

Person Identification Using Text and Image Data

David S. Bolme, J. Ross Beveridge and Adele E. Howe
Computer Science Department
Colorado State University
Fort Collins, Colorado 80523

[bolme,ross,howe]@cs.colostate.edu

Abstract—This paper presents a bimodal identification system using text based term vectors and EBGM face recognition. Identification was tested on a database of 118 celebrities downloaded from the internet. The dataset contained multiple images and two biographies for each person. Text based identification had a 100% identification rate for the full biographies. When the text data was artificially restricted to six sentences per subject, rank one identification rates were similar to face recognition (approx. 22%). In this restricted case, combining text identification and face identification showed a significant improvement in the identification rate over either method alone.

I. INTRODUCTION

Single mode (i.e., image) person identification is highly unusual in ordinary human interaction. Context (e.g., location of meeting, something said) may provide constraints that can supplement the imagery of a person's face. For example, seeing a casual acquaintance at a conference substantially narrows the possible names you might attach to the face. The topic of conversation may also help in identification by providing more information about the person, such as the person's occupation, interests, or institution.

One could imagine a portable device like a PDA or cell phone that uses face recognition, speech recognition, and the text from a conversation to identify a person. By exploiting identifying information gathered by these different sources, it could be possible to automatically record attendance and notes at meetings, index digital content by person, or automatically retrieve a person's identity and other germane information.

Another application is multimedia retrieval. Image retrieval is typically concerned with searching databases of images based on content. Often text or keywords in the surrounding document are exploited to find images of interest [7][18]. Combining face and text identification technologies could improve searches for a particular person. Broadcast video contains image data that can be used for face identification and textual information derived from spoken or onscreen text or captions [14].

This paper describes an investigation into identification of a person using text data that is found in a biography and visual information that is found in a photograph. We show that text information can be used to supplement face identification to produce better recognition rates for difficult identification problems.

The system constructed in this study uses well known techniques for each component: Elastic Bunch Graph Matching (EBGM) [16] for face identification and word frequency vectors for text identification. The similarity measures for both modes are combined using a simple weighted sum.

A bimodal identification system may perform better simply because there is more information available for identification. We show that doubling the information for single mode identification algorithms also dramatically increases the recognition rate. By controlling for the amount of information, we show that bimodal identification still improves the recognition rate over single mode identification. Tests are conducted on a database that was assembled from images and biographies of movie stars.

As far as we know this is the first work that supplements a face recognition algorithm with free form text for person identification. We find that a relatively simple Information Retrieval (IR) method is effective for person identification using text data.

II. RELATED WORK

Both text based information retrieval and face identification are well studied areas. Work has been done in multimedia retrieval that combines visual features from images with keywords and other textual information in surrounding text. Multimodal biometric systems often combine face recognition with other sources of data such as fingerprints, voice, or ears but not unconstrained text. The remainder of this section discusses work that uses combinations of classifiers to solve problems that are similar to those discussed in the introduction.

Several systems have been developed that use caption data to identify people in photographs. Srihari [15] created a system that used spatial cues in caption text to automatically label faces in an image. This system, called Picion, would find the names in the caption and use cues, such as "Person A is to the right of Person B" in the caption to determine which name belonged with each face. Berg *et al.* [1] developed a system that used clustering to associate names in captions with the faces in the photos. In that system each face was labeled with every name in the associated caption. Therefore many names were associated with each face. Clustering based on face similarity was used to determine which of the candidate names belonged with each face.

Hazen *et al.* [8] used devices (PDAs) to validate a person's identity for a login scenario. They found that the microphones and cameras available for the handheld devices were of much lower quality than most sensors used in standard biometric identification systems. In addition, the environment played a key role in the data because of uncontrolled lighting and background noise. As a result verification was very difficult on the PDA collected data. The system used Hidden Markov Models for voice authentication and Support Vector Machines for face authentication. The two similarity scores were combined using a weighted sum. Their work showed that multi-modal identification using both voice and face data reduced the equal error rate by 50%.

Poh *et al.* [11] constructed a similar system that used multi-layer perceptrons for both face and speech authentication. The perceptrons had a binary output indicating if the inputs were from the same person. Verification of a person is successful only if both the face and speech both passed verification individually.

Yang *et al.* [17] developed a multimodal system for recording attendance at meetings. The system used face identification, speaker identification, color appearance identification, person tracking, and sound source identification to identify meeting attendees and automatically assign meeting transcripts to a particular speaker.

III. COMBINED TEXT AND FACE IDENTIFICATION

The identification problem explored in this work is to assign an identity to an unknown person. The algorithm is given an unknown person (probe) and is expected to select the most similar person from a collection (gallery) of known people. Each probe and gallery person has face imagery and text data associated with him or her. The similarity of the probe person to a person in the gallery depends on the similarity of this associated information. The unknown person is then labeled with the identity of the most similar person from the gallery.

The algorithms used in this study were selected because they were well known, effective, and easy to implement. The term vector similarity computation is simple and well established in the information retrieval community. An open source implementation of the EBGM algorithm[4] was used to reduce development time. The weighted sum of similarity scores was used because it was both simple and performed well in other systems[12], [8], [9].

The rest of this section explains how similarity is computed between the text and face data representing a person in the probe set and the text and face data representing a person from the gallery. Section III-A discusses a similarity measure based on the words used in a text. Section III-B discusses face similarity based on EBGM. Section III-C discuss how the two similarity scores were combined using a weighted sum to create a bimodal similarity score.

A. Text Identification

Many modern information retrieval systems are based on a bag-of-words frequency vector model developed by Salton

and Buckley[13]. This method of text retrieval is based on a very simple concept that the number of times a word is used in a text is directly related to the importance of that word.

Each text is represented by a vector such that each element in that vector corresponds to a unique word (called **term** in the IR literature) that is used in that text. A text that has n unique words would have a word frequency vector:

$$\vec{w} = (w_1, w_2, \dots, w_n) \quad (1)$$

where w_t is a weight based on the importance of the term t . Bodies of text that discuss the same topics should have similar word frequency vectors.

To compute similarity between two term vectors Salton and Buckley suggest using the cosine of the angle between the two vectors. The biography similarity measure (S_b) is shown below for term vectors for a probe (\vec{p}) and for a person in the gallery (\vec{g}):

$$S_b(\vec{p}, \vec{g}) = \frac{\sum_t p_t g_t}{\sqrt{\sum_t p_t^2} \sqrt{\sum_t g_t^2}} \quad (2)$$

where t is the union of all terms in \vec{p} and \vec{g} .

A good indicator of term importance is the number of times it is used in the text. The weight is therefore proportional to the term frequency (tf) which is the count of the number of times the term is used in the body of text. If the same term is used in many of the texts, that term is not particularly useful for discriminating between texts. The weight of that term should be reduced by the inverse document frequency (idf), which is based on all of the uses of the term in the corpus. The weight is defined as:

$$w_t = tf_t \cdot idf_t \quad (3)$$

This weighting method is called TFIDF in the IR literature.

Salton and Buckley did an exhaustive study of many different definitions for tf and idf . For this work, we have selected tf_t as the number of times term t is used in the text. The idf_t for each term is estimated using all of the texts:

$$idf_t = \log \frac{N}{n_t} \quad (4)$$

where n_t is the number of texts that contain term t and N is the total number of texts in the corpus.

Not all words in the corpus are useful. If a word is commonly used in many texts it reduces the accuracy of the similarity computation. Such words are called *stop words* and are not included in the word frequency vectors. Typically stop words lists contain 15 to 500 words and are usually words like "AND", "IN", "THAT", "THE", "WHEN", and "WILL". We defined a stop word if it was associated with 50 of the 118 celebrities in the dataset. The names of the celebrities were also removed because the names could make the text trivial to classify. There were 423 stop words and 205 names removed from the dataset, which included words common to celebrity biographies like "ACTING", "DIRECTED", "GLOBE", "MOTHER", "SPOUSES", "THRILLER", and "TV".

B. Face Identification

An EBGM algorithm was chosen for face recognition because it performs relatively well on face datasets with variation in pose and lighting.¹ This research used the open source EBGM algorithm implementation (version 5.1) from Colorado State University’s Evaluation of Face Recognition Algorithms Project²[4], [5], [16]. The algorithm was trained on imagery from the FERET database[10] using a standard configuration that ships with the CSU system. More details on this process can be found in [2].

The EBGM algorithm computes the similarity between two face images. The first step is to geometrically register the image. The normalization process centers the face in a 128 by 128 pixel image tile, and rotates, scales, and translates the face to align the eyes. In these experiments eye finding was performed by hand. The eyes are spaced 24 pixels apart. Histogram equalization and value normalization are also performed on this image tile.

The second step in the algorithm is to locate the 25 interest points shown in Figure 1. Gabor jets are extracted from locations on the face near these landmarks. The information in these jets is compared to similar jets in the training imagery. The phase information is used to refine the estimated location of the landmark.

Once the landmarks have been located, a second set of Gabor jets representing these landmarks are extracted from the image. These jets are used for similarity comparison to other faces. The final similarity measure for the face (S_f) is the average similarity of all 25 jets compared using a correlation like similarity measure.

C. Similarity Combination

The combined similarity score is a weighted sum of the face and biography similarity score:

$$S_c = S_f + 2.0 * S_b \quad (5)$$

where S_c is the combined similarity, S_f is the face similarity, and S_b is the biography similarity.

A series of different weights were tried from 1.0 to 9.0 to determine the optimal weighting between face and text (see Figure 2). The value of 2.0 was chosen for the weight because it was at the peak of the resulting curve and performed slightly better than 1.0 and 3.0.

IV. RESULTS

This section describes the dataset and evaluation performed on the bimodal identification system. The dataset is based on photos and biographies of popular celebrities. A permutation evaluation[3] was used to compare the performance of the combined algorithm to face only and text only algorithms. The evaluations showed that the bimodal algorithm did perform better than the single mode algorithms.

¹Principal Components Analysis was also tried with inferior results.

²<http://www.cs.colostate.edu/evalfacerec>

Recognition Rate for Combination Weights

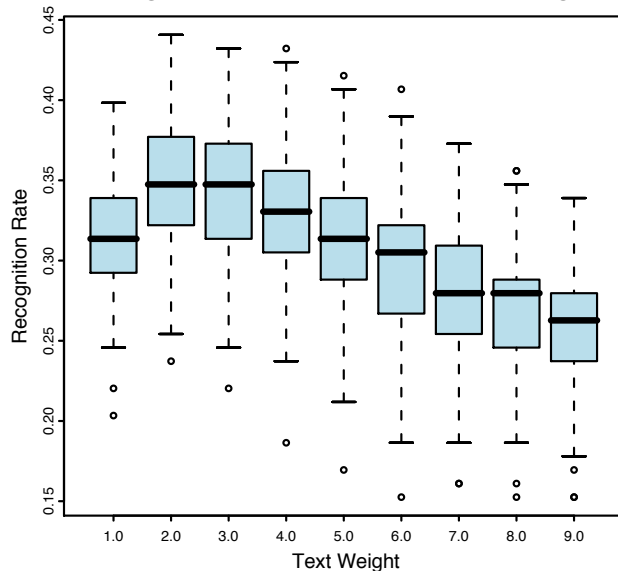


Fig. 2. This figure shows the results of the experiment used to determine the weighting between face and text similarity. The operating point chosen for the experiment is a text weighting of 2.0.

A. Dataset

We required a realistic dataset with both face imagery and associated text information. We assembled a dataset of 118 “popular” celebrities from two online sources: Yahoo Movies³ and Wikipedia⁴. Celebrities were chosen because of the abundance of images and text information that could be found on the Internet.

The dataset was downloaded from each site by automatically searching for the celebrity’s name and then downloading associated text and images. Imagery was only downloaded from Yahoo. Each celebrity had a photo webpage that contained images from the celebrity’s recent films or events such as the Academy Awards. The webpage source for each of the biographies was downloaded and converted to plain text. Text associated with the webpage template was also removed automatically using simple pattern matching leaving only the text associated with the biographical content.

The images used in these experiments were all low resolution (approximately 300 by 400) pixels and were downloaded in a compressed JPEG format. Each photo was inspected and the person’s face was identified and geometrically registered by manually selecting the eye coordinates. Photos that were taken from movies were excluded due to extreme lighting, expressions, costumes, and makeup. Only the last 20 images included on the webpage were considered because those contained most of the event photos. The photos were also separated by events to provide a way to partition the imagery into probe and gallery sets that were collected on different days (see Section IV-B).

³<http://movies.yahoo.com>

⁴<http://www.wikipedia.org>

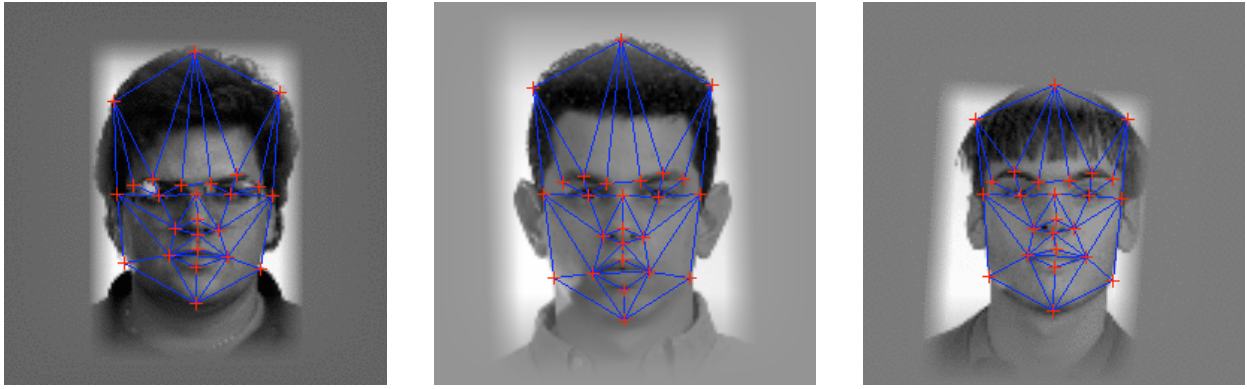


Fig. 1. Landmarks used for the EBGM algorithm.

The biographies typically covered the person’s childhood, personal life, and career. The Yahoo biography’s main focus was on the celebrity’s rise to fame. This typically included a summary of the major roles that influenced the actor’s career. The Wikipedia biographies had a much better balance between career and personal life. Most biographies contained 500 to 3000 words.

The permutation test discussed in the next section requires multiple images and plenty of text for each person in the database. To avoid a person with too little data, the celebrities were chosen that met minimum requirements for number of images, image resolution, and an amount of text. Sentences that had less than 10 words or more than 100 words were removed from the dataset to avoid sentences with too many words or too few words. Some biographies contained filmographies containing many hundreds of words that were parsed as one sentence. Some sentences also had very few words such as section headings. The final composition of the dataset is summarized in Table I.

TABLE I
SUMMARY OF THE DATASET USED IN THE EXPERIMENTS.

Summary	Celebrities	118
	Men	93
	Women	25
	Total Images	1331
Image	Events per Celebrity	2 or more
	Images per Celebrity	4 or more
	Distance Between Eyes	16 pixels
Biography	Words Per Celebrity	400 or more
	Words Per Sentence	10 to 100

B. Evaluation

A non-parametric probe/gallery permutation analysis[3] was used to evaluate the identification algorithms. For each trial, the probe and gallery data were randomly selected from a pool of imagery and biography data for each person. One hundred trials were run to produce average recognition rates and confidence intervals for the algorithms. The dependent variable evaluated in this study is the rank 1 identification

rate, which is the fraction of probes in which the algorithm selects the correct person in the dataset.

Before running experiments on the combined classifier, a series of experiments were done to determine a good weighting between the face and text classifiers. Because of the limited amount of data available for this study, these initial experiments were run on the full dataset using the same permutation analysis method. Text classification using full biographies produced a 100% recognition rate. To make the problem more difficult we reduced the length of the text by randomly selecting sentences from the full biography. Figure 3 compares the recognition rates using only text data of various lengths. Three sentences were chosen for the size of the text data (OneBio) because the recognition rate was very similar to the EBGM algorithm (OneFace).

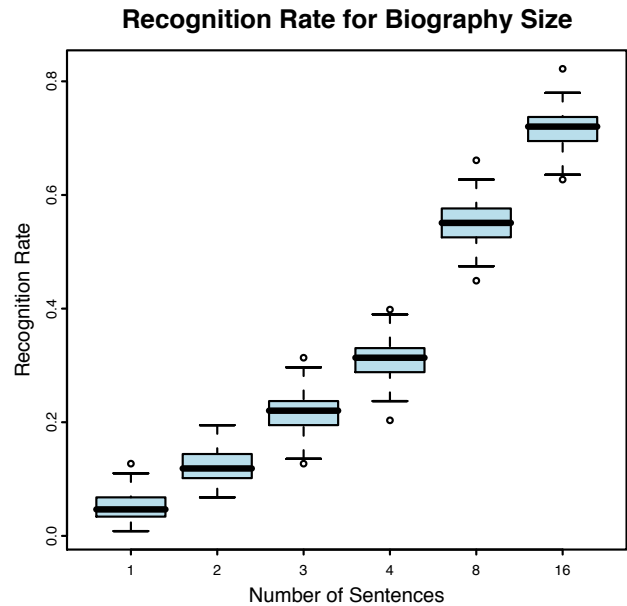


Fig. 3. Adding more text to the biographies causes the performance to go up. The operating point chosen for the experiment is three sentences which is the basis for the OneBio algorithm.

Five specific algorithm variations are studied in this paper.

The first two algorithms are the standard single mode algorithms for face (OneFace) and text (OneBio). The OneFace algorithm is EBGM alone. The OneBio algorithm matches based on three randomly selected sentences of text using the vector model. The third algorithm combines the similarity scores of the OneFace and OneBio algorithms to produce the bimodal algorithm. The comparison is shown in Figure 4.

We expected the Combined algorithm to perform better simply because it uses more non-redundant information. To test if combining the two modes had a significant effect over just adding more information, the combined algorithm was compared to face only and text only algorithms that identify using twice the information. The Combined algorithm was compared to a face only algorithm that matches using two faces (TwoFace), and a text only algorithm that matches using two sets of three sentences (TwoBio). The similarity scores for the single mode algorithms are averaged because there is no reason to prefer one over the other.

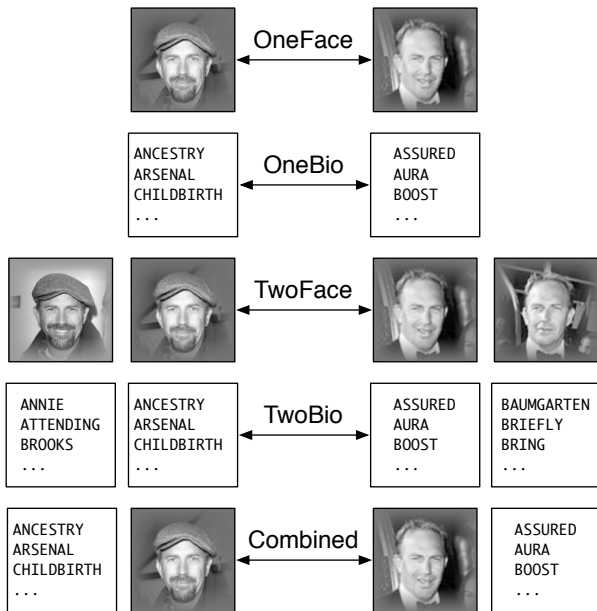


Fig. 4. The five classifiers tested in this evaluation are illustrated above.

Four sets of biography data were randomly sampled to obtain the probe and gallery sets for each person. Each biography set contained three sentences from the person’s biography. Therefore six sentences were used for the probe and another six sentences were used as the gallery. To make sure that the probe and gallery came from different sources, the probe sentences were selected from the Wikipedia dataset and gallery sentences were selected from the Yahoo dataset.

For each person, four face images were randomly sampled from a pool and randomly assigned to two slots in the probe and two slots in the gallery. Many of the images were taken at the same event and therefore had very little variation. To reduce the possibility of getting these easy image pairs in both the probe and gallery the images were resampled until the probe and gallery were selected from different events.

Recognition Rate for Combined Classifier

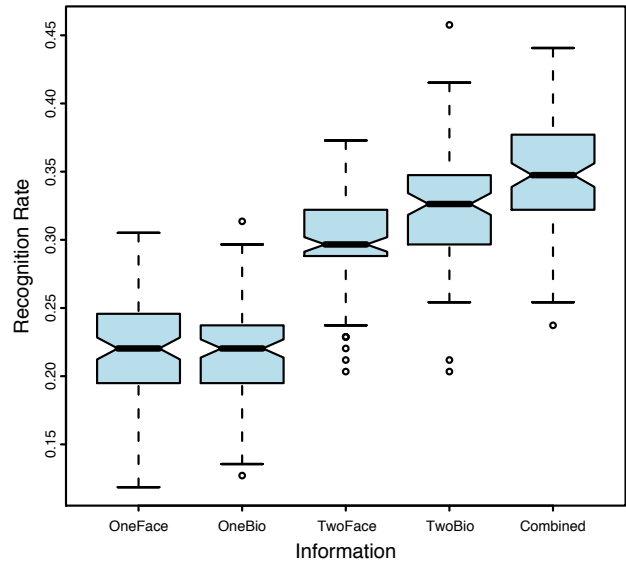


Fig. 5. This figure shows that the combined classifier performs better than text only and face only classifiers even if you double the information used by the single mode classifiers.

Figure 5 shows that the Combined algorithm performs significantly better than either the OneFace or OneBio algorithms. The Combined algorithm also performed better than the TwoFace and TwoBio algorithms, however the difference in the recognition rates are much closer.

A paired t-test (Table II) was used to compare the Combined algorithm to the TwoFace and TwoBio algorithms. Diff. is the difference in the mean recognition rate. The p-values indicate statistical significance. The perm column indicates the number of permutations where the combined algorithm performed better. The 99% interval shows that there is a statistically significant improvement for the combined classifier which means that using face and biography for identification does have an effect over doubling the information for the single mode classifiers.

TABLE II

THIS TABLE SHOWS THE RESULTS OF THE PAIRED T-TEST FOR 100 SAMPLES COMPARING THE TWOBIO AND TWOFACE TO THE COMBINED ALGORITHM.

Alg.	Diff.	p-val	99% Int.	Perm.
TwoBio	0.022	5.6e-07	(0.012,0.034)	64/100
TwoFace	0.046	2.2e-16	(0.034,0.058)	82/100

V. FUTURE WORK

One opportunity for future work could be to expand the types of applications for face and text matching. The identification problems could be reengineered as an IR problem for associating person identities with documents on a file system. Unsupervised clustering could be used to automatically find unique identities in collections of documents and images or

keyword extraction could be a way to automatically associate names with images.

Also, the biography data collected for the dataset may not be representative of typical data that is encountered in the problem domains described above. It was chosen because it was easily collected. The biographies follow a prescribed format that contains a lot of identifying information. For the evaluation, the biography data was sub-sampled to make the identification problem more difficult, but it is possible that these data are still not typical of more general real world problems. A more realistic, but more time consuming alternative might be to download bios and photos from faculty websites.

The biographies used in this study are very structured and contain a variety of information over the subject's whole life. In the absence of a well structured biography, certain types of words or phrases could be important in other sources of text data. An algorithm could automatically learn these phrases and associate them with the person. News articles, for example, would typically contain the person's name and occupation. A person could also be associated with locations, types of events, or other people.

The classification methods used in this work are not state of the art. EBGM is a good but dated face recognition algorithm. TFIDF for computing text similarity is also a very simple implementation that lacks features included in more robust information retrieval systems. A simple linear weighting is used to combine the similarity scores. More complex methods for combining the similarity scores may improve performance.

The text similarity measure has some known problems that have been addressed by the IR community. A simple way to improve the robustness and accuracy of the similarity measure is to use word stemming. For example, the words "dog", "dogs", and "doggy" would all be mapped to the stem "dog" before forming term vectors. Latent Semantic Indexing[6] can also reduce the problems due to words that have multiple meanings (polysemy) and multiple words that have the same meaning (synonymy).

Finally, an investigation into different combination techniques could provide a better overall classification rate. The weighted sum of similarities was a straight forward method but it is not a particularly smart way of doing combination.

VI. CONCLUSIONS

This paper is a proof of concept for combining free text and traditional biometrics for person identification. It was found that the simple term vector similarity measure was able to produce 100% identification using full biographies. When limited text data (three sentences) is available, the text based algorithm performed approximately as well as EBGM face recognition. Combining text and face using the weighted sum of similarity scores outperformed both methods alone, even with the addition of information to the single mode identifiers.

REFERENCES

- [1] T. L. Berg, A. C. Berg, J. Edwards, M. Maire, R. White, Y.-W. Teh, E. Learned-Miller, and D. A. Forsyth. Names and faces in the news. *cvpr*, 02:848–854, 2004.
- [2] J. R. Beveridge, D. Bolme, M. Teixeira, and B. Draper. The CSU face identification evaluation system user's guide: Version 5.0. Computer Science Department Colorado State University, May 2003.
- [3] J. R. Beveridge, K. She, B. Draper, and G. H. Givens. A nonparametric statistical comparison of principal component and linear discriminant subspaces for face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 535 – 542, December 2001.
- [4] D. S. Bolme. Elastic bunch graph matching. Master's thesis, Colorado State University, May 2003.
- [5] D. S. Bolme, J. R. Beveridge, M. L. Teixeira, and B. A. Draper. The CSU face identification evaluation system: Its purpose, features and structure. In *Proc. 3rd International Conf. on Computer Vision Systems*, Graz, Austria, Apr. 2003.
- [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*, 41(6):391–407, 1990.
- [7] W. I. Grosky. Multimedia information systems. *IEEE MultiMedia*, 1(1):12–24, 1994.
- [8] T. J. Hazen, E. Weinstein, and A. Park. Towards robust person recognition on handheld devices using face and speaker identification technologies. In *ICMI '03: Proceedings of the 5th international conference on Multimodal interfaces*, pages 289–292, New York, NY, USA, 2003. ACM Press.
- [9] J. Kittler, M. Hatef, R. P. Duin, and J. Matas. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239, 1998.
- [10] P. Phillips, H. Moon, S. Rizvi, and P. Rauss. The FERET Evaluation Methodology for Face-Recognition Algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(10):1090–1104, October 2000.
- [11] N. Poh and J. J. Korczak. Hybrid biometric person authentication using face and voice features. In *AVBPA '01: Proceedings of the Third International Conference on Audio- and Video-Based Biometric Person Authentication*, pages 348–353, London, UK, 2001. Springer-Verlag.
- [12] F. Roli, J. Kittler, G. Fumera, and D. Muntoni. An experimental comparison of classifier fusion rules for multimodal personal identity verification systems. In *MCS '02: Proceedings of the Third International Workshop on Multiple Classifier Systems*, pages 325–336, London, UK, 2002. Springer-Verlag.
- [13] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513–523, 1988.
- [14] T. Sato, T. Kanade, E. K. Hughes, M. A. Smith, and S. Satoh. Video ocr: indexing digital new libraries by recognition of superimposed captions. *Multimedia Syst.*, 7(5):385–395, 1999.
- [15] R. K. Srihari. Automatic indexing and content-based retrieval of captioned images. *Computer*, 28(9):49–56, 1995.
- [16] L. Wiskott, J.-M. Fellous, N. Kruger, and C. von der Malsburg. Face recognition by elastic bunch graph matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):775–779, July 1997.
- [17] J. Yang, X. Zhu, R. Gross, J. Kominek, Y. Pan, and A. Waibel. Multimodal people id for a multimedia meeting browser. In *MULTIMEDIA '99: Proceedings of the seventh ACM international conference on Multimedia (Part 1)*, pages 159–168, New York, NY, USA, 1999. ACM Press.
- [18] X. S. Zhou and T. S. Huang. Unifying keywords and visual contents in image retrieval. *IEEE MultiMedia*, 9(2):23–33, 2002.