Proceedings of CVPR'97, pp. 368–373, 1997. Copyrighted by the IEEE.

Name-It: Association of Face and Name in Video

Shin'ichi Satoh * Takeo Kanade
School of Computer Science
Carnegie Mellon University
5000 Forbes Ave., Pittsburgh, PA 15213, USA
shinichi.sato@cs.cmu.edu takeo.kanade@cs.cmu.edu

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.

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 [2], feature extraction using Karhunen-Loéve (K-L) transform [4], and video structuring based on scene change detection [8]. 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. Piction system [1] identifies faces within given captioned photos, typically of newspapers. However, Piction deeply depends on text information, in other words, natural language processing, and uses only primitive image processing techniques.

Figure 1. Typical Composition of News Video

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 in news videos. A person can easily say which person is shown in a video. However, this task contains several difficult procedures which 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. Potential applications of Name-It include (1) creation of face-name association database, (2) video annotation by persons' name, etc.

Input videos are composed of transcripts as text information and image sequences. Extraction of face images from image sequences, as well as extraction of name candidates from transcripts are explained in Section 3. 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.

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

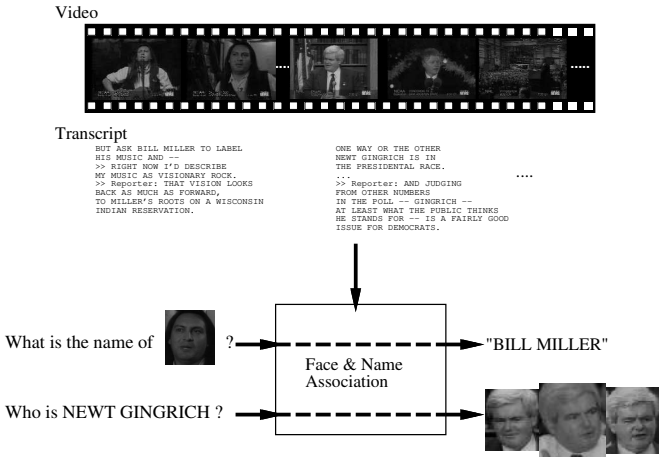

Figure 2. Face and Name Association

2 Overview of Name-It

Figure 1 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 2). 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. Typical face and name association will be carried out by human within news video footage (Figure 1) 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.

Although human seem to be able to do this easily, this process includes several complicated, and high level processes: (1) Scene/topic detection, (2) Face detection, (3) Face identification, and (4) Name extraction. It is very hard for computers to achieve these processes automatically. 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 [5], eigenvector method based face identification [7], 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.

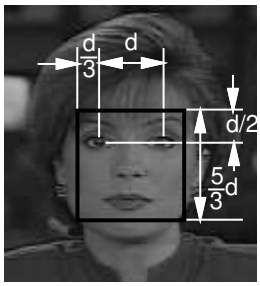

Figure 3. Face normalization based on eye location

3 Preparations

3.1 Face Extraction

First we have to extract faces from a given video footage. We use a neural-network based face detector to locate faces in images [5]. 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. 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 (as shown in Figure 3; black square is a face region). Then face regions are normalized into 64×64 images for face similarity evaluation.

3.2 Face Similarity

We used an eigenvector based method to compute the distance between two faces. This is based on the well known eigenface method [7]. Using this distance, we defined a similarity between two faces.

The eigenface method provides a mapping from face images to multi-dimensional points. First the method is obtained a certain number of fixed training set of faces. Given

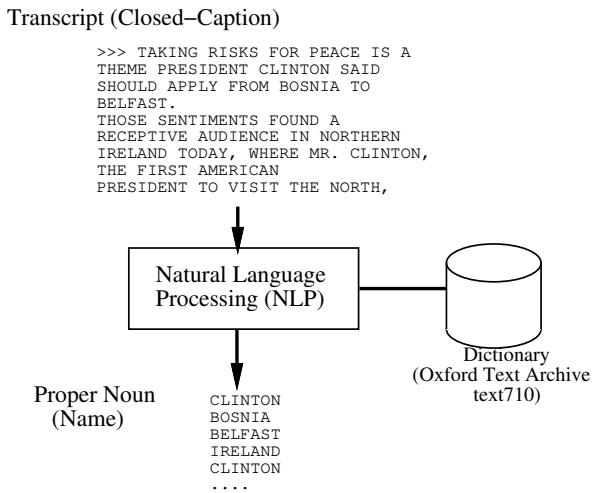

Figure 4. Name Candidates Extraction

M training faces, the method can provide a mapping Φ from faces to $M'(< M)$-dimensional points. Assume that two faces F_i, F_j are given, and corresponding two points ϕ_i, ϕ_j are obtained by Φ. The Euclidean distance between these points is expected to reflect similarity of corresponding faces. Then the similarity between them $S(F_i, F_j)$ is defined as follows;

$$S(F_i, F_j; \sigma_f) \quad = \quad e^{-\frac{\|\Phi(F_i)-\Phi(F_j)\|^2}{2\sigma_f^2}}$$

where σ_f is a standard deviation of Gaussian filter in the eigenface subspace, which will be empirically determined to make the similarity small for identical faces, as well as to make it large for not identical faces. We also have to determine the dimension of the eigenface subspace M', and we used $M' = 16$ for our experimental Name-It embodiment.

3.3 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 4 depicts the diagram of name candidate extraction, an example of a transcript, and extracted proper nouns.

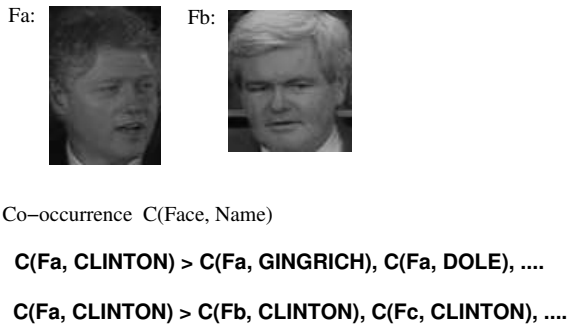

Figure 5. An Example of Face and Name Co-occurrence

4 Face and Name Co-occurrence

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. (See Figure 5.)

4.1 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. The name occurrence function is defined as

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

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 3.2, 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

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

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

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

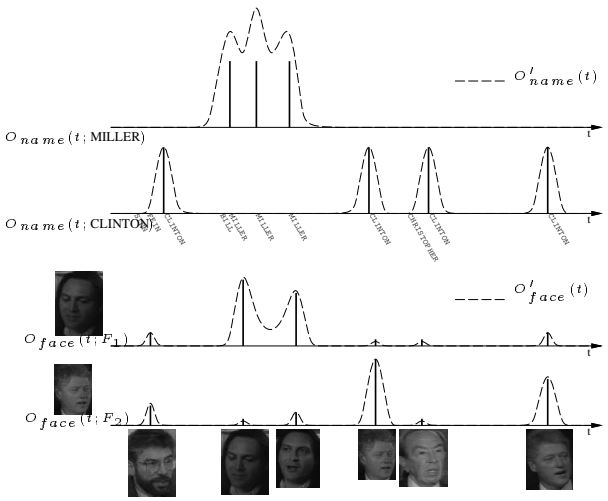

Figure 6. Occurrence Function

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

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. The same thing can be said of $O'_{face}(t;F)$. Figure 6 illustrates the relationship between occurrence of names and faces, and occurrence functions.

4.2 Co-occurrence

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

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

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 6.). 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 larger value 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 $p = 1.5 \sim 2.0$.

5 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. To cope with this problem, we introduce a more efficient algorithm.

5.1 Conversion of the Name Occurrence Function

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

$$\begin{aligned} & \hat{O}_{name}(\phi;N) \\ &= \sum_i [\delta(\phi - \Phi(F_i)) \sum_j e^{-\frac{(t_{N,j}-t_{F_i})^2}{2\sigma_t^2}}] \end{aligned}$$

where $\delta(\phi) = \delta(\|\phi\|^2)$. Then the name occurrence function over the eigenface subspace is dispersed using a Gaussian filter over the eigenface subspace;

$$\begin{aligned} & \hat{O}'_{name}(\phi;N) \\ &= \hat{O}_{name}(\phi;N) \otimes e^{-\frac{\|\phi\|^2}{2\sigma_{face}^2}} \\ &= \sum_i [e^{-\frac{\|\phi-\Phi(F_i)\|^2}{2\sigma_{face}^2}} \sum_j e^{-\frac{(t_{N,j}-t_{F_i})^2}{2\sigma_t^2}}] \\ &= \int_{D_t} O'_{name}(t;N) \cdot O_{face}(t;\Phi^{-1}(\phi))dt. \end{aligned}$$

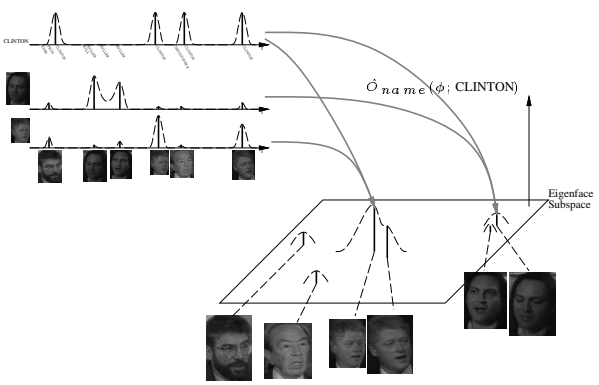

Figure 7. Conversion of Name Occurrence Function

Therefore the co-occurrence factor is given as

$$C(N,F) = \frac{(\hat{O}'_{name}(\Phi(F);N))^p}{\int_{D_t} O_{name}(t;N)dt \int_{D_t} O'_{face}(t;F)dt}. \quad (1)$$

Figure 7 shows composition of name occurrence function over the eigenface subspace from face and name occurrence functions.

5.2 Face-Name Association Algorithm

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

$$\begin{aligned} A^i(\phi) &= e^{-\frac{\|\phi-\Phi(F_i)\|^2}{2\sigma_{face}^2}} \\ B_N^{i,j} &= e^{-\frac{(t_{N,j}-t_{F_i})^2}{2\sigma_t^2}} \\ \hat{O}'_{name}(\phi;N) &= \sum_i (A^i(\phi) \sum_j B_N^{i,j}). \end{aligned}$$

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

While it requires the number of reference faces iterations to compute $\hat{O}_{name}(\phi;N)$, most of the iterations are negligible due to Gaussian filtering, i.e.,

$$\begin{aligned} &\hat{O}_{name}(\phi;N) \\ &= \sum_i A^i(\phi) B_N^i \\ &\simeq \sum_{i \in \{i \mid \|\phi-\Phi(F_i)\| < k\sigma_{face}\}} A^i(\phi) B_N^i \end{aligned}$$

where k is a constant to control reduction of iterations. To select i satisfying $\|\phi - \Phi(F_i)\| < k\sigma_{face}$, it is required to retrieve adjacent points from a given point in multidimensional space since ϕ and $\Phi(F_i)$ are M'-dimensional points. To achieve this retrieval, spatial data structure methods like SR-tree [3] 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. Acceleration of computing the denominator of Eq. (1) can also be achieved as well by using precomputation and the spatial data structure.

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. Resultant faces may be selected among reference face set. Obviously, co-occurrence Eq. (1) 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.

6 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 8. Face-Name Association Results

Figure 8 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. These 4 example queries achieved successful results as shown. 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 in news videos 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 4.2, 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.

7 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.

Acknowledgement

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

[1] R. Chopra and R. K. Srihari. Control structures for incorporating picture-specific context in image interpretation. In *Proc. of International Joint Conference on Artificial Intelligence*, 1995.

[2] 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*, pages 23–32, September 1995.

[3] N. Katayama and S. Satoh. The SR-tree: An index structure for high-dimensional nearest neighbor queries. In *Proc. ACM SIGMOD*, pages 369–380, 1997.

[4] A. Pentland, R. W. Picard, and S. Schlaroff. Photobook: Content-based manipulation of image databases. *International Journal of Computer Vision*, 18(3):233–254, 1996.

[5] 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.

[6] 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.

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

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