Name-It: Association of Face and Name in Video

Shin'ichi Satoh   Takeo Kanade

December 20, 1996

CMU-CS-96-205

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Abstract

This paper proposes a novel approach to extract meaningful content information from video by collaborative integration of image understanding and natural language processing. As an actual example, we developed a system that associates faces and names in videos, called Name-It, which is given news videos as a knowledge source, then automatically extracts face and name association as content information. The system can infer the name of a given unknown face image, or guess faces which are likely to have the name given to the system. This paper explains the method with several successful matching results which reveal effectiveness in integrating heterogeneous techniques as well as the importance of real content information extraction from video, especially face-name association.

This material is based upon work supported by the National Science Foundation under Cooperative Agreement No. IRI-9411299.

Keywords: video indexing, content-based retrieval, face detection and identification.

Given:

Video

Transcript

BUT ASK BILL MILLER TO LABEL
HIS MUSIC AND --
>> RIGHT NOW I'D DESCRIBE
MY MUSIC AS VISIONARY ROCK.
>> Reporter: THAT VISION LOOKS
BACK AS MUCH AS FORWARD,
TO MILLER'S ROOTS ON A WISCONSIN
INDIAN RESERVATION.

THE NATIONAL LEAGUE ROOKIE
OF THE YEAR AWARD WENT
TO A DODGER, PITCHER HIDEO NOMO.
THE JAPANESE-BORN PITCHER POSTED
A 13-6 RECORD AND LED THE LEAGUE
WITH 236 STRIKEOUTS.
NOMO RECEIVED 18 OF THE POSSIBLE
28 FIRST-PLACE VOTES.

ONE WAY OR THE OTHER
NEWT GINGRICH IS IN
THE PRESIDENTAL RACE.
>> THE SPEAKER, EVEN THOUGH
HIS NAME WASN'T ON THE BALLOT.
THE SPEAKER WAS AN ISSUE
ON TUESDAY AND THE SPEAKER WILL
BE AN ISSUE IN EVERY RACE NEXT
YEAR.
>> Reporter: AND JUDGING
FROM OTHER NUMBERS
IN THE POLL -- GINGRICH --
AT LEAST WHAT THE PUBLIC THINKS
HE STANDS FOR -- IS A FAIRLY GOOD
ISSUE FOR DEMOCRATS.

Questions:

What is the name of  ?

Who is NEWT GINGRICH ?

Figure 1: "Who" information

1 Introduction

As digital video libraries are becoming realistic, supported by several technological innovations including MPEG video compression, vast and high speed disk arrays, high speed local/wide area networks, etc., content based video indexing is becoming much more important. Many research efforts are made to achieve this goal. They include image retrieval based on image features including color histogram and texture analysis [1], feature extraction using Karhunen-Loéve (K-L) transform [2], and video structuring based on scene change detection [3]. These approaches provide successful results to some extent, e.g., by using an image retrieval based on color histograms, a forest image may match well with images having trees, a sea side image may match well with marine images, etc. This is because the mapping in color histogram conversion gives adjacency relations similar to a person's cognitive adjacency relations between images, though, these approaches cannot be said as real content-based indexing since they don't extract real content information.

We are aiming at real content information extraction from news videos. News videos give us important content information, e.g., President ⋯ went to ⋯ to attend ⋯ meeting, Prime Minister ⋯ said ⋯ at that meeting, Senate leader ⋯ talked about ⋯, etc. Looking at these types of content, it can be said that "who" information, i.e., the face and name association, is one of the most important information which can be acquired from news videos (Figure 1). A person can easily say which person is shown in a video, e.g., a person may identify Clinton, Gingrich, and Dole by watching and hearing only 5 minutes of news showing them, even though the person does not know their faces beforehand. However, this task contains several difficult steps;

- Scene/topics detection in videos,
- Face detection in videos,
- Identification of faces occur in different scenes/videos,
- Natural language processing to detect names in narrations, and
- Following context to select a person corresponding to a certain name who occurred in video.

These procedures require cooperative integration of video structuring, machine vision, and natural language processing. As the first step to delineate real content information, we introduce a face and name association system, Name-It.

Input videos are composed of transcripts as text information and image sequences. Transcripts may be extracted from speech in the sound track of videos, or from the closed-captions. We use primarily closed-captions as transcripts. Extraction of face images from image sequences, as well as extraction of name candidates from transcripts are explained in Section 4. Then a matching measurement between each face image and name candidate is introduced followed by an actual face-to-name or name-to-face association algorithm. Finally experimental results are shown to evaluate this method.

2 Related Work

Face detection/identification has been researched for a long time, and there are image database systems which can perform face similarity matching; they include MIT Photobook [2] and Virage system. These two systems use the eigenvector based method for face similarity matching [4]. It is noteworthy that Photobook applied their method to more than 7,500 images of about 3,000 people and got successful results. The results reveal that eigenvector based face similarity matching works well to some extent.

Piction system [5] identifies faces within given captioned photos, typically of newspapers. The system extracts faces from a photo and analyzes captions to get geometric constraints among faces which will appear in the photo, then label each face as each name. As far as integrating image processing and text processing, this approach is similar to our method. However, Piction deeply depends on text information, in other words, natural language processing, because the system requires the captions to have geometric descriptions, e.g., "top row, from left, are Michael, Brian, ..." In contrast, Piction uses image processing for face location, but not face identification.

3 Overview of Name-It

Figure 2 shows a typical composition of a news video. A news video consists of image sequences which may contain persons' faces, and transcripts which may contain persons' names. Given these news videos, an ideal face and name association system takes a face as input, then outputs a name of that face, or takes a name then outputs a face of that name (Figure 3). Obviously a human can do this very easily since news videos are created to be understandable for humans; even though we do not know the faces of persons in videos, we can identify each person. Although human seem to be able to do this easily, this process includes several complicated, and high level processes, as follows:

Scene/topic detection: Given videos will be perceived as many scenes, e.g., a scene in which a news caster is giving the introduction of a topic, a scene of live video of the topic, etc. Then those scenes will be organized into a topic, e.g., the visit of president Clinton to Ireland, etc.

Face detection: Face portions within videos are immediately recognized. In addition to this, tracking of each face is performed within an image sequence.

Face identification: If an unknown face is given, while another face which is identical to that face was known to be "foo," then the given face can be thought to be "foo." To achieve this, face identification is necessary. In another case, a footage tells face-A may be "bar," another footage tells face-B probably be "bar" also, and face-A and face-B are identical, the likelihood that both are "bar" is increased.

Name extraction: Names, or proper nouns, will be discriminated from transcripts using a priori knowledge or context tracking that tells which words are most likely proper nouns. Context information may help in distinguishing names which are likely to represent "the person of interest of that topic."

Typical face and name association will be carried out by human within news video footage (Figure 2) as follows:

1. In the first scene, a news caster appears and talks about Mr. Clinton. It shows that this topic is likely to be about Mr. Clinton and his face will appear in the subsequent scenes.

2. The next scene mainly shows a certain person meanwhile the caster is still talking about Mr. Clinton, thus the person may be Mr. Clinton.

3. The next scene shows only one person talking, so this person is likely to be the person of interest of this topic. Also, it is recognized that this person is identical to the previous person, therefore this person who might be the person of interest is probably Mr. Clinton.

Figure 2: Typical Composition of News Video

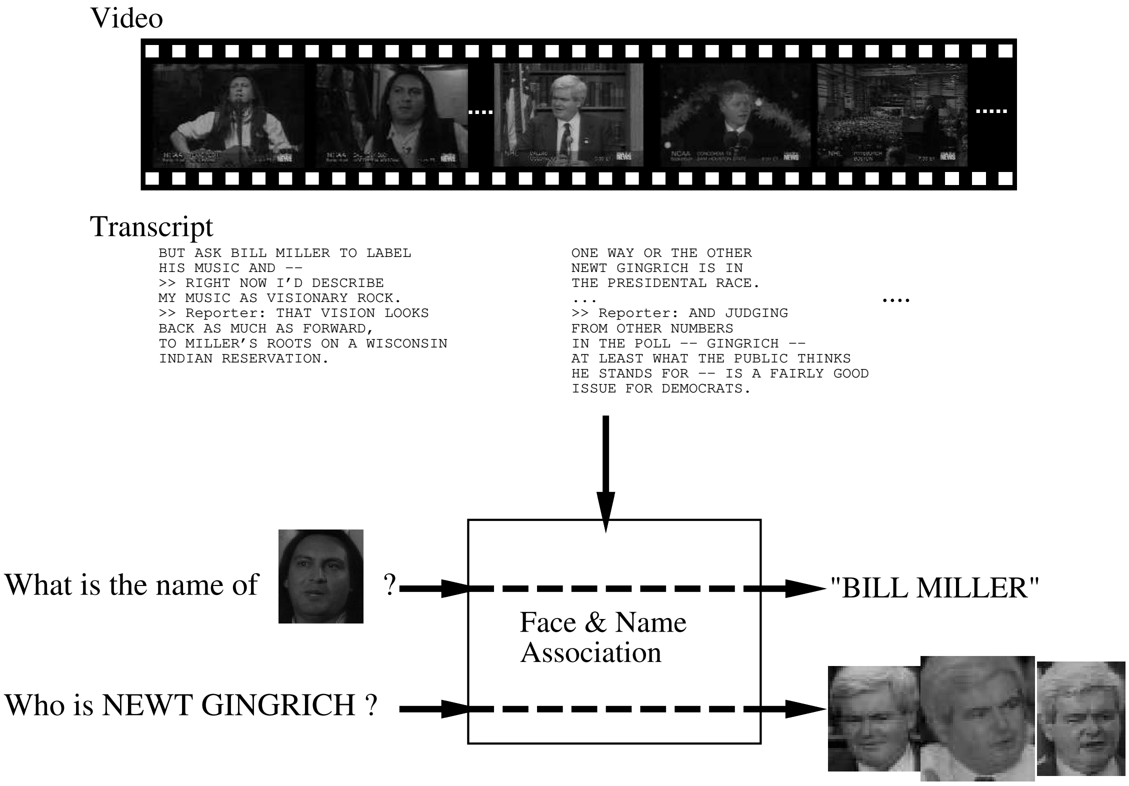


Figure 3: Face and Name Association

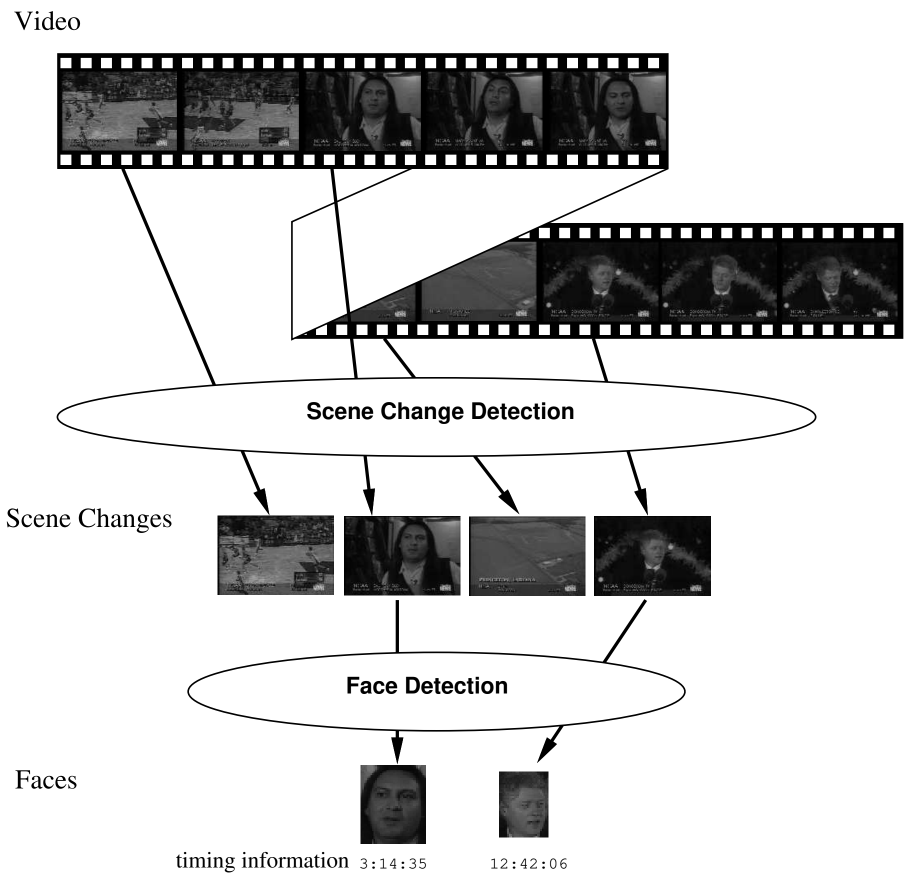


Figure 4: Faces Extraction from Video

Although human can do this easily, it is very hard for computers to achieve automatically. For example, face extraction and identification is a critical vision problem which has not yet been completely resolved, name extraction includes difficult natural language processing problems, and scene/topic detection may contain both vision and natural language processing problems. There are several techniques which achieve the above goals including scene change detection, face detection, etc. They are still far from complete, but it may be promising to properly integrate those techniques to get useful results. We took this strategy to bring face and name association to fruition. We used intensity histogram difference based scene change detection [6], neural-network based face detection, eigenvector method based face identification, and dictionary based name extraction, despite these being still incomplete. Then we integrate these techniques to collaboratively use image sequences and transcript information of video footage.

4 Preparations

4.1 Face Extraction

First we have to extract faces from a given video footage. We use a neural-network

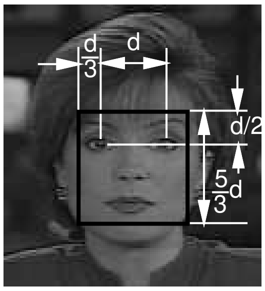


Figure 5: Face normalization based on eye location

based face detector to locate faces in images [7]. This detector can locate front view faces of various sizes with some margin in rotation, though, it takes about 10 seconds with a workstation (MIPS R4400 200MHz) to process a 352×240 image. Thus we first apply an intensity histogram based scene change detector [6] to the video to obtain scene change images, then apply the face detector to scene change images. The overall face extraction process is shown as Figure 4. For example, we processed 4.5 hours news video, including about 490,000 frames (30 fps), and obtained 4,318 scene changes. We applied the face detector to those scene change images to obtain 320 faces. To preserve high quality face images, we use large faces (detected as more than 36 by 36 pixels) of which both eyes are successfully detected by the face detector. In this example, there is no false detection among all detected faces.

Then we normalize faces using eye locations (Figure 5). Let right and left eye location respectively be (r_x, r_y) and (l_x, l_y). Eye distance d, between right and left eyes, is defined as

$$d = l_x - r_x. \tag{1}$$

Normalized face location $(x_l, y_l) - (x_u, y_u)$ is defined as

$$x_l = r_x - \frac{1}{3}d \tag{2}$$

$$x_u = l_x + \frac{1}{3}d \tag{3}$$

$$y_u = \frac{r_x + l_x}{2} + \frac{1}{2}d \tag{4}$$

$$y_l = y_u - \frac{5}{3}d. \tag{5}$$

Then face images are normalized into 64×64 images for face similarity evaluation.

4.2 Face Similarity

We used an eigenvector based method to compute the distance between two faces. This is based on the well known eigenface method [4]. Using this distance, we defined a similarity between two faces.

Let face images be two-dimensional N by N arrays of intensity values. These images also can be regarded as vectors of N^2-dimension. Let the training set of M

face images be $F_1, F_2, \ldots, F_M$. We can define the covariance matrix C as follows;

$$\bar{F} = \frac{1}{M}\sum_{n=1}^{M} F_n \tag{6}$$

$$\tilde{F}_i = F_i - \bar{F} \tag{7}$$

$$\begin{aligned} C &= \frac{1}{M}\sum_{n=1}^{M} \tilde{F}_n \tilde{F}_n^T \\ &= \frac{1}{M} A A^T \end{aligned} \tag{8}$$

where $A = [\tilde{F}_1 \tilde{F}_2 \ldots \tilde{F}_M]$. The eigenvectors, $\mathbf{u}_k$, and corresponding eigenvalues λ_k ($\lambda_1 > \lambda_2 > \ldots \lambda_M$) can be derived from C. We call the eigenvectors, $\mathbf{u}_k$, the eigenfaces.

Assume that an unknown face image, F, is given. Consider $M'(< M)$ eigenfaces having top M' largest eigenvalues. We can define an M'-dimensional eigenface subspace by converting F into

$$\Phi = [\phi_1 \phi_2 \ldots \phi_{M'}]^T \tag{9}$$

where

$$\phi_i = \mathbf{u}_i^T (F - \bar{F}). \tag{10}$$

The distance between two face images annotated by i and j can be derived using corresponding points Φ_i and Φ_j within the eigenface subspace as follows;

$$d_{ij}^2 = \|(\Phi_i - \Phi_j)\|^2. \tag{11}$$

Then the similarity between them is defined;

$$S(F_i, F_j; \sigma_f) = e^{-\frac{d_{ij}^2}{2\sigma_f^2}} \tag{12}$$

where σ_f is a standard deviation of Gaussian filter in the eigenface subspace, which will be empirically determined. We also have to determine the dimension of the eigenface subspace M', and we used $M' = 16$ for our experimental Name-It embodiment.

4.3 Name Candidates Extraction

Along with face extraction, we extract proper nouns, which we use as name candidates, from news transcripts. Though we use closed-caption text as transcripts, we can extract transcripts from sound tracks using a proper speech recognition system. Closed-caption consists of lines, each of which has time code information for when to show each line on the screen.

The system eliminates special characters and numerics from texts, then applies an English dictionary to separate proper nouns. We use the Oxford Text Archive text710 dictionary, which is freely distributed and includes more than 70,000 words. A given word is assumed to be a name (or a proper noun) if (1) the word is annotated as a proper noun in the dictionary, or (2) the word does not appear in the dictionary.

Figure 6 depicts the diagram of name candidate extraction, an example of a transcript, and extracted proper nouns.

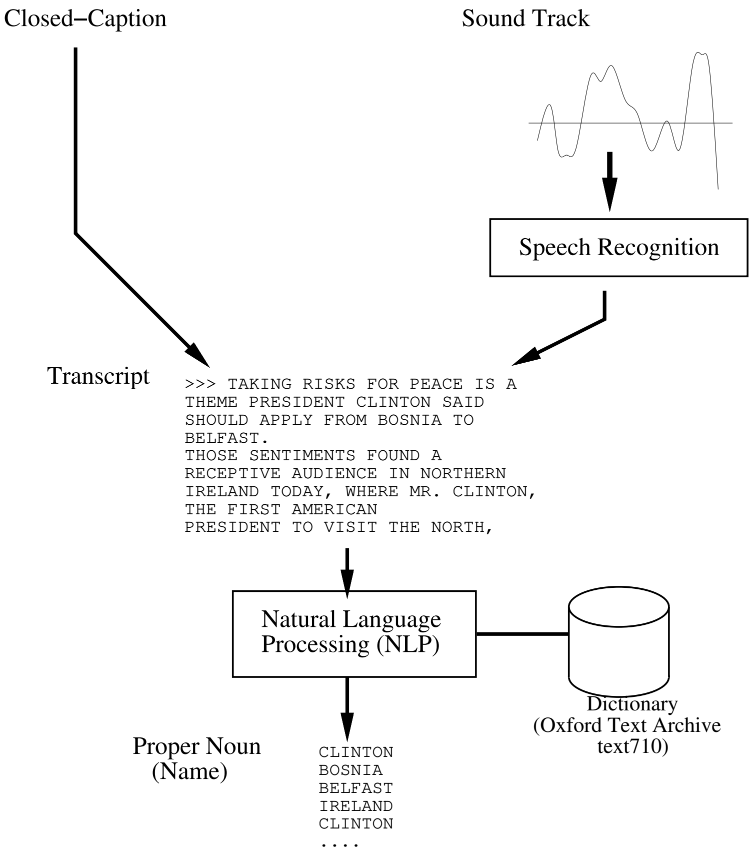
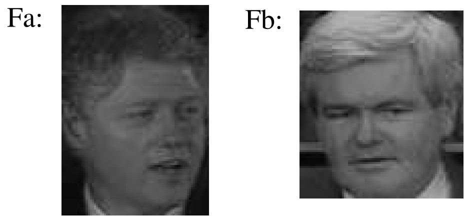


Figure 6: Name Candidates Extraction



Co–occurrence C(Face, Name)

C(Fa, CLINTON) > C(Fa, GINGRICH), C(Fa, DOLE), ...

C(Fa, CLINTON) > C(Fb, CLINTON), C(Fc, CLINTON), ...

Figure 7: An Example of Face and Name Co-occurrence

5 Face and Name Co-occurrence

5.1 Overview of the Method

To achieve face and name association (e.g., to get associated faces by giving a certain name, or vice versa), we introduce a face and name co-occurrence factor $C(N, F)$ of a face F and a name N. To acquire co-occurrence, we first inspect occurrence of extracted names within transcripts and occurrence of extracted faces within videos along the same time scale. Then co-occurrence is calculated to represent how well the name and the face "co-occur" in time, i.e., a face F tends to appear in the videos when a name N appears in the transcripts and vice versa, while F tends not to appear when N does not appear. This factor $C(N, F)$ is expected to give larger value when the face F tends to have the name N. For example, think of faces $F_a, F_b, \ldots$ and names Clinton, Gingrich,... as shown in Figure 7. Obviously F_a corresponds to Clinton, and F_b to Gingrich. In this case, the co-occurrence $C(F_a,$Clinton$)$ gives a larger value than any other combinations of co-occurrence between F_a and any other names, or between any other faces and "Clinton."

Once co-occurrence is defined, face-to-name or name-to-face association can be acquired in a straightforward manner. Consider the situation where a face is given, and the associated names are desired. To achieve this, we compute co-occurrence between the face and every name candidate, then sort names by co-occurrence to get names having top-N largest co-occurrence values. These names may be a good estimation of the name of the face. Name-to-face association can be achieved similarly.

In this section, occurrences of names and faces are defined as functions over time. We call these occurrence functions, and use them to define the co-occurrence factor.

5.2 Occurrence Functions

Assume that all occurrences of a word N are extracted from a given transcript. Let occurrences of N be $t_{N,1}, t_{N,2}, \cdots$, i.e., a word N occurs at $t_{N,1}, t_{N,2}, \cdots$ in the video.

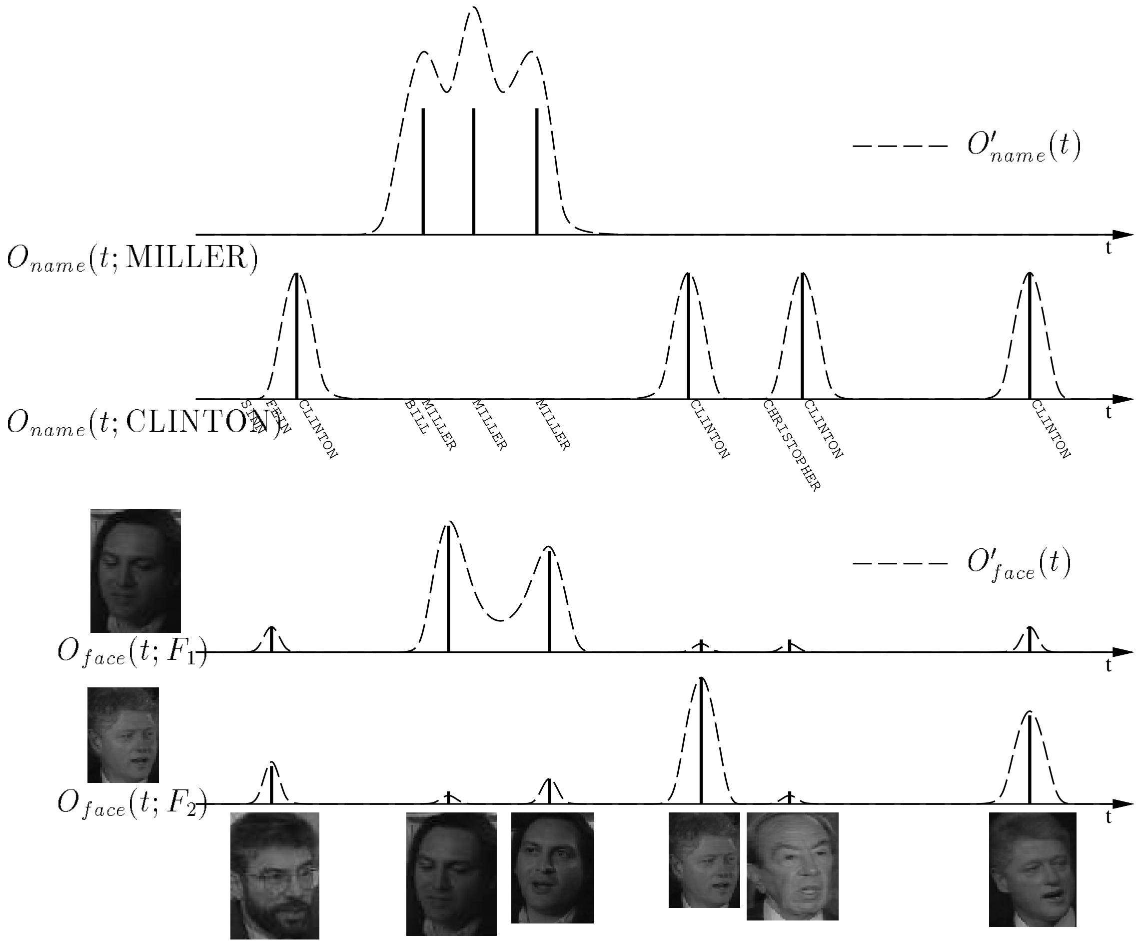


Figure 8: Occurrence Function

The name occurrence function is defined as

$$O_{name}(t; N) \;=\; \sum_i \delta(t - t_{N,i}) \tag{13}$$

where $\delta(t)$ is a Dirac delta function.

Let $F_1, F_2, \cdots$ be a set of faces. We can evaluate face similarity $S(F_i, F_j)$ as given in Section 4.2, e.g., $S(F_i, F_j) > S(F_i, F_k)$ if F_i is more similar to F_j than to F_k. Assume that faces $F_1, F_2, \cdots$ occur at $t_{F_1}, t_{F_2}, \cdots$ within the video footage. The face occurrence function of a face F is defined as

$$O_{face}(t; F) \;=\; \sum_i S(F, F_i)\delta(t - t_{F_i}). \tag{14}$$

Name and face occurrence functions are dispersed using a Gaussian filter.

$$O'_{name}(t; N) \;=\; O_{name}(t; N) \otimes e^{-\frac{t^2}{2\sigma_t^2}}, \tag{15}$$

$$O'_{face}(t; F) \;=\; O_{face}(t; F) \otimes e^{-\frac{t^2}{2\sigma_t^2}} \tag{16}$$

where σ_t is a standard deviation for a temporal Gaussian filter, and $\otimes$ represents the convolution operator. Intuitively, $O'_{name}(t; N)$ represents the likelihood of occurrence of the word N, i.e., if $O'_{name}(t_1;N) > O'_{name}(t_2;N)$, the video at t_1 is more likely about N than at t_2. Similarly, $O'_{face}(t;F)$ represents likelihood of occurrence of the face F, i.e., if $O'_{face}(t_1;F) > O'_{face}(t_2;F)$, the video at t_1 is more likely about F than at t_2. Figure 8 illustrates the relationship between occurrence of names and faces, and occurrence functions.

5.3 Co-occurrence

From defined name and face occurrence functions, we define a name-face co-occurrence $C(N,F)$.

$$C(N,F) = \frac{(\int_{D_t} O'_{name}(t;N) \cdot O_{face}(t;F)dt)^p}{\int_{D_t} O'_{name}(t;N)dt \int_{D_t} O_{face}(t;F)dt} \tag{17}$$

$$= \frac{(\int_{D_t} O_{name}(t;N) \cdot O'_{face}(t;F)dt)^p}{\int_{D_t} O_{name}(t;N)dt \int_{D_t} O'_{face}(t;F)dt} \tag{18}$$

where p is an empirically determined constant ($p > 1$), and D_t is the time domain of the video. We see that when the peaks of $O_{name}(t;N)$ and $O'_{face}(t;F)$ overlap, the numerator will increase, but when they are offset, their product will be near zero and the numerator will be small (See Figure 8.). Suppose that $O'_{face}(t;F) = 1$ almost everywhere, say, white noise. This is the case of an anchor person's face which may coincide with almost any names, i.e., the face occurs anytime without any relation to occurrences of any names. There is no difference between a numerator with white noise face occurrence and an arbitrary $O_{name}(t;N)$, and a numerator with $O'_{face}(t;F)$ having 1 only where $O_{name}(t;N)$ does not equal to zero, i.e., the name occurs. An ideal co-occurrence should be small if $O'_{face}(t;F)$ has a large value where $O_{name}(t;N)$ is small. Thus we normalize the numerator by the size of $O_{name}(t;N)$ and $O_{face}(t;F)$. To understand why the constant p should be greater than 1, consider the case when $O'_{name}(t;N)$ has a value larger than zero at t_0 and $O_{face}(t;F) = k\delta(t-t_0)$ ($k > 0$). If $p = 1$, the value of k has no effect on $C(N,F)$, whereas ideally, as k becomes larger, the co-occurrence should also increase. To accomplish this, we choose a value for p greater than 1. We also note that its magnitude is constrained since a very large value for p will undo the normalization. In the experimental system described later, we used $p = 1.7$, although the system worked fine with $p = 1.5 \sim 2.0$.

6 Face-Name Association Method

The basic method for providing face-name association is simple. To associate a given face F to names, we calculate co-occurrence factor $C(N,F)$ for every name candidates N, sort the names by the factors, and give top-N names as the result. Association of a given name to faces can be given as well. Although this is simple, obtaining a co-occurrence factor requires significant computation, and a number of co-occurrence factors are needed to acquire an association result. It may thus require impractical computation time. Precomputation of the results for possible queries may overcome this drawback, particularly the name to face associations since a finite set of name candidates might be given for target videos. However, face-to-name association cannot be precomputed because a given face is not restricted to any finite set as it is unknown beforehand. To cope with this problem, we introduce a more efficient algorithm. First we give a conversion of the name occurrence function over time into the name occurrence over face space, and reveal that this is equivalent to $C(N, F)$. Then we describe an efficient algorithm to compute co-occurrence factors.

6.1 Conversion of the Name Occurrence Function

Assume f is a variable point within an eigenface subspace ($f \in \hat{F} = \mathrm{R}^{M'}$). Let $\hat{O}_{name}(f; N)$ be the name occurrence function over the eigenface subspace.

$$\hat{O}_{name}(f; N) = \sum_i [\delta(f - f(F_i)) \sum_j e^{-\frac{(t_{N,j} - t_{F_i})^2}{2\sigma_t^2}}] \tag{19}$$

where $\delta(f) = \delta(\|f\|^2)$, and $f(F)$ gives a corresponding point within the eigenface subspace of the face F. Then the name occurrence function over the eigenface subspace is dispersed using a Gaussian filter over the eigenface subspace;

$$\begin{aligned}
\hat{O}'_{name}(f; N) &= \hat{O}_{name}(f; N) \otimes e^{-\frac{\|f\|^2}{2\sigma_{face}^2}} &(20)\\
&= \sum_i [\delta(f - f(F_i)) \sum_j e^{-\frac{(t_{N,j} - t_{F_i})^2}{2\sigma_t^2}}] \otimes e^{-\frac{\|f\|^2}{2\sigma_{face}^2}} &(21)\\
&= \sum_i [e^{-\frac{\|f - f(F_i)\|^2}{2\sigma_{face}^2}} \sum_j e^{-\frac{(t_{N,j} - t_{F_i})^2}{2\sigma_t^2}}]. &(22)
\end{aligned}$$

Next we demonstrate an equivalence between the numerator of co-occurrence factor (18) and $\hat{O}'_{name}(f; N)$.

$$\begin{aligned}
&\int_{D_t} O'_{name}(t; N) \cdot O_{face}(t; F) dt &(23)\\
&= \int_{D_t} \sum_i e^{-\frac{(t - t_{N,i})^2}{2\sigma_t^2}} \cdot \sum_j e^{-\frac{\|f(F) - f(F_j)\|^2}{2\sigma_{face}^2}} \delta(t - t_{F_i}) dt. &(24)\\
&= \sum_{i,j} e^{-\frac{(t_{F_j} - t_{N,i})^2}{2\sigma_t^2}} e^{-\frac{\|f(F) - f(F_j)\|^2}{2\sigma_{face}^2}} &(25)\\
&= \hat{O}'_{name}(f(F); N). &(26)
\end{aligned}$$

Therefore the co-occurrence factor is given as

$$C(N, F) = \frac{(\hat{O}'_{name}(f(F); N))^p}{\int_{D_t} O_{name}(t; N) dt \int_{D_t} O'_{face}(t; F) dt}. \tag{27}$$

Figure 9 shows composition of name occurrence function over the eigenface subspace from face and name occurrence functions.

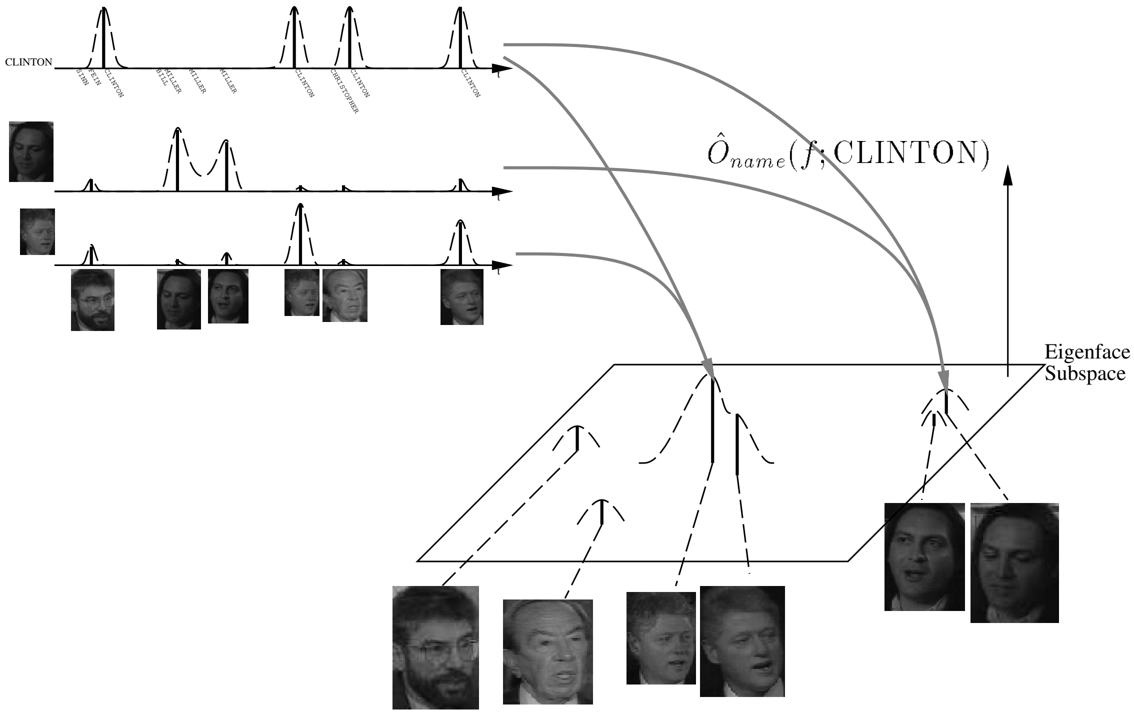


Figure 9: Conversion of Name Occurrence Function

6.2 Face-Name Association Algorithm

The name occurrence function over the eigenface subspace $\hat{O}'_{name}(f;N)$ can be expressed as follows;

$$A^i(f) = e^{-\frac{||f-f(F_i)||^2}{2\sigma^2_{face}}} \tag{28}$$

$$B^{i,j}_N = e^{-\frac{(t_{N,j}-t_{F_i})^2}{2\sigma^2_t}} \tag{29}$$

$$\hat{O}'_{name}(f;N) = \sum_i (A^i(f) \sum_j B^{i,j}_N). \tag{30}$$

Assume that a face F ($f = f(F)$) is given to provide associated names. It is noteworthy that $B^i_N = \sum_j B^{i,j}_N$ can be precomputed for each name candidate N in this situation, and this precomputation will greatly reduce the computation time. We precompute B^i_N for each name candidate N and each reference face F_i. We also prepare slots for B^i_N of the number of name candidates times the number of reference faces. In general the number of name candidates will be several hundred and the number of reference faces will be several thousand, therefore slots to store B^i_N will be on the order of a million and it is thus practical to store them in secondary storage.

To get associated names of a given face, $A^i(f)$ should be computed on demand. While this requires the number of reference faces iterations to compute $\hat{O}_{name}(f;N)$, most of the iterations are negligible due to Gaussian filtering, i.e.,

$$\hat{O}_{name}(f;N) = \sum_i A^i(f) B^i_N \tag{31}$$

$$\simeq \sum_{i \in \{i \mid \|f - f(F_i)\| < k\sigma_{face}\}} A^i(f) B_N^i \tag{32}$$

where k is a constant to control reduction of iterations. To select $A^i(f)$ satisfying $\|f - f(F_i)\| < k\sigma_{face}$, it is required to retrieve adjacent points from a given point in multidimensional space since f and $f(F_i)$ are M'-dimensional points. To achieve this retrieval, spatial data structure methods like R-tree [8] can be used. Roughly speaking, these methods compose tree structures from given M points or (hyper-)rectangles in multidimensional space, and provide spatial retrieval, i.e., they enumerate all data which overlap the given (hyper-)rectangle, with $O(\log M)$ computation.

To compute the co-occurrence factor $C(N, F)$, we need to obtain the numerator of Eq. (27), i.e.,

$$\int_{D_t} O_{name}(t; N) dt = num(N) \tag{33}$$

$$\int_{D_t} O'_{face}(t; F) dt = \sqrt{2\pi}\sigma_t \int_{D_t} O_{face}(t; F) dt \tag{34}$$

$$\int_{D_t} O_{face}(t; F) dt = \sum_i S(F, F_i) \tag{35}$$

$$= \sum_i e^{-\frac{\|f(F) - f(F_i)\|^2}{2\sigma_{face}^2}} \tag{36}$$

where $num(N)$ is the number of occurrences of the word N in the videos which can be precomputed. Note that the given face F is fixed with varying name candidates. In other words, Eq. (35) makes no contribution to the difference among co-occurrence factors of various name candidates. Therefore, this can be omitted when evaluating the co-occurrence factor for selecting associated names of the face.

Finally, to obtain the associated names of the face, we need to evaluate the co-occurrence factor $C(N, F)$ for each name candidate N. In total, this task requires to evaluate co-occurrences the number of candidate words (n_N) times, each of these evaluations requires the number of faces (n_F) iterations. The spatial data structure may drastically reduce the number of iterations with only $O(\log n_F)$ computation. This realizes a sufficiently fast computation, even for interactive systems.

Next, consider obtaining associated faces from a given name N. The basic idea is to select faces F of which co-occurrence factor $C(N, F)$ has large values. Resultant faces may be selected among reference face set $\{F_1, F_2, \cdots\}$. The system computes $C(N, F_i)$ for each reference face F_i, sorts co-occurrences, and selects F_i having the top-N values of co-occurrence. Obviously, co-occurrence Eq. (27) is undefined for a word which is not included in the set of name candidates obtained from videos. Also it is quite reasonable to restrict a given word to be one of the predefined name candidates. Since the number of name candidates is finite and it might not be so large (around several hundred), it is quite practical to precompute associated faces for each name candidate. In addition, the method for computing co-occurrences $C(N, F)$, explained above, can be applied to achieve efficient precomputation.

Who is he?

1. MILLER 30.4 2. NASHVILLE 11.17 3. DON 4.92 (BILL MILLER, singer)	1. AEROSMITH 3.19 2. GARTH 2.37 3. GUY 2.33 (GARTH BROOKS, singer)	1. FEIN 1.47 2. SINN 1.45 3. AUTHER 0.66 (GERRY ADAMS, leader of SINN FEIN)	1. ERHARDT 5.01 2. STEELERS 4.91 3. BEHRING 2.36 (RON ERHARDT, Pittsburgh STEELERS)

Who is "KELLY"?

22.52 21.50 18.83

(GENE KELLY)

```
KELLY MADE HIS FILM DEBUT
WITH JUDY GARLAND IN 1942.
BUT HE'S BEST KNOWN
FOR THE CLASSIC "SINGIN'
IN THE RAIN."
...
```

Figure 10: Face-Name Association Results

7 Experimental Results

We implemented the described algorithm on an SGI Indigo2 workstation (MIPS R4400 200MHz). Since it is an experimental embodiment, we have not yet implemented either search acceleration using spatial data structure in face-to-name association, or precomputation in name-to-face association. The system was applied to 9 CNN Headline News video (30 minutes each); in total 4.5 hours of news videos. From them 4,318 scene changes were detected and 320 faces extracted. The system is also given 251 candidate names which were extracted from all transcripts.

Figure 10 shows several results of face and name association. The upper row shows face-to-name associations. Each image was given to the system as an input face and a list of name candidates is output with corresponding co-occurrence factors. The top 3 name candidates are shown for each face image. The leftmost face is a singer, Bill Miller, and the system output "MILLER" as the first candidate. The next face is also a singer, Garth Brooks, which was its second candidate, "GARTH," while the first candidate "AEROSMITH" was said by reporter many times because the reporter introduced Brooks singing Aerosmith's song. The next face is a leader of Sinn Fein, Gerry Adams. Since "SINN", "FEIN", "GERRY", "ADAMS" were all marked as proper nouns and "SINN" and "FEIN" are frequently spoken, the face matched well with "SINN" and "FEIN." The rightmost face is the former offensive coordinator of the NFL's Pittsburgh Steelers, Ron Erhardt. The system output "ERHARDT" as the first candidate successfully as well as "STEELERS" as the second candidate.

The lower row shows the results of name-to-face association. The name "KELLY" was given to the system and top 3 candidates the system generated are shown with co-occurrence factors. Those faces are all Gene Kelly, a movie actor. Typically, it takes about 7 sec. to get face-to-name association results, and about 2 sec. to get name-to-face association.

These examples are successful results, mostly because they were introduced as special topics about these persons. In those topics, their names are said by anchors/reporters repeatedly, and their footage is also shown with their close-up shots. In such cases, the method extracted meaningful face and name associations. But there are some cases where the method is not applicable. A typical case is with the current American president, "Mr. Clinton" footage. Mr. Clinton is reported almost everyday, several times a day, however, at least in CNN Headline news, most reports are given by anchor persons, and only in their narration without any footage of Clinton. Since the method neither discriminates anchor persons nor recognizes the topics without footage of the person of interest, it thus far cannot associate Clinton's face and name. The method will give less co-occurrence factor to anchors which coincide with many names as described in Section 5.3, and thus the method does not need to give special treatment to anchors. However, it is necessary to discriminate anchors to cope with this "Mr. Clinton" case. In addition, much higher natural language processing techniques and knowledge of news video structure may be needed.

8 Conclusions

This paper describes a face and name association method in videos. The method is provided with news videos showing persons of interest, then given a name and returns associated faces, or given a face and returns associated names. The successful results demonstrate that face and name association provides useful video content information, thus Name-It contributes to automated video indexing. In addition, the successful results of Name-It reveal that our methodology of integrated use of image understanding techniques and natural language processing techniques is headed in the right direction to achieve our goal of accessing real contents of multimedia information. Because the method is still not fully robust, i.e., there are some cases the method cannot analyze, we are improving this method by using higher level natural language processing, i.e., applying a thesaurus and parsing, as well as much higher image processing, i.e., face tracking and extracting face occurrence duration, and enhancement of face identification method. This research was done with support from the Informedia digital video library project, and more than 100 hours of news videos are available. We would like to apply this method to this archive after the method acquires more robustness. This might explain a practical effectiveness of the method, and as an outcome of this, we can obtain a faces and names association database containing more than several hundred of people.

References

[1] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huand, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker, "Query by image and video content: The QBIC system," *IEEE Computer*, pp. 23–32, September 1995.

[2] A. Pentland, R. W. Picard, and S. Schlaroff, "Photobook: Content-based manipulation of image databases," *International Journal of Computer Vision*, vol. 18, no. 3, pp. 233–254, 1996.

[3] H. Zhang, C. Low, and S. Smoliar, "Video parsing and indexing of compressed data," *Multimedia Tools and Applications*, vol. 1, pp. 89–111, 1995.

[4] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.

[5] R. Chopra and R. K. Srihari, "Control structures for incorporating picture — specific context in image interpretation," in *Proceedings of IJCAI '95*, 1995.

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

[7] H. Rowley, S. Baluja, and T. Kanade, "Human face detection in visual scenes," Tech. Rep. CMU-CS-95-158, School of Computer Science, Carnegie Mellon University, 1995.

[8] A. Guttman, "R-TREES: A dynamic index structure for spatial searching," in *SIGMOD*, pp. 47–57, 1984.