# RECOGNIZING FACES IN RECORDED MEETINGS VIA MRC-BOOSTING

*Xun Xu[†] , Yong Rui[‡]  and Thomas S. Huang[†]*

† Beckman Institute
University of Illinois at Urbana-Champaign
Urbana, IL 61801
*{xunxu, huang}@ifp.uiuc.edu*

‡ Microsoft Research
One Microsoft Way
Redmond, WA 98052
*yongrui@microsoft.com*

## ABSTRACT

Person-based indices and timelines can enable fast and non-linear access to recorded meetings. This paper focuses on how to automatically construct those indices and timelines by using face recognition techniques. While there exist extensive research in generic face recognition, recognizing faces in recorded meetings is still an understudied area. Real-world meeting videos impose several interesting and unique challenges including complex lighting, low imaging quality, and large variations in head pose and size. In this paper, a promising approach based on MRC-Boosting is presented to address these challenges, which achieves encouraging performance on real-world meeting videos and shows superior accuracy and robustness compared to two popular existing approaches.

## 1. INTRODUCTION

Meetings are one of the most common activities in business. However, because of people's busy schedule, it can be difficult for all the team members to find a common available time to meet. Whenever live (synchronous) meetings are not possible, recorded (asynchronous) meetings can come to help. That is, for those who missed a meeting they can watch the meeting off-line at a later time.

Traditionally, off-line meeting reviewing experience is not satisfactory. For example, most of the existing meeting recording systems capture a meeting into WMV or AVI format that can be played back by Windows Media Player or RealPlayer. This traditional viewing mode only provides a linear access to the meeting content, which is far from effective. In the past few years, several new meeting recording systems emerged, and the trend is to provide rich and non-linear access to recorded meetings so that off-line reviewers can have similar experience as those who were in the live meetings [2]. For example, in [9][11], the authors used



Figure 1: RingCam: an inexpensive omni-directional camera and microphone array designed for capturing meetings.

focus of attention and audio speaker identification (ID) to construct indices for recorded meetings. In [1][4], microphone array sound source localization was used to segment speakers and construct meeting timelines. These timelines and indices not only provide a non-linear way to access the meeting, but can also signify interesting events in a meeting [1][11]. They allow off-line viewers to quickly find segments of interest and skip un-related segments. In this sense, participating in a rich off-line meeting can sometimes more time efficient than attending a live meeting.

While these timelines and indices are useful, constructing them automatically and reliably is not an easy task. For example, sound source localization can only tell where the sound is coming from, but cannot address the question "I only want to view the segments where my boss John was talking". By using audio speaker ID, e.g., [9], it is possible to construct person-based indices. However, the accuracy of audio speaker ID is still far from satisfactory, especially when the training data is limited. In [11], the authors used Eigenface and dynamic-space-warping based face recognition for speaker ID, and the preliminary results were reasonable. However, the meeting environment where they tested the recognition is relatively easy. For meetings that we recorded in real life, e.g., regular weekly team meetings, the environment is much more difficult.

Figure 1 shows a meeting recording device, called RingCam, that Microsoft Research developed to record 360-degree audio and video in a meeting [2]. An example recorded video frame is shown in Figure 2. Across the video frames, there are large variations in lighting conditions, people's head poses and head sizes. While there exist rich research in generic face recognition, recognizing faces in recorded meetings is still an understudied area, and that is the focus of this paper. In Section 2, we introduce a new face recognition framework, MRC-Boosting [10], which is able to handle large appearance variations in face images. In Section 3, we analyze unique challenges and opportunities in recorded meetings and propose the pre- and post-processing steps. In Section 4, we report experimental results, comparing the performance of MRC-Boosting and two other representative methods. We conclude the paper in Section 4.

## 2. MRC-BOOSTING FOR FACE RECOGNITION

Our previous research [10] demonstrated that face recognition can be modeled as a "target detection" problem, which is a special category in the two-class discrimination problems. Elad *et al* [3] showed that for a "target detection" type problem, Maximal-Rejection-Classifier (MRC) is an effective approach to find the most discriminative projection vectors. MRC is an iterative

Figure 2: Panoramic meeting video captured by the RingCam

method. In the training stage, a linear projection vector $\mathbf{w}^*$ (which we call *MRC feature*) is obtained through solving:

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \frac{\mathbf{w}^T \mathbf{R}_X \mathbf{w}}{\mathbf{w}^T \left( \mathbf{R}_X + \mathbf{R}_Y + (\mathbf{m}_X - \mathbf{m}_Y)(\mathbf{m}_X - \mathbf{m}_Y)^T \right) \mathbf{w}},$$

where $(\mathbf{m}_X, \mathbf{R}_X)$ and $(\mathbf{m}_Y, \mathbf{R}_Y)$ are the mean-covariance pairs of the target and clutter class respectively. Since the functional to be minimized is a generalized Rayleigh quotient, the optimal $\mathbf{w}^*$ can be conveniently computed through generalized eigenvalue decomposition.

Although MRC approach is able to find the discriminative component classifiers based on the MRC features, the final classifier is obtained by combining the component classifiers via simple AND rule. Therefore, it can only construct a convex (more precisely, a parallelogram polytope) decision region for the target class, which has limited capability of tackling complex classification problem. To address this issue, we put the MRC features into the boosting framework, so that a strong classifier with good generalization capability and computational efficiency can be constructed.

The MRC-Boosting approach for face recognition follows the general framework suggested by Moghaddam et al [5], where face recognition is reduced to a two-class (intra-/extra-personal) classification problem. In the training stage we are given a set of training faces $\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\}$, with known identities $\{c_1, c_2, \cdots, c_N\}$. Taking difference between each pair of training faces generates $N^2$ differences which constitute the training sample set. The difference $\mathbf{d}_{ij} \equiv \mathbf{x}_i - \mathbf{x}_j \ (i, j = 1, 2, \cdots, N)$ is called intra-personal if $c_i = c_j$, and extra-personal otherwise. The training task is to learn a classifier that can discriminate samples from the two classes.

Like other boosting method such as AdaBoost, the training of MRC-Boosting is iterative. Each difference sample $\mathbf{d}_{ij}$ in the training set carries a weight $w_{ij}$ (which will be dynamically adjusted through the training procedure). Note that $w_{ij} = w_{ji}$ since two symmetric differences $\mathbf{d}_{ij}$ and $\mathbf{d}_{ji} = -\mathbf{d}_{ij}$ are equivalent. In each iteration, a discriminative projection vector (i.e. MRC feature) is computed from the *weighted* training samples. In order to find the MRC feature, we only need the covariance matrix of weighted intra-personal differences $\mathbf{S}_I$ and that of the weighted extra-personal differences $\mathbf{S}_E$. It is shown in [10] that direct calculation of these matrices is expensive and a far more efficient way is as follows. We define intra-personal and extra-personal weighting matrices: $\mathbf{W}_I(i,j) = \begin{cases} w_{ij} & ,c_i = c_j \\ 0 & ,c_i \neq c_j \end{cases}$,

$\mathbf{W}_E(i,j) = \begin{cases} w_{ij} & ,c_i \neq c_j \\ 0 & ,c_i = c_j \end{cases}$. Then the covariance matrices can be

obtained as $\mathbf{S}_I = 2\mathbf{X}\tilde{\mathbf{W}}_I \mathbf{X}^T$ and $\mathbf{S}_E = 2\mathbf{X}\tilde{\mathbf{W}}_E \mathbf{X}^T$, where $\tilde{\mathbf{W}}_I = \mathrm{diag}(\mathbf{W}_I \mathbf{e}) - \mathbf{W}_I$ and $\tilde{\mathbf{W}}_E = \mathrm{diag}(\mathbf{W}_E \mathbf{e}) - \mathbf{W}_E$. The constant vector $\mathbf{e} = \begin{bmatrix} 1, & \cdots, & 1 \end{bmatrix}^T$ is introduced for compact expression, and $\mathrm{diag}(\mathbf{v})$ denotes a diagonal matrix formed by the elements of vector $\mathbf{v}$.

Once the two covariance matrices are calculated, generalized eigenvalue decomposition is used to find the MRC feature. Then a weak classifier is obtained. In the end of an iteration, the weights of the training samples are adjusted according to whether the weak classifier correctly classifies them. The complete algorithm is given in Figure 3.

In the recognition phase, the learned MRC-Boosting classifier $S(\mathbf{p}, \mathbf{g}) = \sum_{k=1}^{K} \alpha_k f_k(\mathbf{p}, \mathbf{g})$ is applied to measure the similarity between a probe face $\mathbf{p}$ and each gallery face $\mathbf{g}$. Finally $\mathbf{p}$ is recognized to be of the same identity as $\mathbf{g}$ that gives the largest $S(\mathbf{p}, \mathbf{g})$.

## 3. FACE RECOGNITION IN MEETINGS

We have shown in our previous work [10] that MRC-Boosting works very well for generic face recognition tasks, e.g., on the CMU-PIE database. However, recognizing faces in recorded meetings imposes several new challenges and opportunities.

- *Lighting*: the lighting can change significantly in a meeting. It can be bright when the meeting room lights are on (e.g., people are discussing or writing on the whiteboard). It can also be dark, when a presentation is on going. Furthermore, the colors of the slides can also affect the lighting/reflection.
- *Face resolution*: while RingCam captures very high resolution images, the resolution of the images used for face recognition can be low. First, because of DSP chip bandwidth constraints, RingCam does not provide a resolution as high as that of a still digital camera. Secondly, some meeting attendees can sit far from the camera, resulting in face images with especially low resolution (e.g. 10x10 pixels).
- *Head poses*: while this is considered a generic problem in face recognition, much bigger pose variations appear in recorded meetings. For example, people may turn their head significantly to talk or write on the whiteboard.
- *Temporal coherence*: unlike the above three challenges, this is an opportunity in recorded meetings. By observing and enforcing temporal coherence we can increase the recognition accuracy.

### 3.1. Pre-processing for lighting and head sizes

To handle the varying environmental lighting, intensity normalization is applied to all the face images extracted from the video. For a face image $I(x, y)$, the normalized version is given by:

**Input:** Training faces $\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\}$ and known identities $\{c_1, c_2, \cdots, c_N\}$.

**Initialize:** $w_{ij} = \begin{cases} \frac{1}{2N_I} & , c_i = c_j \\ \frac{1}{2N_E} & , c_i \neq c_j \end{cases}$ where $N_I$ and $N_E$ are the numbers of the intra-personal and extra-personal differences, respectively. The maximal number of weak classifiers $K$.

For $k = 1, 2, \cdots, K$

- Compute intra-personal and extra-personal covariance matrices $\mathbf{S}_I$ and $\mathbf{S}_E$, using the efficient way suggested.

- Find optimal weighted MRC vector $\mathbf{w}^*$ through solving
$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \frac{\mathbf{w}^T \mathbf{S}_I \mathbf{w}}{\mathbf{w}^T \mathbf{S}_E \mathbf{w}}.$$

- Obtain a weak classifier:
$$f_k(\mathbf{x}_i, \mathbf{x}_j; \mathbf{w}, T) = \begin{cases} +1 & , \left| \mathbf{w}^T(\mathbf{x}_i - \mathbf{x}_j) \right| \leq T \\ -1 & , else \end{cases}.$$

  The threshold $T$ is determined by minimizing classification error $\varepsilon_k = \sum_{i=1}^{N} w_{ij} I\left( f_k(\mathbf{x}_i, \mathbf{x}_j) \neq \lambda_{ij} \right)$,

  where $\lambda_{ij} = \begin{cases} +1 & , c_i = c_j \\ -1 & , c_i \neq c_j \end{cases}$.

- Updating weights: $w_{ij} \leftarrow \frac{1}{Z_k} w_{ij} \exp\left[ -\alpha_k \lambda_{ij} f_k(\mathbf{x}_i, \mathbf{x}_j) \right]$, where

  $\alpha_k = \frac{1}{2} \ln \frac{1 - \varepsilon_k}{\varepsilon_k}$ and $Z_k$ is a normalization factor to ensure $\sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij} = 1$.

**Output:** Strong classifier $F(\mathbf{p}, \mathbf{g}) = \text{sgn}\left[ S(\mathbf{p}, \mathbf{g}) \right]$, where $S(\mathbf{p}, \mathbf{g}) = \sum_{k=1}^{K} \alpha_k f_k(\mathbf{p}, \mathbf{g})$ is the similarity measure of two faces $\mathbf{p}$ and $\mathbf{g}$.

Figure 3: MRC-Boosting training algorithm for face recognition

$$\tilde{I}(x, y) = \left[ I(x, y) - \mu \right] / s,$$

where $\mu = \frac{1}{N} \sum_{x,y} I(x, y)$ and $s = \sqrt{\frac{1}{N} \sum_{x,y} \left[ I(x, y) - \mu \right]^2}$ are the mean and standard deviation of the $I(x, y)$'s pixel intensities, respectively. Through this operation, the intensity ranges of all images are normalized, so that the variations caused by lighting changes are alleviated.

The small face size is another significant problem encountered in practice. The consequence of the low face resolution is that after the face images are extracted from the video and normalized to be of a common size (24x24 in our experiments), many images appear to be blurred. Since in the blurred images the neighboring pixels are highly *correlated*, the actual dimensionality of their ensemble is much lower than the number of the pixels $D$ ($D$=576 in our case). To address this problem, the first step of training involves the use of PCA to perform dimensionality reduction to the face images, thus for each face image $\mathbf{x}_i$ we can obtain a $d \ll D$ dimensional feature vector $\tilde{\mathbf{x}}_i$, i.e.

$\tilde{\mathbf{x}}_i = \mathbf{P}\mathbf{x}_i$, where $\mathbf{P}$ is the $d$-by-$D$ PCA dimensionality reduction matrix. The standard MRC-Boosting training algorithm is then applied to the feature vectors $\{\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, \cdots, \tilde{\mathbf{x}}_N\}$ instead.

After the training, the projection vectors learned via MRC-Boosting can be transformed back into $D$ dimensional projection vectors:
$$\mathbf{w}' = \mathbf{P}^T \mathbf{w},$$
which will replace the original $d$ dimensional $\mathbf{w}$ and be applied on the face images in the recognition phase.

### 3.2. Post-filtering for temporal coherence

An advantage we can take in recorded meetings is that the recognition is based on a video sequence, instead of many independent still images. Therefore, the *temporal* correlation between neighboring frames can be utilized to improve the recognition accuracy. While there exist sophisticated methods, e.g., [12], they are computationally expensive, thus may not be feasible to analyze long meeting video sequences. Therefore, a much faster scheme was employed in our experiments, which we call *identity filtering*. What needs to be done is a post-processing after face recognition is done on all the frames *independently*. Suppose for a video sequence containing $T$ frames, image based face recognition gives the identities $\{c_t \in \mathcal{C} \mid t = 1, 2, \cdots, T\}$, where the subscripts indicate the frame indices and $\mathcal{C}$ is the set of all candidate identities. The filtered version of the identities is obtained as:
$$\tilde{c}_t = \max_{\chi \in \mathcal{C}} \sum_{i=t-K}^{i=t+K} \mathbf{I}(c_i = \chi),$$
where $\mathbf{I}(\bullet)$ is the indicator function equaling to 1 if the condition is satisfied and 0 otherwise. Intuitively, the filtered identity is the one that received most votes from the neighboring frames. In this way, spurious identities given by image based recognition can be corrected. It will be shown in Section 4 that this post-processing is able to significantly reduce the error rates of image-based face recognition algorithms.

### 4. EXPERIMENTS

Video sequences (~9,000 frames) captured from three different real-world meetings using RingCam were used in our experiments. To evaluate the recognition accuracy, we also hired external contractors to ground-truth the identity and face location of all the meeting attendees every 15 frames. Using the ground truth, the face region images were cropped out from video frames, so that we have a face database containing 14 people, and 3534 images in total, all with the resolution of 24x24. Sample face images are shown in Figure 4. It can be observed that there exist large variations in the appearance of the face images, due to partial occlusion (from hands), and the drastic changes of lighting condition, head pose and facial expression (including the effect caused by speaking), all of which are common in real-world meetings. Also can be noticed is the low resolution of the images of some subjects.

Following the standard protocol for face recognition experiments [6], the 3534 face images were randomly partitioned into three *disjoint* sets: the training set, the gallery, and the probe set. In our experiments, 50 images per person (700 in total) were used for training, the gallery contains 10 images for each person,

Figure 4: Sample images from the face database used in our experiments.

and the rest images were used for testing. This experimental setting is quite challenging, since compared to the large variations in the probe images, a gallery size of 10 images/person is rather small. This setting intends to simulate the true scenario, where it is often not possible to collect a lot of gallery images for each person.

We compared our algorithm with two popular existing methods, namely Eigenface [8][7] and the Bayesian method [5]. Eigenface is the most widely used traditional method for face recognition, and we employ it as a baseline approach. The Bayesian method is an influential method proposed more recently, which has been shown to be good at modeling the variations in face appearance. We performed face recognition experiment using these three algorithms separately with the setting stated above, and recorded the rank-1 recognition accuracy of each algorithm. As we mentioned in Subsection 3.1, before the MRC-Boosting training, PCA is applied to reduce the dimensionality of all face images from $D$=576 to a lower one $d$=150. For fair comparison, Eigenface method also employs a 150 dimensional PCA subspace. And for the Bayesian method, both of the intra-personal and extra-personal subspaces are of 75 dimensions. The same experiment was performed 20 times, each time with a different random partition of the data. The average and standard deviation of the recognition error rates achieved by three methods are listed in Table 1.

|  | Error Rate (%) | Std. Dev. (%) |
|---|---|---|
| MRC-Boosting | 6.034 | 0.443 |
| Bayesian | 9.592 | 1.796 |
| Eigenface (PCA) | 44.33 | 1.956 |

Table 1: The performance of three face recognition methods

Eigenface did a rather poor job, with an error rate of more than 40%. This is not surprising because there are very large variations in the probe face images, due to the complex environments in our real-world meeting videos. Eigenface does not have the ability to discriminate faces of different people under this condition. Bayesian method did a much better job, since it directly models the variations. However, the proposed MRC-Boosting algorithm achieved the best performance. Furthermore, it also showed higher robustness than the Bayesian method, with a standard deviation lower than a quarter of the latter's.

We also applied the *identity filtering* scheme to fuse the recognition results of adjacent frames, as we discussed in Subsection 3.2. As shown in Table 2, the recognition accuracy of all the three methods was improved by this post-processing scheme. Specifically, the error rate of MRC-Boosting algorithm was lowered by nearly an order of magnitude. The improvement for the other two methods was not as significant. Again, MRC-

Boosting was shown to be more robust than both Bayesian and Eigenface, giving a much smaller standard deviation.

|  | Error Rate (%) | Std. Dev. (%) |
|---|---|---|
| MRC-Boosting | 0.724 | 0.512 |
| Bayesian | 1.954 | 1.838 |
| Eigenface (PCA) | 41.35 | 2.581 |

Table 2: The performance of three face recognition methods (with identity filtering)

## 5. CONCLUSION

In this paper, we presented a promising face recognition algorithm combined with pre- and post-processing modules specifically designed for unique challenges in recorded meetings, e.g., changing lighting conditions, partial occlusions, and large variations in head pose and size. Experiments showed that the proposed approach achieved encouraging performance on real-world meeting videos, and is more accurate and robust than two representative and popular traditional approaches.

As for future work, one immediate direction is to integrate our face detection/tracking sub-system [2] with the proposed face recognition sub-system. Another interesting direction is to develop new sensor fusion techniques for better identity recognition. RingCam captures both audio and video. While face recognition utilizes video data for identity recognition, we are working on new spectrum-based speaker ID using captured audio data. We envision that by giving off-line viewers useful meeting indices and timelines, we can significantly enrich their experience.

## REFERENCES

[1]. T. Chen, W. Yin, X.-S. Zhou, D. Comaniciu, T. S. Huang, "Illumination Normalization for Face Recognition and Uneven Background Correction Using Total Variation Based Image Models", *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.

[2]. R. Cutler, Y. Rui, et. al., "Distributed meetings: a meeting capture and broadcasting system", *Proc. of ACM Multimedia*, pp. 503-512, 2002

[3]. M. Elad, Y. Hel-Or, and Renato Keshet, "Pattern Detection Using a Maximal Rejection Classifier", *Pattern Recognition Letters*, Vol. 23, No. 12, pp. 1459-1471, October 2002.

[4]. D. Lee, B. Erol, J. Graham, J. Hull, N. Murata, "Portable meeting recorder", *Proc. of ACM Multimedia*, pp. 493-502, 2002

[5]. Moghaddam, B.; Jebara, T.; Pentland, A., "Bayesian Face Recognition", *Pattern Recognition*, Vol 33, Issue 11, pps 1771-1782, November 2000

[6]. P. Jonathon Phillips, Hyeonjoon Moon, Syed A. Rizvi and Patrick J. Rauss, "The FERET Evaluation Methodology for Face-Recognition Algorithms", *IEEE Transactions on Pattern Analysis and Machine Intelligence,* pp 1090-1104, v22, n10, Oct. 2000.

[7]. Turk, M., Pentland, A., "Eigenfaces for Recognition", *Journal of Cognitive Neuroscience*, Vol. 3, No. 1, 1991

[8]. L. Sirovich and M. Kirby, "Low-Dimensional Procedure for the Characterization of Human Faces", *J. Optical Soc. Am. A*, vol. 4, pp. 519-524, 1987.

[9]. R. Stiefelhagen, J. Yang and A. Waibel, "Modeling focus of attention for meeting indexing", *Proc. of ACM Multimedia*, pp. 3-10, 1999

[10]. Xun Xu and Thomas S. Huang. "Face Recognition with MRC-Boosting", pp. 1770-1777, *Tenth IEEE International Conference on Computer Vision (ICCV'05)* Volume 2, 2005.

[11]. J. Yang, X. Zhu, R. Gross, J. Kominek, Y. Pan, and A. Waibel, "Multimodal people ID for a multimedia meeting browser", *Proc. of ACM Multimedia*, pp. 159-168, 1999

[12]. S. Zhou and R. Chellappa, "Simultaneous tracking and recognition of human faces from video," *Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing*, 2003.