Multiple Face Recognition from Omnidirectional Video

Fernando De la Torre Carlos Vallespi Paul E. Rybski Manuela Veloso Takeo Kanade
Robotics Institute. Carnegie Mellon University. 5000 Forbes Ave.Pittsburgh, PA 15213
ftorre@cs.cmu.edu, cvalles@cs.cmu.edu, prybski@cs.cmu.edu, veloso@cs.cmu.edu, tk@cs.cmu.edu

Abstract

Meetings are an integral part of business life. In previous work, we have developed a physical awareness system called CAMEO (Camera Assisted Meeting Event Observer) to record and process audio/visual information of a meeting. A very important task in meeting understanding is to know who is attending to the meeting and CAMEO's task is to infer people's identity from video. In this paper, we present an approach to identify people from an omnidirectional video sequence. Two main novelties are proposed: first a new dimensionality reduction technique MODA (Multimodal Oriented Discriminant Analysis) is used to perform fast matching and second we show that using multiple spatio-temporal constraints the recognition performance greatly improves. The effectiveness and robustness of the proposed system is demonstrated over several real time experiments and a large data set of videos.

1 Introduction

Meetings are an integral part of business life. In fact, approximately 11 million business meetings are held every day in the United States [1, 9]. A mid-level manager or professional spends around 35% of his time in meetings, and this percentage increases as a person advances up the company ladder [9]. On the other hand, meetings are not always as productive as expected. Among professionals who meet on a regular basis, 96% miss all or a part of a meeting, 73% have brought other work to the meeting, 39% have dozed during a meeting, and many of those attending a meeting need to clarify miscommunications. Having systems that help to review and share meetings can help to improve these undesirable situations. In previous work, we have introduced CAMEO (Camera Assistant Meeting Event Observer) a hardware/software system to record and process audio-visual information as a first step towards understanding human interactions in meeting [4, 10].

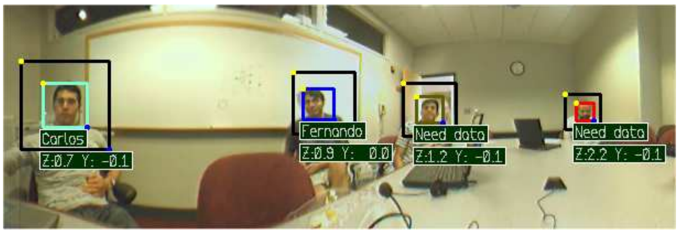

Figure 1: Multiple face recognition from a panoramic image.

A very important task in meeting understanding is to know who is assisting to the meeting and CAMEO's task is to infer people's identity from the mosaic images. Face recognition from images/video is quite complex problem which suffers from misalignment, high dimensionality of the visual data, occlusions, facial expression changes and illumination variations. Due to its difficulty and usefulness as a biometric, there exist a huge literature and there are many available techniques for face recognition from images (see [19] for a review). In our particular application, we are interested in developing efficient recognition methods which can work in real time. In this paper, we use a recently introduced dimensionality reduction technique, Multimodal Oriented Discriminant Analysis (MODA) to optimally reduce the dimension of the data for fast recognition. Moreover, in order to deal with unexpected variation changes, several strategies which exploit spatio-temporal redundancy and are able to recognize several people simultaneously are introduced to improve performance. Figure 1 shows an example where several people are recognized simultaneously.

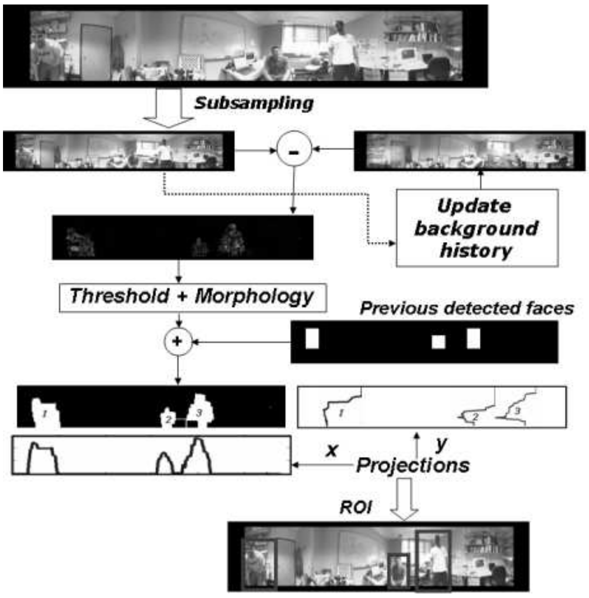

Figure 2: Algorithm to detect regions of interest.

2 Detecting and gathering faces

In order to track and recognize faces efficiently, a first step is to detect regions where potentially can be moving objects. In this section, we describe a simple but effective region of interest (ROI) algorithm, assuming no camera motion (the camera is static in the center of the table). Details of the hardware and mosaic construction are explained in [4].

When CAMEO starts, it computes an estimate of the background by averaging the incoming mosaic images during several seconds. Once an estimation of the background is given, we subtract a multi-resolution version of the mosaic image from the estimated background, this will provide a first estimate of the ROI. The difference is thresholded and an opening (morphological operator) with an structural element of 3 × 3 used to eliminate spurious noise. At this point, we have a binary image with blobs corresponding to compact regions with graylevel changes and the previously detected/tracked faces' area are added to enforce temporal consistency (in case that the person do not move the head). In order to label the blobs, we project the image into its x and y coordinates. From segmenting the x-projection, it is easy to estimate the y component and build a bounding box around the area of interest. Finally, the background regions which do not belong to the ROI are used to update the estimation of the background. Figure 2 illustrates the ROI algorithm.

Once the ROIs are computed, the next step for face recognition is to construct a statistical model of facial expression/pose/illumination variations of a person. In order to gather data, a person sits in front of one CAMEO's camera and performs different facial expressions under several pose/illumination changes (approximately 1 minute of video is recorded). After the ROIs are computed, we run the Scheniderman face detector [11, 12] to detect frontal faces. Figure (3.a) shows some original frontal faces gathered (60×60 pixels), approximately 800 frontal faces.

The face detector occasionally gives some false positives. In order to filter these potential outliers, a simple scheme based on color [17] is used. Using the normalized r, g components ($r = \frac{R}{R+G+B}$, $g = \frac{G}{R+G+B}$), a Gaussian distribution $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is fitted to a set of training faces, where $\boldsymbol{\mu} \in \Re^{2\times 1}$ and $\boldsymbol{\Sigma} \in \Re^{2\times 2}$ are the mean and covariance that approximate the skin color. Given a new face, CAMEO computes the percentage of skin color pixels inside the patch containing the face, and the percentage of skin color pixels in the surrounding area. In an ideal face, the ratio of skin pixels in the area containing the face versus the percentage of skin pixels in the surrounding area should be big. If this ratio is below a certain threshold the sample face is discarded.

3 Preprocessing

Once the frontal faces have been gathered, in order to compensate for small scale and translational factors, we register the data using parameterized component analysis [2] to achieve geometric invariant learning. After geometrically align the faces, we perform dimensionality reduction in samples and features, which is a common technique to filter noise and makes algorithms more computationally tractable.

Let $\mathbf{D}^i \in \Re^{d\times n_i}$[1] a data matrix containing the gathered images for class i, usually $n_i \approx 800$. There are two type of "dimensions" that would be interesting to reduce. The first one consist on selecting the p most representative samples where ($p << n_i$). Let $\mathbf{D} \in \Re^{d\times(p\times c)}$ the matrix containing the data from all the classes (c), where each column $\mathbf{d}_i$ is a vectorized image. The second dimensionality reduction will consist on reducing the existing redundancy in the column space of $\mathbf{D}$ by finding k basis ($k << d$), which spans the subspace of maximum variation.

[1] Bold capital letters denote a matrix $\mathbf{D}$, bold lower-case letters a column vector $\mathbf{d}$. $\mathbf{d}_j$ represents the j column of the matrix $\mathbf{D}$. All non-bold letters will represent variables of scalar nature. *diag* is an operator which transforms a vector to a diagonal matrix. $\mathbf{I}_k \in \Re^{k\times k}$ is the identity matrix. $tr(\mathbf{A}) = \sum_i a_{ii}$ is the trace of the matrix $\mathbf{A}$. $||\mathbf{A}||_F = tr(\mathbf{A}^T\mathbf{A}) = tr(\mathbf{A}\mathbf{A}^T)$ designates the Frobenious norm of a matrix. $N_d(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ indicates a d-dimensional Gaussian on the variable $\mathbf{x}$ with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$.

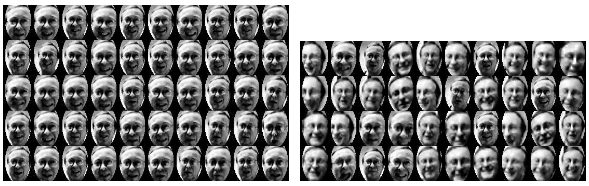

Figure 3: a) Original training images (800). b) 40 out of 75 clusters.

When processing large videos of the same person, the amount of redundant facial expression/poses becomes an issue for several reasons. Firstly, we do not necessarily have a uniform sampling of all the possible facial expressions/poses. This can bias the posterior classifier, and secondly and more importantly, the huge amount of data would make very computationally intensive training any classifier. To avoid this phenomena, once the images are registered, we find the most representative prototypes (≈ 75) by clustering $\mathbf{D}^i$, using the recent advances in multi-way normalized cuts [18]. Figure (3.a) shows some images of the original 800 samples and (3.b) represents 40 (out of 75) prototypes. Observe that the variations are mostly due to facial expression/scale/illumination changes.

Each gathered image is rescaled to be 60×60 pixels, which can be represented in a 3600-dimensional vector, where most of the dimensions are highly correlated. This redundancy will be an inconvenient for several reasons, the first one is that any type of discriminative learning (e.g. classifiers) will suffer over-fitting effects with lots of correlated data and secondly we are interested in reducing the computational burden of the overall algorithm. In order to reduce the dimensions of the column space of $\mathbf{D}$, we use principal component analysis (PCA) , previous normalization of the data ($||\mathbf{d}_i|| = 1 \ \forall \ i$ and subtract the mean). PCA projects the data into the subspace spanned by the eigenvectors of the covariance matrix $\mathbf{D}\mathbf{D}^T \in \Re^{d \times d}$, however for big amounts of data where ($d >> n$), it is more numerically convenient to compute the eigenvectors of $\mathbf{D}^T\mathbf{D} \in \Re^{n \times n}$ [15]. $\mathbf{D}\mathbf{D}^T$ has the same eigenvalues as $\mathbf{D}^T\mathbf{D}$ and the eigenvectors are related by $\mathbf{D}$. Observe that projecting onto the principal components some discriminatory power could be lost, however it is worth to point out a couple of aspects. First, by projecting onto the principal components the generalization performance in the case of $d >> n$ probably will be better and secondly the true dimensionality of $\mathbf{D} \in \Re^{d \times n}$ when $d >> n$ is n, so by projecting into the first k eigenvectors which eigenvalues are different from zero no discriminatory power is lost. We project on the PC which preserve 99% of the energy.

4 Multimodal Oriented Discriminant Analysis

Given a test image our goal is to perform fast classification into one of the c classes. One naive solution will be to match each test image with each of the prototypes (75) in the class i, however this nearest neighbor classifier is not very efficient since k dimensions (number of principal components) have to be matched. In this section, we use Multimodal Oriented Discriminant Analysis (MODA) a recently introduced method [3] which generalizes Linear Discriminant Analysis (LDA). LDA is only optimal for Gaussian distributed classes, whereas MODA accomodates multimodal distributed classes with different covariances. MODA will allow to perform fast matching ($m << k$), avoids overfitting and improves recognition performance w.r.t. LDA.

For each class we have around 800 original images and 75 prototypes. Given the original 800 images, we further cluster ([18]) into s clusters, typically between 2 and 5, which mostly cluster pose and scale changes. Each of these clusters is going to be modeled as a high dimensional Gaussian $N(\mathbf{x}; \boldsymbol{\mu}_i^r, \boldsymbol{\Sigma}_i^r)$. MODA seeks for a low dimensional projection $\mathbf{B} \in m \times k$, common to all the classes (i.e. $N(\mathbf{B}^T\boldsymbol{\mu}_i^r, \mathbf{B}^T\boldsymbol{\Sigma}_i^r\mathbf{B}) \ \ \forall i, r$) that maximizes the Kullback-Leibler (KL) divergence [5] between the clusters of different classes in the low dimensional space, but do not impose distance constraints between the clusters of the same class. That is, MODA maximixes:

$$E(\mathbf{B}) = \sum_i \sum_{j \neq i} \sum_{r_1 \in C_i} \sum_{r_2 \in C_j} KL_{ij}^{r_1 r_2} \qquad (1)$$

where $KL_{ij}^{r_1 r_2} = tr((\mathbf{B}^T\boldsymbol{\Sigma}_i^{r_1}\mathbf{B})^{-1}\mathbf{B}^T\boldsymbol{\Sigma}_j^{r_2}\mathbf{B} + (\mathbf{B}^T\boldsymbol{\Sigma}_j^{r_2}\mathbf{B})^{-1}\mathbf{B}^T\boldsymbol{\Sigma}_i^{r_1}\mathbf{B} - 2\mathbf{I}) + (\boldsymbol{\mu}_i^{r_1} - \boldsymbol{\mu}_j^{r_2})^T\mathbf{B}(\mathbf{B}^T\boldsymbol{\Sigma}_j^{r_2}\mathbf{B})^{-1} + (\mathbf{B}^T\boldsymbol{\Sigma}_i^{r_1}\mathbf{B})^{-1})\mathbf{B}^T(\boldsymbol{\mu}_i^{r_1} - \boldsymbol{\mu}_j^{r_2})$ is the Kullback-Leibler (KL) divergence between the cluster r_1 of class i and the cluster r_2 of class j after projecting with $\mathbf{B}$. After some algebraic arrangements, it can be shown that [3]:

$$E(\mathbf{B}) = \sum_i \sum_{r_1 \in C_i} tr\big((\mathbf{B}^T\boldsymbol{\Sigma}_i^{r_1}\mathbf{B})^{-1} \qquad (2)$$

$$(\mathbf{B}^T(\sum_{j \neq i} \sum_{r_2 \in C_j} (\boldsymbol{\mu}_i^{r_1} - \boldsymbol{\mu}_j^{r_2})(\boldsymbol{\mu}_i^{r_1} - \boldsymbol{\mu}_j^{r_2})^T + \boldsymbol{\Sigma}_j^{r_2})\mathbf{B})\big)$$

$r_1 \in C_i$ sums over all the clusters belonging to class i. Eq. 3 is quite hard to optimize w.r.t $\mathbf{B}$, second order type of gradient methods (e.g. Newton or conjugate gradient) do not scale well with huge matrices. Moreover, in this particular energy function the second derivative is quite complex. A bound optimization method called iterative majorization [7] is used instead.

Iterative majorization (similar to Expectation Maximization type of algorithms) is able to monotonically reduce the value of the energy function. For details of the optimization check [3].

Despite having reduced the dimensionality of the data from d to k with PCA, fitting discriminative models like MODA can easily suffer from over-fitting problems and lack of generalization. In order to be able to generalize better and do not suffer from storage and computational requirements, we approximate the covariance matrices as the sum of outer products plus a scaled identity matrix $\mathbf{\Sigma}_i \approx \mathbf{U}_i \mathbf{\Lambda}_i \mathbf{U}_i^T + \sigma_i^2 \mathbf{I}_k$. $\mathbf{U}_i \in \Re^{k \times l}$, $\mathbf{\Lambda}_i \in \Re^{l \times l}$ is a diagonal matrix. In order to estimate the parameters σ_i^2, $\mathbf{U}_i$, $\mathbf{\Lambda}_i$, a fitting approach is followed by minimizing $E_c(\mathbf{U}_i, \mathbf{\Lambda}_i, \sigma_i^2) = ||\mathbf{\Sigma}_i - \mathbf{U}_i \mathbf{\Lambda}_i \mathbf{U}_i^T - \sigma_i^2 \mathbf{I}_k||_F$. It can be shown that the optimal solution satisfies $\mathbf{U}_i \mathbf{\Sigma}_i = \mathbf{U}_i \hat{\mathbf{\Lambda}}_i$, $\sigma_i^2 = tr(\mathbf{\Sigma}_i - \mathbf{U}_i \hat{\mathbf{\Lambda}}_i \mathbf{U}_i^T)/k - l$, $\mathbf{\Lambda}_i = \hat{\mathbf{\Lambda}}_i - \sigma_i^2 \mathbf{I}_k$, [3]. The same expression could be derived from probabilistic assumptions [8, 14].

Observe that the original covariance has $k(k+1)/2$ free parameters, and with the factorized matrices $(\mathbf{U}_i, \mathbf{\Lambda}_i, \sigma_i^2)$ the number of parameters is reduced to $l(2k - l + 1)/2$ (assuming orthonormality of $\mathbf{U}_i$), so we need much less data to estimate these parameters and hence it is not so prone to over-fitting.

In order to test our approach, we have gathered a database of 23 people over two different days and different illumination conditions. Figure 4 shows some images of people in the database, variations are mostly due to facial expression, pose, scale and illumination conditions. The training set consist on the data gathered during the first day under three different illumination conditions (varying lights in the recording room), scale and rotation changes. The testing data consist on the recording of the second day (a couple of weeks later) under similar conditions (16 people were collected). Figure 5 illustrates the recognition performance using PCA, LDA and MODA, similarly table 1 give some detailed numerical values for different number of basis and techniques.

Basis	5	10	20	30	40	50
PCA	0.26	0.43	0.55	0.58	0.59	0.59
LDA	0.36	0.48	0.56	NA	NA	NA
MODA	0.38	0.50	0.59	0.61	0.62	0.63

Table 1: Recognition performance of PCA/LDA/ODA (23 classes)

Once the images are projected with MODA a nearest neighboor classifier is used for classification. Several metrics have been tested (e.g. Mahalanobis, Euclidian, Cosine, etc), however in our experiments the Euclidean distance is the one which performs the best. MODA is computed clustering each of the classes into 2 clusters and using the Euclidian distance for nearest neighborhood classifier (with 75 prototypes). For the same number of basis, MODA outperforms PCA/LDA. Also, observe that LDA can extract just classes-1 features (22 features) and that just matching a 20-dimensional vector the recognition rate achieves almost its upper bound.

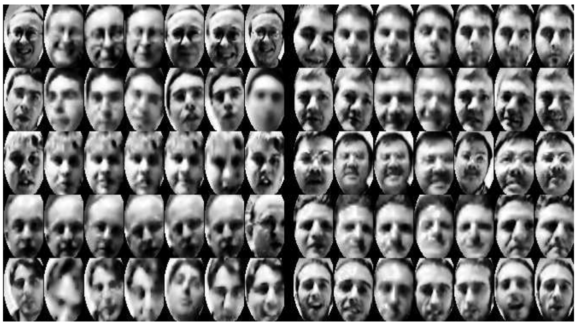

Figure 4: Some samples of the training data.

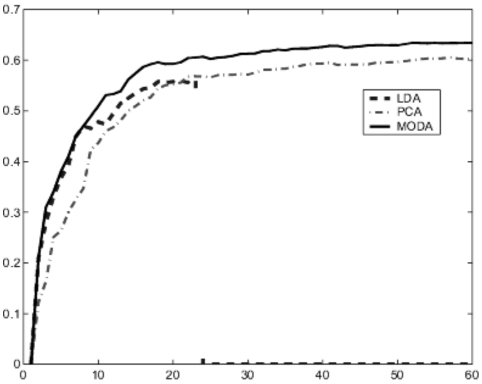

Figure 5: PCA/LDA/MODA

5 Improving performance

Although MODA is the optimal linear dimensionality reduction technique that preserves discriminability power among classes, it does not handle very well changes in the statistical properties of the training data due to undesirable noise (e.g. different illumination, hair configuration, misregistration). Three strategies have been tested in order to handle such unexpected situations:

- Use a verification step to reject samples (outlier detection).
- Integrate temporal information into the classification process.
- Recognize multiple people simultaneously.

5.1 Pattern verification

Face recognition from video has the advantage of having a lot of temporal redundancy that can be exploited to improve performance. In order to construct robust systems against surprise or untrained situations, not all of the adquired samples are going to be classified, only the ones that are reliable. To decide if a sample should be used for classification or not, a simple outlier detection strategy is used.

When CAMEO detects the face of a person, the first step it performs is to determine if the person is in the database or not. Two thresholds (one generative and the other discriminative) are used in order to determine if the sample belongs to any of the classes. Both thresholds are computed from the covariance of the projected data. That is, the training data have been projected onto the discriminative or generative subspace (i.e. $\mathbf{C} = \mathbf{B}^T\mathbf{D}$), and a threshold is established by the variance of this distribution, that is $\mathbf{C}\mathbf{C}^T$. The covariance naturally becomes diagonal (due to a first PCA step). Outliers will be considered the ones that are 4 times far away from this standard distribution.

If the first global step is not an outlier, the second one uses a verification method to check if the data belong to a local subspace. For each class i the principal components that preserve 80% of energy are computed, $\mathbf{U} \in \Re^{d\times k_i}$. The average error ($\mu_i$) and the variance ($\sigma_i$) of the distance from the subspace (DFS) ($||\mathbf{I}-\mathbf{U}\mathbf{c}||_2^2 = ||\mathbf{I}||_2^2 - \mathbf{c}^T\mathbf{c}$) for class i, are computed, as well as the average error (μ_o) and the average variance (σ_o) of the distance from the subspace for all the classes but class i. Once the mean and variance for the inter class DFS of class i, and the mean and variance for the intra class DFS for class i are obtained, a quadratic classifier that minimizes the classification error is calculated. That is, an optimal threshold (T), which reduces the classification error ($\int_T^\infty P_i p_i(x)dx = \int_{-\infty}^T P_o p_o(x)dx$), where P_i and P_o, are the a priori probabilities of each class. Assuming that $p_i(x)$ and $p_o(x)$ are Gaussian, the optimal threshold is given by the solution of the following second order equation:

$$T^2(\frac{1}{\sigma_i^2}-\frac{1}{\sigma_o^2})+2T(\frac{\mu_i}{\sigma_i^2}-\frac{\mu_o}{\sigma_o^2})+(\frac{\mu_i^2}{\sigma_i^2}-\frac{\mu_o^2}{\sigma_o^2}-2ln(\frac{P_i\sigma_o}{P_o\sigma_i}))$$

Figure 6 shows the comparison between MODA and MODA plus a generative verification step. The x-axis are the number of bases and the y-axis represents the recognition rate. Table 2 shows more detailed values for some basis and the percentage of data that has not been discarded. Discarding approximately half of the data greatly improves the recognition performance. In the real time implementation, all the thresholds, the global and local ones, vary proportionally to one parameter, which is adapted depending on the amount of data that has been classified.

Basis	5	10	20	30	40	50
MODA2	0.35	0.64	0.70	0.72	0.73	0.73
Per(%)	44%	49%	50%	51%	52%	50%

Table 2: Recognition performance of MODA and MODA+outliers. The % indicates the percentage of inlier data.

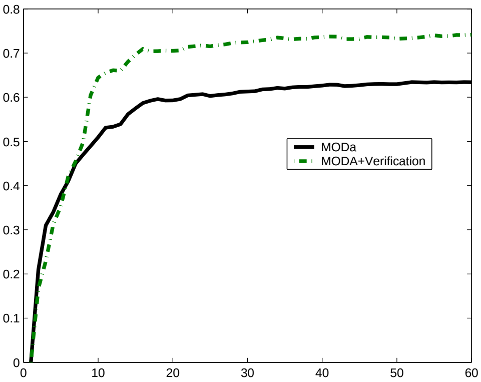

Figure 6: MODA and MODA + a verification step

5.2 Temporal consistency

Up to now CAMEO has used one single image to classify faces; however, a key aspect to improve recognition performance consist of integrating all the evidence coming from the video stream.

Preliminary work has been done to match sets of images rather than individual images. Yamaguchi et al. [16] propose to compute the principal angle as a distance between two sets. However, this measure is too sensitive to possible outliers, since it is based on computing the minimum correlation between two vector (one on each subspace). Shakhmarovich et al. [13] took into account a more probabilistic approach, modeling both sets as a high dimensional Gaussian, and then computing the Kullback-Leibler distance between distributions. In this paper, a simple voting scheme is used. The data is first projected with the basis obtained with MODA, and the class with less error is chosen. After the verification step, if the reconstruction error for this class falls within a range, the sample is used for classification. The same procedure is followed with the rest of the samples, and when 10 samples are collected CAMEO assigns the set of images to the class where more samples have been classified. Figure 7 compares the recognition performance w.r.t. just using MODA. Table 3 shows some numerical results.

Basis	5	10	20	30	40	50
MODA3	0.46	0.76	0.81	0.84	0.84	0.84

Table 3: Recognition peformance using MODA + Verification step + Temporal voting.

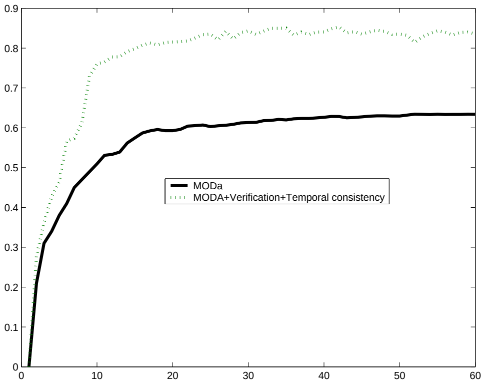

Figure 7: MODA and MODA + Verification step + Temporal voting

5.3 Multiple face recognition

Recognizing several people simultaneously greatly improves the recognition performance, since the constraint that two people can not have the same identity is explicitly imposed. To incorporate these constraints into the classification problem, the multiple face recognition problem is posed as the classical assignment (transport) problem.

A matrix **A** with m rows (number of people tested) and n columns (number of people in the database) is created, where each position a_{ij} correspond to the average error of the last 10 samples for person i w.r.t class j. The error is computed as the average of the minimum error from the prototypes of class i. Once this matrix is created, the assignment problem is the task of finding a permutation $\pi[0]$, $\cdots$, $\pi[n-1]$ of $\{0, \cdots, n-1\}$ such that $\sum_{k=0}^{n-1} a_{k\pi[k]}$ is minimized [6]. This is typically solved when $(m = n)$ with the Hungarian algorithm. In our case, $m < n$, and $n - m$ dummy rows with zero value are added, which do not affect the result of the algorithm. The code for the Hungarian algorithm has been copied from [6].

Fig. 8 shows the recognition performance exploiting spatio-temporal information and outlier detection vs. using just MODA. As we can see, around 40% of increase in performance is achieved. Imposing constraints in the multiple face recognition improves on average 5% of recognition performance and it can be time consuming for real time applications.

Basis	5	10	20	30	40	50
MODA4	0.52	0.74	0.87	0.87	0.89	0.90

Table 4: Multiple face recognition with a window of 10 samples.

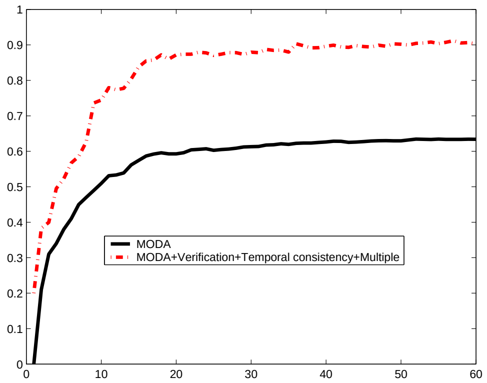

Figure 8: MODA and MODA + Verification + Temporal consistency + Multiple face Recognition.

6 On line learning new people

If CAMEO does not recognize the person during execution time, it will start gathering faces coming from the face detector. Running the face detector [11, 12] is quite costly computationally speaking, once a face is detected, a normalized template matching will be used to track the last template. However, correlation based trackers will not be robust to changes in pose or facial expression. Once CAMEO detects that the tracker is lost (the correlation coefficient decreases), it will run the frontal face detector [11, 12] again over the region of interest in a separate thread, until ≈ 100 faces are gathered this way. As in the off-line method, the faces are detected, but not perfectly registered and in order to construct a good linear generative model of the face of a person we need to properly register the images. In the off-line version, we used parameterized component analysis [2], however it can be computationally expensive to run on real time. Instead, we cluster the data into *cl* clusters and register for the translational component with normalized correlation w.r.t. the set of clusters (the one which gives less error). Later, the clusters are recomputed and we proceed that way until convergence. This procedure is not as optimal as [2], but it is suitable for fast on-line registration.

In order to select the number of clusters, we have measured the compactness of the data after the registration step is performed. The compactness is measured by computing the energy of a set of basis when the SVD is performed, The first row indicates the number of basis and the second row and beyond the percentage of energy preserved with this number of basis. We compare the % of energy rather than the total reconstruction error since due to interpolation errors the energy of two sets with different registration algorithms is not necessarily equal, so we can not compare, but a percentage of the energy is comparable. Table 5 shows the result.

Number of basis	2	3	6	12
1 cluster(mean)	0.14	0.24	0.43	0.63
2 clusters	0.32	0.39	0.54	0.69
4 clusters	0.27	0.41	0.56	0.70
6 clusters	0.27	0.39	0.58	0.72
8 clusters	0.26	0.36	0.55	0.70

Table 5: Energy vs. number of basis for different number of clusters.

As expected with more clusters the compactness is higher. However, registering w.r.t. 2 clusters usually gives a good trade off between speed and quality of the clustering. The results indicate that probably there exist two modes in the distribution of faces (likely to be due to scale factors).

7 Experiments

Several experiments have been performed to test our approach. 1 minute video of a typical meeting has been recorded and it can be downloaded from www.salleURL.edu/~ftorre/recognition3.avi . In this video, we can see 4 people having a meeting. CAMEO has a pre-stored facial model for all these 4 people. CAMEO is able to track the faces in real time [4] and after few frames it starts recognizing the people. At the beginning CAMEO is not sure of the labels and the face is labeled as 'Need data'. CAMEO sucessfully has recognized the 4 out of 10 people in the database. The system runs about 7-10 frames per second. CAMEO also computes the distance to the closer camera (lower black rectangle).

In this video, we have not imposed the multiple face recognition constraint, since it only increases 5% the recognition performance (in our database experiments) and it can be time consuming for real time applications. If two users have the same identity, we choose the one with less error. Although not the optimal solution, it is a reasonable approximation for real-time applications.

8 Conclusions and Future work

In this paper we have introduced an efficient and effective multiple face recognition system for CAMEO. Several novelties for dimensionality reduction and strategies to improve recognition from video sequences have been proposed. Also, on-line learning capabilities have been demonstrated. However, several aspects remain to be researched and extended. For instance, since face to face meeting encompasses several modalities such as speech, gesture or handwriting, these should be added to CAMEO's capabilities and use them as a biometric measure.

Acknowledgment

Thanks to Jordi Casoliva for the fast energy normalization code and preliminary versions of the face tracker based on normalized correlation. Thanks to R. Patil, J. White and F. Tamburrino for helpful comments and discussions. This work has been partially supported by National Business Center of the Department of the Interior under a subcontract from SRI International and U.S. Department of Defense contract N41756-03-C4024, NIMH Grant R01 MH51435. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, by the NBC, DOI, SRI or the US Government.

References

[1] M. Conferencing. Meetings in america: A study of trends, costs and attitudes toward business travel, teleconferencing, and their impact on productivity. In *A network MCI Conferencing White Paper.*, 1998.

[2] F. de la Torre and M. J. Black. Robust parameterized component analysis: theory and applications to 2d facial appearance models. *Computer Vision and Image Understanding*, 91:53 – 71, 2003.

[3] F. de la Torre and T. Kanade. Multimodal oriented discriminant analysis. In *tech. report CMU-RI-TR-05-03, Robotics Institute, Carnegie Mellon University, January 2005.*

[4] F. de la Torre, C. Vallespi, P. E. Rybski, M. Veloso, and T. Kanade. Omnidirectional video capturing, multiple people tracking and recognition for meeting understanding. Technical report, Robotics Institute, Carnegie Mellon University, January 2005.

[5] K. Fukunaga. *Introduction to Statistical Pattern Recognition, Second Edition.* Academic Press.Boston, MA, 1990.

[6] D. E. Knuth. *The Standford GraphBase.* Addison-Wesley Publishing Company, 1993.

[7] J. D. Leeuw. *Block relaxation algorihtms in statistics.* H.H. Bock, W. Lenski, M. Ritcher eds. Information Systems and Data Analysis. Springer-Verlag., 1994.

[8] B. Moghaddam and A. Pentland. Probabilistic visual learning for object representation. *Pattern Analysis and Machine Intelligence*, 19(7):137–143, July 1997.

[9] R. B. Nelson and P. Economy. Better business meetings. In *McGraw-Hill.*, 1995.

[10] P. E. Rybski, F. de la Torre, R. Patil, C. Vallespi, M. Veloso, and B. Browning. Cameo: Camera assisted meeting event observer. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2004.

[11] H. Schneiderman. Feature-centric evaluation for cascaded object detection.. In *CVPR*, 2004.

[12] H. Schneiderman. Learning a restricted bayesian network for object detection. In *CVPR*, 2004.

[13] G. Shakhnarovich, J. W. Fisher, and T. Darrell. Face recognition from long-term observations. In *ECCV*, 2002.

[14] M. Tipping and C. M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society B*, 61:611–622, 1999.

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

[16] O. Yamaguchi, K. Fukui, and K. Maeda. Face recognition using temporal image sequence. In *IEEE International Conference on Automatic Face and Gesture Recognition.*, 1998.

[17] J. Yang and A. Waibel. A real-time face tracker. In *Proceedings 3rd IEEE Workshop on Applications of Computer Vision*, pages 142–147, 1996.

[18] S. Yu and J. Shi. Multiclass spectral clustering. In *ICCV*, 2003.

[19] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *ACM computing surveys*, 35(4):399–458, 2003.