Proceedings of IJCAI-97, pp. 1488–1493, 1997. Copyrighted by International Joint Conferences on Artificial Intelligence, Inc.

Name-It: Naming and Detecting Faces in Video by the Integration of Image and Natural Language Processing

Shin'ichi Satoh* and **Yuichi Nakamura**† and **Takeo Kanade**
School of Computer Science
Carnegie Mellon University
5000 Forbes Ave., Pittsburgh, PA 15213, USA
satoh@rd.nacsis.ac.jp yuichi@image.is.tsukuba.ac.jp tk@cs.cmu.edu

Abstract

We have been developing Name-It, a system that associates faces and names in news videos. First, as the only knowledge source, the system is given news videos which include image sequences and transcripts obtained from audio tracks or closed caption texts. The system can then either infer the name of a given face and output the name candidates, or can locate the faces in news videos by a name. To accomplish this task, the system extracts faces from image sequences and names from transcripts, both of which might correspond to key persons in news topics. The proposed system takes full advantage of advanced image and natural language processing. The image processing contributes to the extraction of face sequences which provide rich information for face-name association. The processing also helps to select the best frontal view of a face in a face sequence to enhance the face identification which is required for the processing. On the other hand, the natural language processing effectively extracts names by using lexical/grammatical analysis and knowledge of the news video topics structure. The success of our experiments demonstrates the benefits of the advanced image and natural language processing methods and their incorporation.

1 Introduction

Recent years have seen an increased demand for multimedia applications, including: video on demand, digital libraries, video editing/authoring, etc. The currently available multimedia data consists of a vast amount of image, video, audio, and text information, into which a modicum of essential "content" has been absorbed. An essential part of handling this large pool of information is to investigate the best way to access its contents. A content of a multimedia data may vary from person to person, and depends on its application; it may be key words of multimedia documents, key persons of news videos, etc. Without a doubt, vision/image processing and natural language processing play an important role in handling the contents of multimedia information. However, these techniques, by themselves, are still too immature to sufficiently handle contents. Since multimedia information is a mixture of video, audio, text, etc., a combination of these techniques is quite effective in achieving the desired goal.

To accomplish this task, Satoh et al. proposed Name-It [Satoh and Kanade, 1997], a system which associates names and faces in given news videos. Name-It's basic function is to guess "which face corresponds to which name" in given news videos. The use of Name-It demonstrated successful results and revealed the importance of combining image and text information. However, Name-It used only preliminary image and text processing.

In this paper, we describe how we extended Name-It by incorporating advanced image and natural language processing techniques. As with the former system, we assume that given news videos consist of video images and transcripts. Transcripts could be obtained from audio by using speech recognition; instead, we use closed-caption texts as transcripts. Potential applications of Name-It include

- Face candidate retrieval by name, and vice versa,
- Automated video indexing by the person's name,
- Automated creation of thousands of face-name correspondences database from thousands of hours of news videos.

We implemented the first of these as an example. The successful results we achieved showed the effectiveness of the integration of advanced image processing and natural language processing.

2 Overview of Name-It

The purpose of Name-It is to associate names and faces in news videos. A potential benefit might include, for example, naming all the politicians shown in Inauguration Day videos, even if they were not mentioned but had appeared in past videos. However, for our purposes here, we consider relatively simple applications, i.e., the system provides name candidates for a given face, or face candidates for a given

*National Center for Science Information Systems (NACSIS), 3-29-1 Otsuka, Bunkyo, Tokyo 112, Japan. The author had been a visiting scientist at CMU from April 1995 to April 1997.

†University of Tsukuba, Tsukuba City, 305, Ibaraki, Japan. The author had been a visiting scientist at CMU from March 1996 to December 1996.

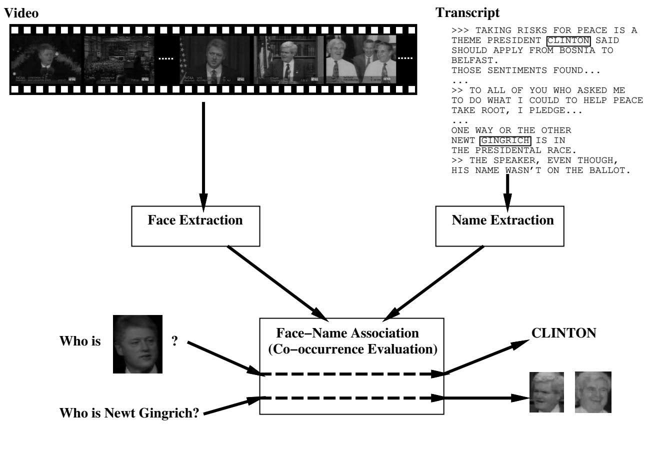


Figure 1: Architecture of Name-It

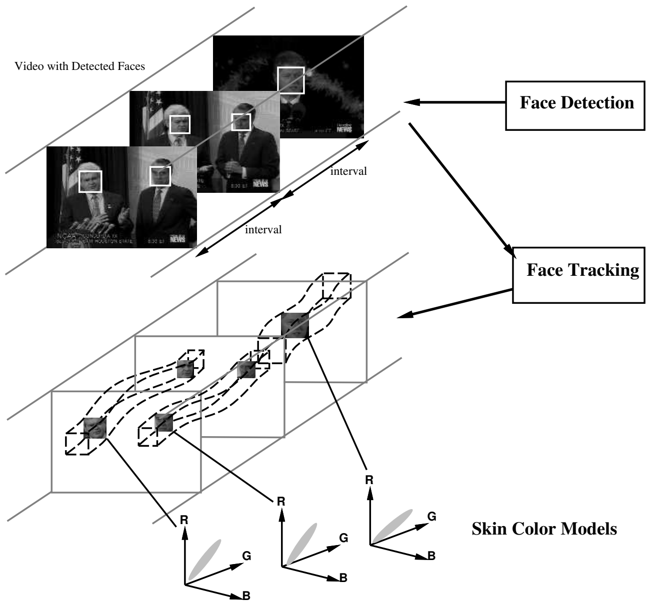


Figure 2: Face Tracking

name. To achieve this goal, the following procedures are required:

- From video images, the system extracts faces of persons who might be mentioned in transcripts,
- From transcripts, the system extracts words corresponding to persons who might appear in videos; then
- The system evaluates the association of the extracted names and faces.

Since both names and faces are extracted from videos, they furnish additional timing information, i.e., at what time (in frames) 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 well the name coincides with the face.

In the earlier version of Name-It, the *face extraction* was made by applying the face detector only to scene change images. It is clear that the system fails to extract a face which appears within a scene but not at scene changes. Moreover, it could not provide face duration information which would give rich hints for evaluating how well the face coincided with each name. Therefore, we extended the image processing portion primarily by incorporating face tracking. On the other hand, in the former version, the *name extraction* was made by using a dictionary to select proper nouns from transcripts. We enhanced its performance by incorporating more in-depth lexical/grammatical analysis that uses a dictionary, a thesaurus, and a parser.

Figure 1 shows the overall architecture of Name-It. The system is first given news videos; then it analyzes these videos, using the face extraction sub-system and the name extraction sub-system. After considering the results, the face-name association sub-system calculates co-occurrence and realizes retrieval of face-to-name candidates and name-to-face candidates.

3 Image Processing

The image processing portion of Name-It is necessary for extracting faces of persons who might be mentioned in transcripts. Those faces are typically shown under the following conditions: (a) frontal, (b) close-up, (c) centered, (d) long duration, (e) frequently. Given a video as input, the system outputs a two-tuple list: timing information (start $\sim$ end frame), and face identification information. Some of the conditions above will be used to generate the list; others will be evaluated later using information provided by that list.

3.1 Face Tracking

Face tracking consists of 3 components; face detection, skin color model extraction, and skin color region tracking (See Figure 2.). The following sub-sections describe the face tracking components.

Face Detection

First, Name-It applies face detection to every frame within a certain interval of frames. This interval should be small enough so that the detector does not fail to detect any important face sequences, yet at the same time large enough to ensure a reasonable processing time. Optimally, we apply the face detector at the intervals of 10 frames. The system uses the neural network-based face detector [Rowley *et al.*, 1995] which detects size-free, position-free, almost frontal, any number of faces in a given image. The detected face is output as a rectangular region that includes most of the skin, but excludes the hair and the background. The face detector can also detect eyes; we use only faces in which eyes are successfully detected to ensure that the faces are frontal and close-up. A detected face is tracked bi-directionally timewise to get a face sequence.

Skin Color Model Extraction/Tracking

Once a face is detected, the system extracts the skin color model. In several cases, researchers used the Gaussian model in (r, g) space ($r = R/(R+G+B)$, $g = G/(R+G+B)$) as

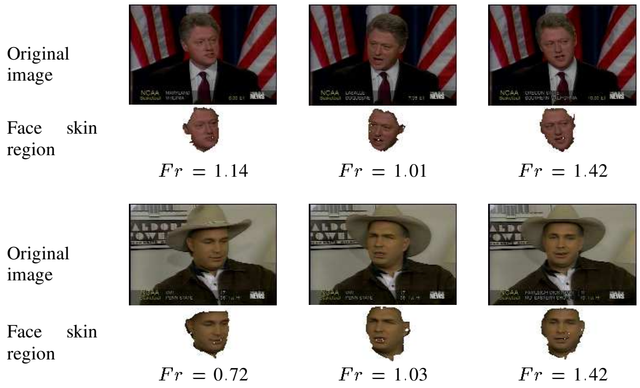


Figure 3: Frontal Face Selection

a general skin color model for face tracking [Yang and Waibel, 1995; Hunke, 1994]. Instead, for our research, the Gaussian model in (R, G, B) space is used because this model is more sensitive to brightness of skin color, and thus is much more suitable for the model tailored for each face.

Let F be the detected face region, and $I(x, y)$ be color intensities $[R\ G\ B]^T$ at (x, y). A skin color model consists of a covariance matrix C, a mean M, and a distance d.

$$M = \frac{1}{N} \sum_{(x,y)\in F} I(x, y)$$

$$C = \frac{1}{N} \sum (I(x, y) - M)(I(x, y) - M)^T$$

where N is the number of pixels in F. We used a constant for d. A model is extracted for each detected face, and is used to extract skin candidate pixels in the subsequent frame. A pixel $I(x, y)$ is a skin candidate pixel if $(I(x, y) - M)^T C(I(x, y) - M) < d^2$. Then a binary image of skin candidate pixels is composed, and noise reduction with region enlarging/shrinking and tracing contours of regions are applied to get skin candidate regions. The overlap between each of these regions and each of the face regions of the previous frame is evaluated to decide whether one of the skin candidate regions is the succeeding face region. In addition, the scene change detection based on the sub-region color histogram matching method [Smith and Kanade, 1995] is applied; this face region tracking is continued until a scene change is encountered or until no succeeding face region is found.

3.2 Face Identification

To infer the frequent occurrence of a face, face identification is necessary, i.e., we need to determine whether one face sequence is identical to another.

The Most Frontal Face Selection

To make face identification work most effectively, we need to use frontal faces. Although detected faces are not necessarily frontal enough, the best frontal view of a face, i.e., the *most frontal face* could be chosen from each face sequence. To choose the most frontal face from all the detected faces, the face skin region clustering method is first applied. For each detected face, cheek regions which are sure to have the skin color are located by using the eye locations. Using the cheek regions as initial samples, the region growing in the (R, G, B, x, y) space is applied to obtain the face skin region. We assume the Gaussian model in the (R, G, B, x, y) space; (R, G, B) contributes by making the region have skin color, and (x, y) contributes by keeping the region almost circular. Then the center of gravity (x_f, y_f) of the face skin region is calculated. Now let the locations of the right and left eyes of the face be (x_r, y_r), (x_l, y_l), respectively. We assume that the most frontal face has the smallest difference between x_f and $(x_l + x_r)/2$, and the smallest difference between y_l and y_r. To evaluate these conditions, we calculate the frontal factor Fr for every detected face;

$$w = \frac{5}{3}(x_l - x_r)$$

$$Fr = 1 - \frac{|2x_f - x_r - x_l|}{w} + \frac{1}{2}(1 - \frac{|y_l - y_r|}{w})$$

where w is the normalized face region width. The factor for an ideal frontal face is 1.5. The system chooses the face having the largest Fr to be the most frontal face of the face sequence. Figure 3 shows example faces, extracted face skin regions, and frontal factors.

Eigenface-Based Face Identification

We choose the eigenface-based method to evaluate face identification [Turk and Pentland, 1991]. Each of the most frontal faces is normalized into a 64 by 64 image by using the eye positions, then converted into a point in the 16-dimensional eigenface space. Face identification can be evaluated as the face distance, i.e., the Euclidean distance between two corresponding points in the eigenface space.

3.3 Evaluation

Figure 4 shows several results of the extended face extraction method compared with the former method. The start and the

Figure 4: Face Extraction Results

end frames of a face sequence and the selected frontal face frame are shown as the new face extraction results; a face frame is shown for the old face extraction. Figure 4(a) and (b) show that the former system failed to detect corresponding faces. The failure of (a) is due to the fact that the person looked down in the first frame, and the failure of (b) is due to scene changes using special effects (wiping, turning over) which could not be detected. Figure 4(c) and (d) show that, while the former system detected corresponding faces, the faces were not sufficiently frontal; the new system, on the other hand, extracted faces that were much more frontal. Total time for an SGI workstation (MIPS R4400 200MHz) to process a 30 minutes video was roughly thirty hours.

4 Natural Language Processing

The natural language processing portion extracts from transcripts name candidates corresponding to persons who might appear in videos. We will describe how the improved name candidate extraction uses lexical/grammatical analysis and the knowledge of the structure of a topic in news transcripts.

4.1 Typical Structure of News Videos

We use CNN Headline News videos for our experiments. The largest components of news deal with individual topics. We call these components simply topics. Each topic contains one or more paragraphs. A paragraph roughly corresponds to a scene. In closed-caption texts of CNN Headline News, the components can easily be distinguished; a topic is led by >>>, and a paragraph is led by >> (See Figure 5.). A typical paragraph at the beginning of the topic is an anchor paragraph, in which an anchor person gives an overview of the topic. After an anchor paragraph, live video paragraphs, which are actual videos related to the topic, or speeches by the person of interest in that topic, are typically presented. A live video paragraph, especially one that includes someone's speech, is quite important for Name-It; this paragraph almost certainly contains a close-up scene of that person. However, we should note that the person rarely mentions his/her own name in the speech; thus corresponding transcripts may not contain desired name. The extra care needed to handle this situation is described in the following sub-sections.

4.2 Conditions of Name Candidates

Each name candidate should satisfy some of the following conditions:

1. The candidate should be a noun that represents a person's name or that describes a person (president, fireman, etc.).
2. The candidate should preferably be an agent of an act, especially an act of speech, of attendance at a meeting, or of a visit. For example, a speaker is usually centered in the speech scene, while the other people are not always shown in videos even if they are mentioned.
3. The candidate tends to be mentioned earlier than others in the topic in transcripts. (In a news video, important information which might have corresponding images is usually mentioned earlier, rather than later.)

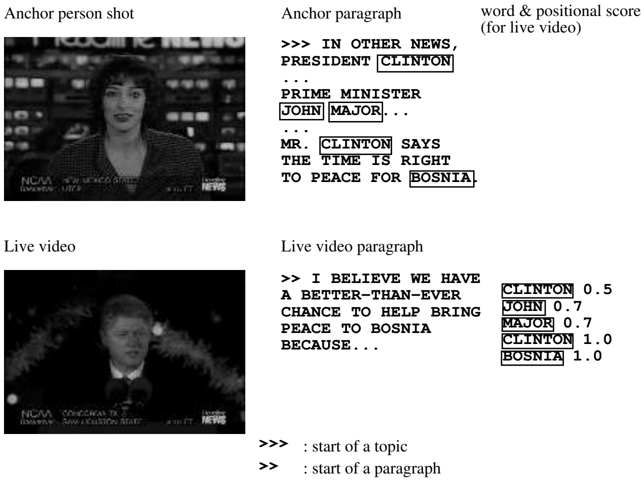


Figure 5: Positional Score for Live Video

4. The candidate tends to be mentioned just before a live video is shown. The person appeared in a live video rarely mentions his/her own name. Instead, just before the live video, an anchor person tends to appear and introduce him/her (See Figure 5.).

The system evaluates these conditions for each word that occurs in transcripts by using the dictionary (the Oxford Advanced Learner's Dictionary [Oxford]), the thesaurus (WordNet [Miller, 1990]), and the parser (Link Parser [Sleator, 1993]). Finally, the system outputs the three-tuple list: a word, timing information (frame), and a normalized score.

4.3 Score Calculation

Referring to the dictionaries and the parsing results, the system calculates the score for each word in transcripts. The score is normalized and that close to 1.0 corresponds to a word which very likely corresponds to a face. The score calculation is defined as follows:

Grammatical Score: After consulting the dictionary, the system gives 1.0 to proper nouns, 0.8 to common nouns, and 0 to other words. And by consulting the parsing results, the system gives 1.0 to nouns, and 0.5 to other words. The net grammatical score is the product of these two.

Lexical Score: After consulting the thesaurus, the system gives 1.0 to persons, 0.8 to social groups, and 0.3 to other words.

Situational Score: The act corresponding to the word is represented by the verb in the sentence which includes the word. By looking the verb up in the thesaurus, the system gives 1.0 to speech, 0.8 to attendance at meetings, and 0.3 otherwise.

Positional Score: The system gives 1.0 to words that appear in the first sentence in a topic, 0.5 to words that appear in the last sentence, and linearly interpolated score to other words according to the position of the sentence where the word appears. As for a live video, the system also outputs the same tuples as those of the paragraph which appears before the live video (possibly the anchor paragraph), replacing the timing information with that of the live video (See Figure 5.). In addition, it replaces the positional score according to the position of the sentence in the anchor paragraph: 1.0 for the sentence just before the live video, 0.5 for the first sentence of the topic, and linearly interpolated score otherwise.

Finally, the net score is calculated as the product of all 4 scores. The execution time for a 30-minute news video is approximately 1.5 hour on an SGI workstation (MIPS R4400 200MHz). Most of that time is consumed by parsing.

5 Face-Name Association

5.1 Algorithm

In this section, the algorithm for retrieving face candidates by a given name is described. We use the co-occurrence factor [Satoh and Kanade, 1997] with an extension to handle face duration and name score. Let N and F be a name and a face, respectively. The co-occurrence factor $C(N, F)$ is expected to have a degree which represents the fact that the face F is likely to have the name N. Think of the faces $F_a, F_b, \cdots$ and the names $N_p, N_q, \cdots$, and F_a corresponds to N_p. Then $C(F_a, N_p)$ should have the largest value among co-occurrence factors of any combinations of F_a and the other names (e.g., $C(F_a, N_q)$, etc.), or of the other faces and N_p (e.g., $C(F_b, N_p)$, etc.). Retrieval of face candidates by a given name is realized as follows using the co-occurrence factor:

1. Calculate co-occurrences of combinations of all face candidates and the given name.
2. Sort co-occurrences.
3. Output faces that correspond to the top-N largest co-occurrences.

Retrieval of name candidates by a face is realized as well.

5.2 Co-occurrence Calculation

In this section, the co-occurrence factor $C(N, F)$ of a face F and a name N is defined. Assume that we have the two-tuple list of face sequences (timing, face identification): $\{(t_{F_i}, F_i)\} = \{(t_{F_1}, F_1), (t_{F_2}, F_2), \ldots\}$, and the three-tuple list of name candidates (word, timing, score): $\{(N_j, t_{N_j,k}, s_{N_j,k})\} = \{(N_1, t_{N_1,1}, s_{N_1,1}), (N_1, t_{N_1,2}, s_{N_1,2}), \ldots (N_2, t_{N_2,1}, s_{N_2,1}), \ldots\}$. Note that t_{F_i} has duration ($t_{start,F_i} \sim t_{end,F_i}$); so we can then define the duration function as $dur(t_{F_i}) = t_{end,F_i} - t_{start,F_i}$. Also note that a name N_j may occur several times in a video, so each occurrence is indexed by k. Let $d_f(F_i, F_j)$ be the distance between the points in the eigenface space corresponding to the faces F_i, F_j, and $d_t(t_{F_i}, t_{N_j,k})$ be the distance between the timing of the face F_i and the word N_j of the k-th occurrence. Actually $d_t(t_{F_i}, t)$ is defined as follows;

$$d_t(t_{F_i}, t) = \begin{cases} 0 & (t_{start,F_i} \le t \le t_{end,F_i}) \\ t_{start,F_i} - t & (t < t_{start,F_i}) \\ t - t_{end,F_i} & (t_{end,F_i} < t) \end{cases}$$

Then the co-occurrence factor $C(N, F)$ of the face F and the name candidate N is defined as follows;

$$S_f(F_i, F_j) = e^{\frac{d_f^2(F_i, F_j)}{2\sigma_f^2}}$$

$$S_t(t_i, t_j) = e^{\frac{d_t^2(t_i, t_j)}{2\sigma_t^2}}$$

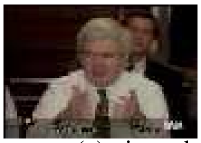
$$C(N, F) = \frac{(\sum_i S_f(F_i, F) \sum_k s_{N,k} S_t(t_{F_i}, t_{N,k}))^p}{\sum_i S_f(F_i, F) \sum_k s_{N,k} dur(t_{F_i})}$$

where σ_t and σ_f are standard deviations of the Gaussian filter in time and in the eigenface space, respectively, and p is a constant ($p > 1$). Intuitively, the numerator of $C(N, F)$ becomes larger if F is identical to F_i AND F_i coincides with N having the larger score. To prevent "anchor person problem," (An anchor person coincides with almost any name. A face/name coincides with any name/face should correspond to NO name/face.) $C(N, F)$ is normalized with the denominator. p should be greater than 1 to make the co-occurrence of a face and a name which frequently coincide larger than the co-occurrence of a face and a name which coincide only once. $p = 1.5 \sim 2.0$ worked fine with our experiments. The detailed explanation of the equations is appeared in [Satoh and Kanade, 1997].

6 Experiments

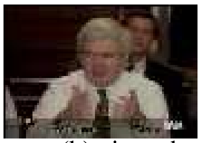
We implemented the Name-It System on an SGI workstation. We processed 10 CNN Headline News videos (30 minutes each) in a total of 5 hours. From them, the system extracted 556 face sequences, and was given 752 name candidates. Name-It performs name candidate retrieval by a given face, and face candidate retrieval by a given name as example applications. Since the face extraction is evaluated with the results in Section 3.3 and its contribution to face-name association is obvious, we demonstrated the effect of the improved name extraction. The results obtained by taking full advantage of the improved methods are compared with the results obtained by using a combination of the improved face extraction and the former name extraction.

Figure 6(a) and (b) show the results of name candidate retrieval by using the face of Newt Gingrich, and Figure 6(c) and (d) show the results of face candidate retrieval by the name "CLINTON" in order from left to right. Figure 6(a) and (c) are obtained by using the new name extraction while Figure 6(b) and (d) are obtained by using the old name extraction. The right answer "GINGRICH" is listed as the second candidate in (a), while it did not appear in the top thirty candidates in the results of (b). Figure 6(c) shows that the top 3 candidates are correct; in fact, 6 out of the top 8 are correct. On the other hand, Figure 6(d) shows that only the third candidate is correct; moreover, it is the only correct candidate among the top 16 candidates. Three of Clinton's faces shown in (c) appeared along with his speech; thus the live video treatment mentioned in Section 4.3 worked well in this example. Figure 6(e) and (f) show the other examples.



1	SPEAKER	0.00035289
2	GINGRICH	0.000214526
3	RACE	0.000155937
4	NUMBERS	0.000154024

(a) given the face w. new name extraction



1	GOP	0.000114514
2	CEO	8.13667e-05
3	IRS	4.74364e-05
4	COSSACK	4.26947e-05

(b) given the face w. former name extraction

(c) given "CLINTON" w. new name extraction

(d) given "CLINTON" w. former name extraction

(e) given "ERHARDT" w. new name extraction

Ron Erhardt, NFL's Pittsburgh Steelers

1	BROOKS	0.000269475
2	GARTH	0.000151196
3	FANS	0.000105675
4	PENTAGON	9.6434e-05

1	CHRISTOPHER	0.00107707
2	CONGRESS	0.000162332
3	WARREN	0.000149605
4	OFFICIALS	0.000136414

1	ROOKIE	0.00271504
2	SUTTON	0.00134173
3	NOMO	0.00112312
4	PITCHER	0.000913124

(f) given the faces w. new name extraction

Garth Brooks, singer, Warren Christopher, Secretary of State, and Hideo Nomo, pitcher of L.A. Dodgers

Figure 6: Face-Name Association Results

7 Conclusions

This paper describes Name-It, a system that associates faces and names in news videos. The system has been extended by incorporating advanced image processing and natural language processing. The image processing contributes to extracting face sequences, and to selecting the most frontal face in a face sequence for improving face identification. The natural language processing utilizes a dictionary, a thesaurus, and a parser for lexical/grammatical analysis as well as knowledge of the news video topics structure. The enhancement achieved by those techniques is demonstrated by providing actual sample results. Those successful results reveal the importance of an approach that integrates image and natural language processing, and show that we are headed in the right direction to achieve our goal of accessing real contents of multimedia information.

Acknowledgement

The authors whould like to thank Michiyo Kimoto and Imari Sato for their help.

This material is based upon work supported by the National Science Foundation under Cooperative Agreement No. IRI-9411299. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

[Oxford] The Oxford Text Archive. http://ota.ox.ac.uk/.

[Hunke, 1994] H. Martin Hunke. Locating and tracking of human faces with neural networks. Technical Report CMU-CS-94-155, School of Computer Science, Carnegie Mellon University, 1994.

[Miller, 1990] G. Miller. Wordnet: An on-line lexical database. *International Journal of Lexicography*, 3(4), 1990.

[Rowley *et al.*, 1995] H. Rowley, S. Baluja, and T. Kanade. Human face detection in visual scenes. Technical Report CMU-CS-95-158, School of Computer Science, Carnegie Mellon University, 1995.

[Satoh and Kanade, 1997] Shin'ichi Satoh and Takeo Kanade. Name-It: Association of face and name in video. Proc. of CVPR'97, 1997.

[Sleator, 1993] D. Sleator. Parsing english with a link grammar. In *Third International Workshop on Parsing Technologies*, 1993.

[Smith and Kanade, 1995] M. Smith and T. Kanade. Video skimming for quick browsing based on audio and image characterization. Technical Report CMU-CS-95-186, School of Computer Science, Carnegie Mellon University, 1995.

[Turk and Pentland, 1991] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.

[Yang and Waibel, 1995] Jie Yang and Alex Waibel. Tracking human faces in real-time. Technical Report CMU-CS-95-210, School of Computer Science, Carnegie Mellon University, 1995.