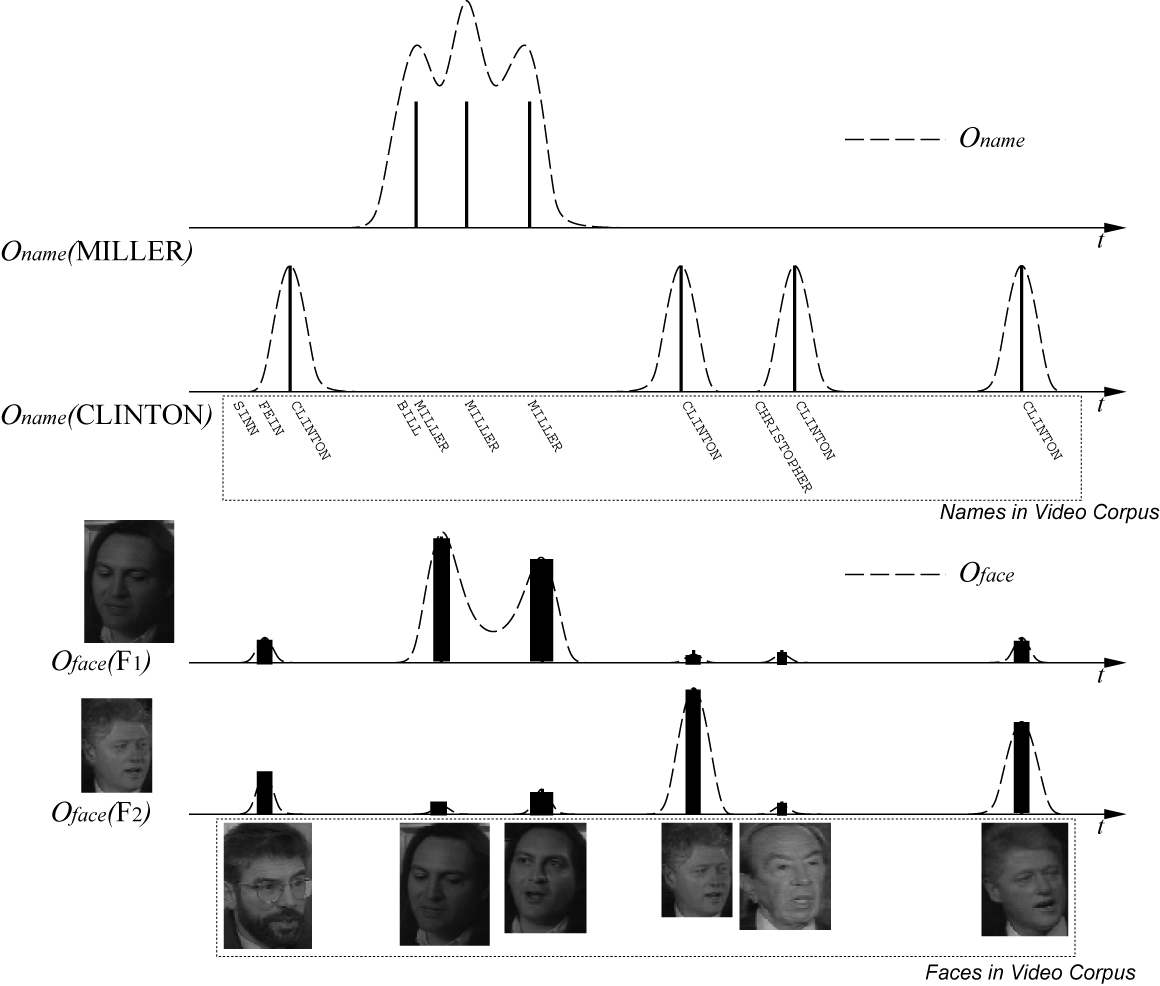


Figure 3. Face and name occurrence patterns.

Usage of the other modalities is also promising. In addition to images, closed-caption text, and video captions, speaker identification provides a powerful cue for face-name association for monologue shots [, ].

In integrating face and name detection results, Name-It uses co-occurrence, which is based on coincidence. However, as mentioned before, since news videos are concise and easy to understand for people, relationship between corresponding faces and names is not so simple as coincidence, but may yield a kind of video grammar. In order to handle this, the system ultimately needs to “understand” videos as people do. In [] an attempt to model this relationship as temporal probability distribution is presented. In order to enhance the integration, we need much elaborated video grammar, which intelligently integrate text processing results and image processing results.

It could be beneficial if corpus-based video analysis approach is applied to general objects in addition to faces. However, obviously it is not feasible to realize detection and recognition of many types of objects. Instead, in [] one of the promising approaches is presented. The method extracts interest points from videos, and then visual features are calculated for each point. These points are then clustered by features into “words,” and then a text retrieval technique is applied for object retrieval for videos. By this, the method symbolizes objects shown in videos as “words,” which could be useful to extend corpus-based video analysis to general objects.

## References

1. M. Li, D. Li, N. Dimitrova, and I. Sethi, “Audio-Visual Talking Face Detection,” Proceedings of the International Conference on Multimedia and Expo (ICME2003), 2003.
2. C. G. M. Snoek and A. G. Haptmann, “Learning to Identify TV News Monologues by Style and Context,” CMU Technical Report, CMU-CS-03-193, 2003.
3. J. Yang, M. Chen, and A. Hauptmann, “Finding Person X: Correlating Names with Visual Appearances,” Proceedings of the International Conference on Image and Video Retrieval (CIVR'04), 2004.
4. S. Satoh, Y. Nakamura, and T. Kanade, “Name-It: Naming and Detecting Faces in News Videos,” IEEE MultiMedia, Vol. 6, No. 1, January-March (Spring), 1999, pp. 22-35.
5. R. Baeza-Yates and B. Ribeiro-Neto, “Modern Information Retrieval,” Addison Wesley, 1999.
6. S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer and R. Harshman, “Indexing by Latent Semantic Analysis,” Journal of the American Society for Information Science, Vol. 41, 1990, pp. 391-407.
7. R. C. Verma, C. Schmid, and K. Mikolajcayk, “Face Detection and Tracking in a Video by Propagating Detection Probabilities”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 25, No. 10, 2003, pp. 1216-1228.
8. M.-H. Yang, D. J. Kriegman, and N. Ahuja, “Detecting Faces in Images: A Survey,” IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 24, No. 1, 2002, pp. 34-58.
9. J. Sivic and A. Zisserman, “Video Google: A Text Retrieval Approach to Object Matching in Videos,” Proceedings of the International Conference on Computer Vision (ICCV2003), 2003.
10. W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, “Face recognition: A literature survey,” ACM Computing Surveys, Vol. 35, No. 4, 2003, pp. 399-458.

Index terms (alphabetically):

Closed-caption news

Face name association

News video face

Etc.