

TOWARDS FACE RECOGNITION AT A DISTANCE

Simon J.D. Prince¹, James Elder², Y. Hou², M. Sizinstev² and E. Olevskiy²

¹ Department of Computer Science, University College London, United Kingdom. s. prince@cs.ucl.ac.uk

² Center for Vision Research, York University, Toronto, Canada. jelder@yorku.ca

Keywords: Biometrics, Face Recognition, Security, Pose

Abstract

Current face recognition algorithms require the tacit cooperation of users, who must position themselves in a small area of space and face the camera. Face recognition in uncontrolled conditions, such as in security camera footage presents two extra challenges. First, it is difficult to capture good quality images of faces in this setting. Second, the pose of the face is relatively uncontrolled which causes most face recognition algorithms to fail. In this paper, we present a series of solutions to address these problems. High quality face images are captured using a foveated wide field sensor, in which a narrow-field camera is directed towards faces using information from a static wide-field camera. Feature points corresponding to the eyes/nose etc. are accurately localized and face shape is normalized. A novel algorithm is introduced to identify these (typically non-frontal) faces from a test gallery of frontal faces. Results are demonstrated to be superior to contemporary approaches.

1 Introduction

Commercial face recognition systems typically operate in restricted circumstances: they require the user to place themselves in front of the device and to face the camera. In this paper we summarize our results on determining identity using face recognition in a security setting. Consider surveillance equipment in a large open area such as an outdoor car park.

Challenge 1: In order to visualize the area the camera needs a wide field of view. Unfortunately, wide field of view comes at the expense of image resolution. For example, a human face at 5m will subtend only about 4×6 pixels on a 640×480 sensor with 130° field of view. This is insufficient resolution for biometric tasks. One solution is to have two cameras. A wide-field *pre-attentive* camera observes the whole scene at low resolution. We use data the output to orient a narrow field *foveal* camera towards faces in the scene using a pan-tilt motor.

There have been a number of prior efforts to combine narrow-field and wide-field sensors for human activity tracking [14, 6]. However, functionality of prior systems has been limited in various ways. For example, Scassellati's [14] system was designed to detect frontal human faces at very close range (3-4 feet). Greiffenhagen [6] et al. relied on extensive modeling of the scene and assumed people were standing upright. A major goal of our work has been to demonstrate how this kind

of sensor architecture can be useful for realistic surveillance problems.

Challenge 2: A second major problem is that the user is not necessarily facing the camera, and the lighting cannot of the face cannot easily be determined. Both of these factors cause significant degradation in performance in current face recognition algorithms [17]. Indeed, horizontal pose variation may be particularly extreme under these circumstances, and frequently change through $\pm 90^\circ$.

There have been two major approaches to the problem of face recognition across pose: one approach is to create a full 3D head model for the subject based on just one image [13, 2, 3] and compare 3D models to those in a database. This approach is feasible, although the computation involved is significant. An alternative approach is to build a linear model of the covariance of pixel intensity in frontal and non-frontal images [7] and use this to predict frontal from non-frontal images. This approach is fast, but the recognition accuracy of previous studies has been limited.

In this paper, we describe a system that meets both of these challenges: it is capable of gathering high quality face data over a wide area and performing face recognition which is relatively robust to even severe pose variations.

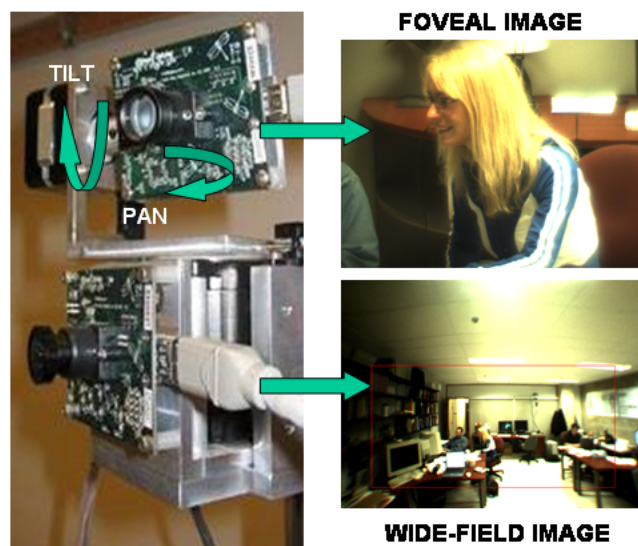


Figure 1: Our foveated sensor consists of a stationary camera (bottom) with a 130° field of view and a moving foveal camera (top) which is driven by pan and tilt motors to orient towards faces detected in the stationary camera.

2 Wide Field Sensor and Person Detection

The sensor consists of two 30Hz RGB Point Grey Dragonfly cameras (see Figure 1). The wide field camera is fixed in position and has a 2.1 mm lens subtending a 130 deg horizontal field of view. The foveal camera is mounted on a pan/tilt platform and has a 24 mm lens with a 13 deg horizontal field of view. The pan and tilt motors are Lin Engineering step motors, with a step size of 0.1 degrees. We aim to use the output from the wide-field camera to orient the foveal camera at faces in the scene.

The dominant paradigm in face/body detection is to combine weak classifiers using Adaboost to build a strong classifier based on supervised data [15, 8]. A number of factors limit the utility of these methods for far-field or wide-field face detection. First, they assume a minimum scale of at least 12×16 pixels. However, in our relatively unconstrained indoor testing environments, the median face size is 4×5 pixels, and many faces subtend less than a pixel (Figure 2). Moreover, faces are not restricted to being frontal or profile and heads may even face away from the camera. We would still like to orient the foveal camera towards them in this context as the head may turn toward the camera in subsequent frames. A different approach would be to detect the full body as in [16]. However, such dynamic approaches are designed to detect walking pedestrians, translating along a ground plane. In contrast, we are interested in wide field detection of people whether they are walking, turning, or remaining relatively still, whether they are standing or sitting, and whether or not they are partially occluded by a desk or other object.

We have developed a robust approach to face detection in these challenging circumstances which produces a map of the posterior probability that a human head is present at each point. This is derived from a vector \mathbf{D} of complementary cues. Letting \mathcal{H}_h denote the hypothesis that a head is present at a given pixel, and $\mathcal{H}_{\bar{h}}$ denote the absence of a head we have:

$$Pr(\mathcal{H}_h|\mathbf{D}) = \frac{Pr(\mathbf{D}|\mathcal{H}_h)Pr(\mathcal{H}_h)}{Pr(\mathbf{D}|\mathcal{H}_h)Pr(\mathcal{H}_h) + Pr(\mathbf{D}|\mathcal{H}_{\bar{h}})Pr(\mathcal{H}_{\bar{h}})} \quad (1)$$

Each individual cue in \mathbf{D} is defined by a modality, scale and offset. We employ three different modalities:

- 2-frame motion differencing - we calculate the absolute difference in intensity between subsequent frames at each pixel.
- background subtraction - we build an adaptive model on the background based on a mixture of Gaussians formulation and calculate the posterior probability that each pixel is foreground.



Figure 2: Conventional face/body detectors fail in the wide field image as pose variation and occlusion is too great, and many of the faces/bodies are very small. (a,b,c) show the largest (95×78 pixels), median (52×14) and smallest (15×2) bodies identified by hand in a training set (d,e,f) show the largest (34×31), median (5×4) and smallest (1×1) faces. Arrows in lower image depict other difficulties. From L to R, some people face away from the camera, others are partially occluded or are in unusual poses

- skin detection - we calculate a posterior probability that the pixel belongs to a skin region, based on non-parametric models of skin colour and background colour.

Each of these is initially derived independently for each pixel in the image. These cues are integrated over the a rectangular region representing the body (for motion detection and background subtraction) or head (for skin detection). The distribution of these cues when a face was present and absent was learnt from a training database of 2095 hand marked images. The probability of these cues under \mathcal{H}_h and $\mathcal{H}_{\bar{h}}$ was modelled as a mixture of three Gaussian distributions.

Example likelihood ratios from these distributions can be seen in Figure 3. No modality alone is sufficient. Motion detection fails when people are still, and the background subtraction method is adaptive, so eventually static people are incorporated into the background model. Skin detection fails when people turn away from the camera and for this scene there are large parts of the background that are approximately skin colored. However, the combination of all three modalities produces a likelihood ratio that considerably more robust than any of the three alone. For the example pictured, there are strong peaks in the joint likelihood ratio at 3 out of 4 of the actual head positions despite severe occlusion of all of these

individuals. We combine this likelihood with priors favoring likely locations in the scene to improve performance.

Our system works at 6 frames per second and we pass images from the foveal camera to a second machine where face detection takes place. We find that 93% of saccadic targets defined by the global maxima of the posterior were genuinely heads. This performance is further improved by incorporating attentive feedback. Scenes containing many people generate multiple extrema in the posterior map. We would like to prevent long fixations on a single individual, but we also want to ensure that the sensor fixates on an individual for sufficient time to gather enough high resolution data for biometric identification. We introduce a *fixation prior* which causes the sensor to dwell on an individual, tracking them as they move around the scene. We apply the method of [8] to the footage from the foveal camera. When a face is successfully detected, a feedback map is generated which temporarily inhibits this location and causes the sensor to move to a different position.

To evaluate this sensor, we selected 30-second segments of continuous data (at 6 fps) from two test videos. The number of people in the scene ranged from 2 to 20. There are several people who are present for less than 2 seconds. The system fixated 52 of the 63 people present (83% hit rate). Most of the errors occur for visibly difficult cases: individuals who are unusually low in the scene, largely occluded, extremely distant, or appear very briefly (1 frame). Example detected faces from the foveal video stream are indicated in Figure 3. For further details of this system refer to [5].

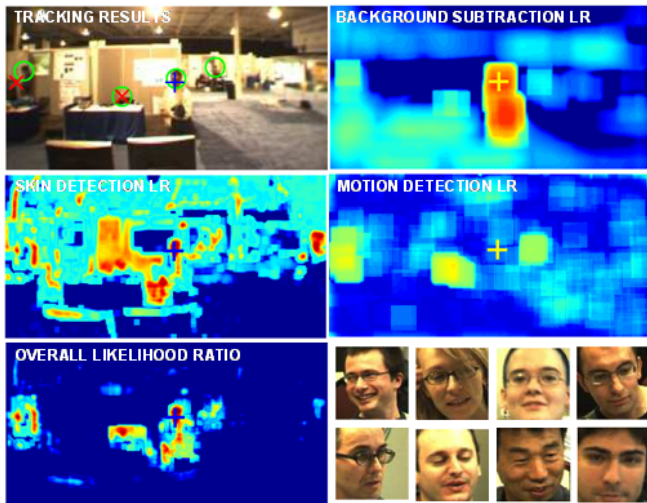


Figure 3: Results for face tracking. Crosses depict hypothesized faces (blue cross is the most probable), circles represent true face positions. Overall likelihood ratio at each pixel is determined by combining output from background subtraction, skin detection and motion detection subsystems. Local maxima in the likelihood ratio map are considered possible saccade locations. Bottom right shows examples of faces detected in the foveal camera.

3 Pre-processing Faces

In the previous section we have discussed how to gather high quality face footage over a wide area. In this section we discuss how to pre-process faces so that they are suitable for face recognition. A cascade-based face detector [8] is employed. This provides an approximate estimate of scale and position and the image is face by these estimates to a standard size. However, for accurate face recognition we need high quality registration. We identify distinct face features such as the eyes, nose etc. A warp is performed by triangulating the image based on these features, and mapping the vertices of the triangle to the standard shape. The color of each pixel at given barycentric coordinates remains the same after this warp (see Figure 4).

We aim to find 20 distinct points on the face. These include the corners of the eyes, tip of the nose, nostrils, corners of the mouth etc. (see Figure 4 top right). In profile faces where these features are occluded only the visible features are sought. Unfortunately, localizing these face features automatically is non-trivial - for example, local image data in the mouth region varies considerably depending on the individual, expression, presence of facial hair etc. For each feature we learn a local likelihood model based on a training dataset. Our measurements consist of the real and imaginary responses of 40 Gabor filters centred at that point. The probability of the local image data given a hypothesized position is an 80 dimensional Gaussian mixture model based on these responses, which we learn from training data with the EM-Algorithm [4]. We combine this with a Gaussian prior on each feature position that is also learnt from the training data. For a new face we estimate the maximum a-posteriori position.

Unfortunately, this local model is insufficient to always identify feature position accurately and some features are still misplaced. Hence, we also learn a Gaussian model of the covariance of the twenty feature positions. This embodies a priori knowledge such as the expectation that the eyes will not be found below the mouth. Ideally, we would like to represent the full covariance between each feature and every other. Unfortunately, for this fully-connected graphical probability representation, it is computationally intractable to find the optimal solution: if there are 20 features and 1000 possible pixel positions per feature then we must enumerate all 1000^{20} possible solutions exhaustively to find the most likely position.

We approximate the full covariance relations by a subgraph which takes the form of a tree (i.e. has no loops). Part of this tree is shown in the second panel of Figure 4 together with the search regions for seven of the nodes. To create the tree, we cut edges from the fully connected graph of covariance relations. We prune edges using mutual information: we aim to retain the tree with the maximum pairwise mutual information between nodes. We construct the maximum spanning tree using this criterion using Prim's algorithm [10]. Once the problem has been simplified to a tree, we can efficiently perform exact

inference using dynamic programming techniques. The system is extremely robust, and can cope with changes in expression, blinks and facial hair. We construct one of these feature detectors for each pose under investigation (here poses of -90° , -67.5° , -22.5° , 0° , 22.5° , 67.5° , 90° from horizontal).

Given the positions of the facial features and having warped the image to a standard shape, we now need to extract information suitable for performing face recognition. In particular, we would like to extract features that are likely to be correlated across different poses. We place a regular grid of 25 squares centered on each facial feature and extract the average gradient in this region in eight directions, and also the mean intensity level. This is repeated within each color channel, to yield a vector of length $25 \times (8 + 1) \times 3 = 675$ for each of the facial feature positions. For each face feature we project this 675×1 vector into a subspace of dimension 100 based on training data of features gathered *across all poses*. This retains variation that is common across poses.

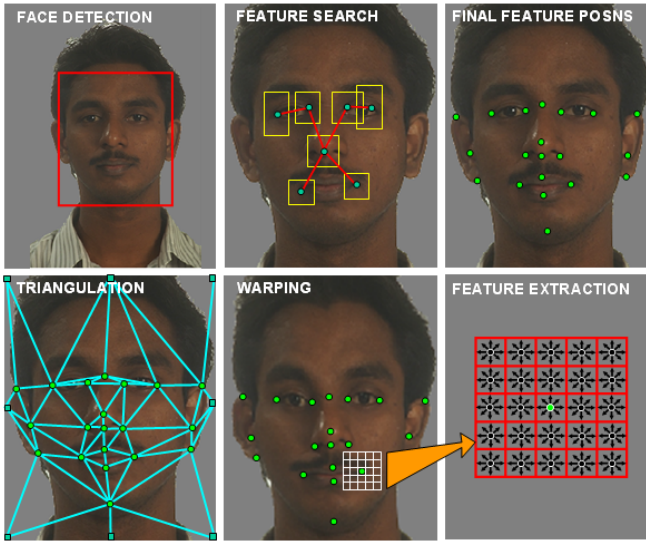


Figure 4: Preprocessing for faces. Face detection is performed using the method of [8]. The image was scaled by the estimated size from the detector. Image features (eyes etc.) are identified by dynamic programming on a tree. Each node of the tree consists of one feature, which may take several different positions (yellow boxes depict range). Branches of the tree represent constraints on the relative position of features. Twenty image feature positions are identified. The image is triangulated and warped to a standard shape. Blurred local gradient and mean intensity features are measured in 25 positions around each feature point.

4 Face Recognition across Pose

In this section we discuss the hard problem of how to establish reliable face recognition performance in the presence of severe pose changes. Conventional face recognition algorithms

perform very unreliably when the pose of the probe face is different from the stored face because typical feature vectors vary more with pose than with identity. This is also true of the feature vectors described in the previous section. The emphasis in previous approaches to this problem has been on creating a model which can predict how a given face will appear when viewed at different poses [7, 3]. Prince and Elder [11], presented a heuristic algorithm to construct a single feature *which does not vary with pose*. This seems a natural formulation for a recognition task and we have recently extended this to a full Bayesian probabilistic setting.

We propose a generative model that creates a one-to-many mapping from an idealized identity space \mathbf{c} to the observed data space \mathbf{x} (the image measurements). In the identity space, the representation for each individual does not vary with pose. We define the particular mapping from this space \mathbf{F}_k to the feature space is contingent on the discretized pose, k . In this way, the variation of the observed image features with pose is explained away by the generative model. In particular, we choose a linear transformation between the two spaces to give the forward model:

$$\mathbf{x}_k = \mathbf{F}_k \mathbf{c} + \mathbf{m}_k + \epsilon_k \quad (2)$$

where \mathbf{x}_k is the measured feature vector observed at pose k , \mathbf{c} is the constant position in identity space for this individual, \mathbf{F}_k is a pose-contingent linear transformation, \mathbf{m}_k is the mean observation vector for pose k and ϵ_k is Gaussian distributed noise, with diagonal covariance matrix Σ_k . We term this model “tied factor analysis” [12] since the choice of transformation (the factors, F) depends on pose but the loadings c are constant. The generative model is illustrated in Figure 5.

We learn the parameters, $\theta = \{\mathbf{F}_{1..K}, \mathbf{m}_{1..K}, \Sigma_{1..K}\}$ using training data where we know which face corresponds to which. We seek to increase the joint likelihood $Pr(\mathbf{x}, \mathbf{c} | \theta)$ of the measured image data \mathbf{x} and the invariant vectors, \mathbf{c} given these parameters. Unfortunately, we cannot observe the invariant vectors directly: we can only infer them, and this in turn requires the unknown parameters, θ . This type of chicken-and-egg problem is suited to the EM algorithm[4]. We iteratively maximize:

$$Q(\theta_t, \theta_{t-1}) = \sum_{i=1}^I \sum_{k=1}^K \int Pr(\mathbf{c}_i | \mathbf{x}_{i1..ik}, \theta_{t-1}) \log[Pr(\mathbf{x}_{ik} | \mathbf{c}_i, \theta_t) Pr(\mathbf{c}_i)] d\mathbf{c}_i$$

where t represents the iteration index. The first probability term on the r.h.s. is the probabilistic inversion of Equation 2, the second term is the generative model itself, and the prior on the identity space is defined to be Gaussian with mean zero and identity covariance. The term \mathbf{x}_{ik} represents training face data for individual i at pose ϕ_k . For notational convenience

we assume that we have one training face vector, \mathbf{x}_{ik} for each individual i at every pose ϕ_k . In practice this is not a necessary requirement: if data is missing (all individuals are not seen at all poses) these terms are simply dropped from the summation.

The EM algorithm alternately finds the expected values for the unknown pose-invariant vectors \mathbf{c} (the Expectation- or E-Step) and then maximizes the overall likelihood of data as a function of the parameters θ (the Maximization- or M-Step). More precisely, the E-Step calculates the expected values of the invariant vector \mathbf{c}_i for each individual i , using the data for that individual across all poses, $\mathbf{x}_{i1\dots iK}$. The M-Step optimizes the the values of the transformation parameters $\{\mathbf{F}_k, \mathbf{m}_k, \Sigma_k\}$ for each pose, k , using data for that pose across all individuals, $\mathbf{x}_{1k\dots Ik}$. These steps are repeated until convergence.

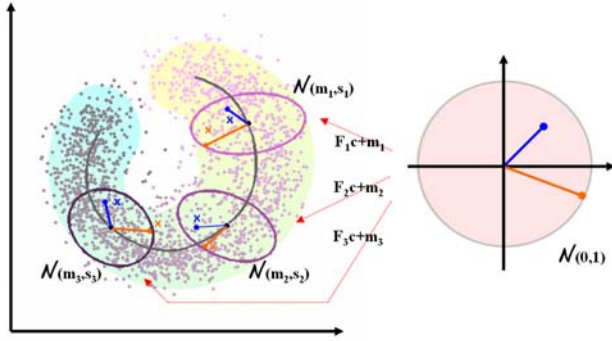


Figure 5: Observed features (left) vary with the pose. The orange crosses represent the same individual viewed at three different poses. The blue individuals represent a different person at the same three poses. The data is modelled as being generated from an ‘‘identity’’ space (right) where there is one vector for each individual. In particular, it is assumed that this vector has been transformed by a different linear transformation for each pose with additive Gaussian noise.

Having learnt these parameters, we need to use this generative model to perform face recognition. One obvious way to do this is to simply compare the identity vectors \mathbf{c} . For example we could simply take the nearest neighbor in this pose-invariant space. In practice, the identity vectors $\mathbf{c}_{1\dots N}$ for the N test faces and the vector \mathbf{c}_p for the probe face are uncertain. We can create an effective probabilistic metric by integrating out this uncertainty in a Bayesian fashion. Denoting the hypothesis that the probe face matches the n 'th test face by \mathcal{H}_n , we can write the likelihood as:

$$Pr(\mathbf{x}_1 \dots \mathbf{x}_N, \mathbf{x}_p | \mathcal{H}_n) = \int Pr(\mathbf{x}_{1\dots N}, \mathbf{x}_p | \mathbf{c}_{1\dots N}, \mathbf{c}_p = \mathbf{c}_n) d\mathbf{c}_{1\dots n} \quad (3)$$

where $\mathbf{x}_1 \dots \mathbf{x}_n$ represent the n observed image feature vectors in the test database and \mathbf{x}_p represents the measured feature for

the probe face. A posterior probability for the likely match can be calculated using Bayes' rule:

$$Pr(\mathcal{H}_n | \mathbf{x}_{1\dots N}, \mathbf{x}_p) = \frac{Pr(\mathbf{x}_1 \dots \mathbf{x}_N, \mathbf{x}_p | \mathcal{H}_n) Pr(\mathcal{H}_n)}{\sum_{i=1}^N Pr(\mathbf{x}_1 \dots \mathbf{x}_N, \mathbf{x}_p | \mathcal{H}_i) Pr(\mathcal{H}_i)} \quad (4)$$

This expression can be calculated in closed form since the integrated terms are all Gaussians. In this way we can calculate a posterior over the possible matches.

5 Face Recognition Results

We built binary pose models comparing frontal faces with one other pose. In each case we trained the system with 220 faces from the FERET dataset. We built a separate model each for 10 of the face features that were visible in both profile and frontal faces. These sub-models are assumed to be independent, and the likelihoods are multiplied together under each hypothesis. The test database consisted of 100 different frontal faces from the FERET dataset. On each trial a single probe face is presented at a different pose. We report the proportion of first choice correct matches to the database. Results are given in Figure 6. Performance at $\pm 22.5^\circ$ pose difference was perfect, but dropped to an average of 95% at $\pm 67.5^\circ$ and 86% at $\pm 90^\circ$.

Our results compare favorably with previous studies. Gross et al. [7] report 75% first match results over 100 test faces from a different subset of the FERET database with a mean difference in absolute pose of 30° , and a worst case difference of 60° . Our system gives 96% performance with a pose difference of 67.5° for every pair. Blanz et al. [3] report results for a test database of 87 subjects with a horizontal pose variation of either $\pm 45^\circ$ from the Face Recognition Vendor Test 2002 database. They investigate both full co-efficient based 3D recognition (84.5%, square symbols in Figure 6) and estimating the 3D model and creating a frontal image to compare to the test database (86.25% correct, circular symbols). Our system produces better performance over a larger pose difference in a larger database.

6 Discussion: The Challenges Ahead

In this paper we have sketched a series of solutions for building reliable face recognition solutions which do not require the co-operation of the subject. However, there are extensions required before we have a full useful system. First, in the examples in this paper, we have classified the pose of the face by hand, and this needs to be automated. Second, we have not dealt with lighting variation. The system described for creating invariance to pose variations could be adapted to eliminate variation due to illuminance. However, it is harder



Figure 6: Face recognition results comparing subset of FERET database containing 100 frontal faces to probe faces at different poses. Our system gives significantly better performance than the competing techniques of [7] and [3]

to bin luminance variation into discrete categories, and there is little data available to train across lighting variations. Ideally, an unsupervised technique would be used which establishes the effect of lighting variation without the need for supervised classification of training data. Furthermore, the amount of training data required becomes very large if we must make recognition jointly invariant to pose and lighting.

A crucial aspect of our face recognition system is that it produces a full posterior over the possible matches. This means that in a practical system data can be accumulated until there is a sufficient degree of certainty as to the identity of the individual. In the context of our attentive sensor, we can feed this information back to help drive the sensor to track individuals until their identity is established and then saccade to other people. A further aspect of the sensing stage that could be improved is to explicitly track the individuals in the low resolution footage once their identity has been established to prevent fixating them again at subsequent times unnecessarily.

Another future challenge is to establish the number of individuals who entered an area in a certain time period. This is a model selection problem: we would like to establish how many different people were in the scene and when. A strong advantage of our probabilistic decision metric is that we can meaningfully compare models with different numbers of parameters and address such questions. This is valid because we have integrated out the unknown identity parameters c , so explanations of the data with more identity parameters do not necessarily explain the data better. Finally, it should be noted that our face recognition scheme does not currently have any notion of “within-subject” variance (other than due to pose).

This is known to improve face recognition results [1].

References

- [1] P.N. Belhumeur, J. Hespanha and D.J. Kriegman, “Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection,” *PAMI*, Vol. 19, pp. 711-720, 1997.
- [2] V. Blanz, S. Romdhani and T. Vetter, “Face identification across different poses and illumination with a 3D morphable model,” *Int’l Conf. Face and Gesture Rec.* pp. 202-207, 2002.
- [3] V. Blanz, P. Grother, P. J. Phillips and T. Vetter, “Face Recognition Based on Frontal Views Generated from Non-Frontal Images,” in *Proc. CVPR*, pp. 454-461, 2005.
- [4] A. Dempster, N. Laird and D. Rubin, “Maximum likelihood from incomplete data via the EM algorithm,” *J. Roy. Statist. Soc. B*, Vol. 39, pp. 1-38, 1977.
- [5] J.H. Elder, S.J.D. Prince, Y. Hou, M. Sizinstev and E. Olevskiy, “Pre-Attentive and Attentive Detection of Humans in Wide-Field Scenes,” *IJCV*, in press.
- [6] M. Greiffenhagen, V. Ramesh, D. Comaniciu, and H. Niemann, “Statistical modeling and performance characterization of a real-time dual camera surveillance system,” in *Proc. CVPR*, pp. 335-342, 2000.
- [7] R. Gross, I. Matthews and S. Baker, “Appearance-Based Face Recognition and Light Fields,” *IEEE PAMI*, Vol. 26, pp. 449-465, 2004.
- [8] R. Lienhart and J. Maydt, “An extended set of haar-like features for rapid object detection,” in *Proc ICIP*, pp. 900-903, 2002.
- [9] L. Marchesotti, L. Marcenaro, and C. Regazzoni, “Dual camera system for face detection in unconstrained environments,” in *Proc. ICIP*, Vol. 1, pp. 681-684, 2003.
- [10] R. C. Prim. “Shortest connection networks and some generalizations,” *Bell System Technical Journal*, Vol. 36, pp. 1389-1401, 1957.
- [11] S.J.D. Prince and J. Elder, “Invariance to nuisance parameters in face recognition,” in *Proc. CVPR*, pp. 446-453, 2005.
- [12] S.J.D. Prince and J. Elder, “Tied factor analysis for face recognition across large pose changes,” *BMVC*, submitted.
- [13] S. Romdhani, V. Blanz and T. Vetter, “Face identification by fitting a 3D morphable model using linear shape and texture error functions,” in *Proc. ECCV*, 2002.
- [14] B. Scassellati, “Eye finding via face detection for a foveated active vision system” in *AAAI/IAAI*, pp. 969-976, 1998.
- [15] P. Viola and M.J. Jones, “Rapid object detection using a boosted cascade of simple features,” in *Proc. CVPR*, pp. 511-518, 2001.
- [16] P. Viola, M.J. Jones, and D. Snow, “Detecting pedestrians using patterns of motion and appearance,” in *Proc. ICCV*, pp. 734-741, 2003.
- [17] W. Zhao, R. Chellappa, A. Rosenfeld and J. Phillips, “Face Recognition: A literature Survey,” *ACM Computing Surveys*, Vol. 12, pp. 399-458, 2003.