# FACE DETECTION AND CLUSTERING FOR VIDEO INDEXING APPLICATIONS*

*Csaba Czirjek, Noel O'Connor, Sean Marlow and Noel Murphy*

{czirjekc, oconnorn, marlows, murphyn}@eeng.dcu.ie
Centre for Digital Video Processing
Dublin City University, Ireland

## ABSTRACT

This paper describes a method for automatically detecting human faces in generic video sequences. We employ an iterative algorithm in order to give a confidence measure for the presence or absence of faces within video shots. Skin colour filtering is carried out on a selected number of frames per video shot, followed by the application of shape and size heuristics. Finally, the remaining candidate regions are normalized and projected into an eigenspace, the reconstruction error being the measure of confidence for presence/absence of face. Following this, the confidence score for the entire video shot is calculated. In order to cluster extracted faces into a set of face classes, we employ an incremental procedure using a PCA-based dissimilarity measure in conjunction with spatio-temporal correlation. Experiments were carried out on a representative broadcast news test corpus.

## 1. INTRODUCTION

When searching for content in a large multimedia repository, browsing through linear segments of video is often inefficient and time consuming from the user's perspective. On the other hand, video queries tend to be very broad in scope, ranging from a general form such as e.g. locating an indoor or outdoor scene, to ones which address more specific targets e.g. locating the occurrence of a particular soccer player or a favorite actor/actress in a movie. Analysis tools that are able to locate the presence of particular objects or events can facilitate the task of associating a semantic meaning to video segments. Consequently, browsing and querying can be subsequently greatly improved by combining high-level semantic features with low-level audio-visual descriptors.

In this paper we focus on detecting and matching a particular object in generic scenes – the human face. Our detection method follows a coarse-to-fine approach, whereby in each step we impose particular constraints on possible candidates, in order to minimize false alarms and increase precision. The test corpus consists of broadcast news programs, as the presented face detection and matching tools are being developed in the context of a larger project to segment individual news stories.

The remainder of this paper is organized as follows: Section 2 describes our face detection strategy within continuous video shots. Section 3 covers our preliminary experiments on face clustering and the performance obtained. Our results for both approaches are presented in Section 4. Finally, conclusions and a summary of the paper are presented in Section 5.

## 2. FACE DETECTION IN COLOUR IMAGES

Given a video programme, the task is to recognize whether or not a face (or faces) occur within each shot. Many approaches have been proposed to locate faces in generic scenes, which use shape, color, texture, motion, edge and statistical analysis [1]. When working with digital video, face detection algorithms have to deal with a significant constraint – computation time. In general, 30 minutes of MPEG-1 video contains approximately 55,000 individual video frames, thus the algorithm applied to an individual frame should be able to classify it as face/non-face in the shortest time possible. To reduce the computation time, typically only every $N^{th}$ frame is selected for processing. Such sampling seems natural, considering the temporal redundancy between adjacent frames.

In our approach to face detection, we use colour information as the primary tool for locating face regions. We make use of this source of information based on our test content and also considering computational constraints. The processing chain involves a sequential pruning of candidate regions until all criteria are satisfied. In general, skin colour based face detection algorithms are computationally modest, but are

* The support of the Informatics Research Initiative of Enterprise Ireland is gratefully acknowledged. This material is based upon work supported by the IST programme of the EU in the project IST-2000-32795 SCHEMA.

sensitive to noise, lighting conditions and the colour content of the objects present in the scene. The face detection algorithm is illustrated in Figure 1 and described in the following sections.

### 2.1. Skin Tone Filtering

It has been shown that skin colour, under normal lighting conditions, falls into a relatively narrow band of the colourspace [2]. A study on the properties of human skin under different lighting conditions can be found in [3]. The range of skin possibly detected by a certain camera is used to enhace the detection rate. Many colour models have been used in algorithms to locate the presence of humans. In the literature, normalized RGB, YUV, HSV, CIEL, L*a*b*, XYZ etc. are used for this purpose. When using color information, each pixel in the image should be classified as skin or non-skin. In [4] the skin colour distribution in CIE LUV is modeled by a bivariate Gaussian, using only the chrominance information. Earlier research [5] proved that skin regions can be located in the $YC_bC_r$ colorspace using an appropriate set of thresholds. In general, if $(\mu_{Cb},\sigma_{Cb})$ and $(\mu_{Cr},\sigma_{Cr})$ are the parameters of the Gaussian distribution, then a pixel in the image will be classified as skin if it satisfies the following relations:

$$\left|p_{Cb}(x,y)-\mu_{Cb}\right| < k\sigma_{Cb} \quad (1)$$

$$\left|p_{Cr}(x,y)-\mu_{Cr}\right| < k\sigma_{Cr} \quad (2)$$

where $p_{Cb}(x,y)$ and $p_{Cr}(x,y)$ are the chrominance values of a pixel at coordinates (x,y) and $k$ a sub unitary constant. In our approach, we decided to detect skin-like pixels in YCbCr colorspace in order to avoid additional computations during colorspace transformations. A similar method has been used in the HSV colorspace [6]. Alternative methods make use of a skin probability map [7] instead of performing a binary classification of each pixel so that the weight corresponding to a pixel represents the likelihood of it being a skin sample. However, in order to create such a map a large labeled skin database is required [8].

### 2.2. Morphological Filtering

Because of the nature of the classification used, the output of skin tone filtering – a skin-mask – will be populated with many isolated pixels. To eliminate this undesirable effect, we apply a morphological opening and closing with a square kernel of $N \times N$ (experimental results have indicated a suitable value of $N=5$). After filtering, we obtain smoothed homogeneous areas of connected pixels. Connected component labeling is then performed, which gives the number of regions used in the next stages of the algorithm.

### 2.3. Skin Region Processing

Even after applying morphological filtering to the skin-map, regions with a small number of pixels may be present. In order to reduce the number of false candidate regions, areas less than 625 pixels are ignored. We have chosen this threshold based on the assumption that a face with size smaller than $25 \times 25$ pixels should not be detected by our approach. Horizontal and vertical strips, which are less likely to contain a human face, are also ignored. These regions are detected by identifying a significant difference between the width and height of the region's bounding box, with the condition that the smaller dimension does not exceed 25 pixels.

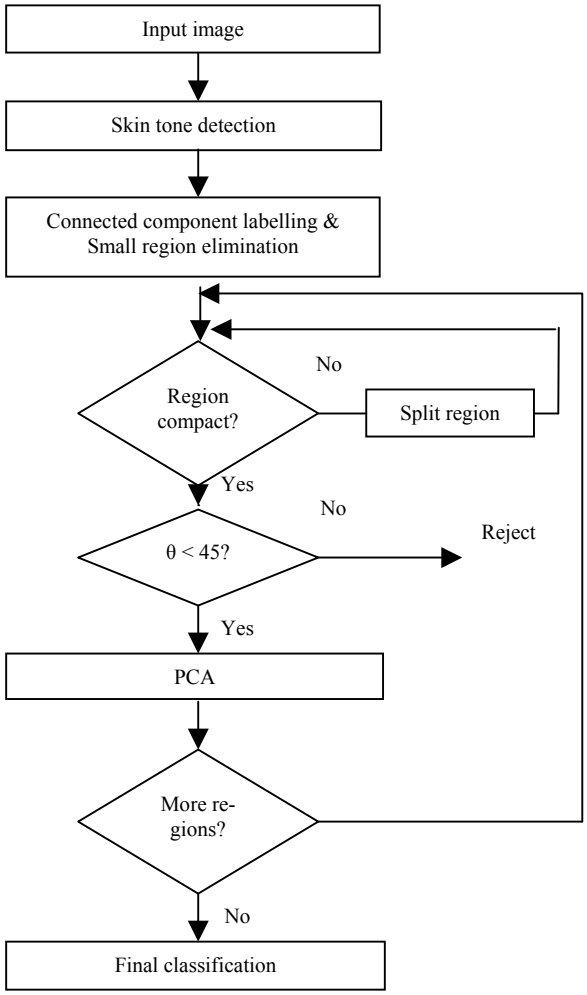


Figure 1: The face detection algorithm

It is possible that other objects in a visual scene have similar colour characteristics to human skin, or that other objects are merged with the face (e.g. hands,

background, etc). In general, when dealing with skin-colour based face detection, the following scenarios can occur:

1. a face candidate region is separated (not connected to other objects).
2. a face region is merged with other objects due to their similar colour characteristics.
3. a candidate region is a false alarm.
4. a candidate region is split into unconnected regions which are not valid for detection by themselves, as they should be part of the same entity. This situation usually occurs when there is a close-up of a face in the scene.
5. a candidate region is not detected due to insufficient colour information (black/white movies or poor lighting conditions).

In our approach candidates belonging to scenarios 1 and 3 are handled by the principal component analysis module outlined below. We do not address scenarios 4 or 5 in this work. In order to address scenario 2, where the face is merged with other parts of the scene, we perform an iterative splitting procedure on the connected component. To this end, we introduce a measure of region compactness as the ratio between the area of the connected component and its bounding box. This value indicates the degree of convexity of a region.

The compactness of a candidate region *S* signals if it should be partitioned into sub-regions or not. If the ratio falls below a threshold, k disjoint sub-regions are formed by maximizing the compactness of each subsequent sub-region[1]:

$$S_i, i = 1..k \;,\; \bigcup_i^k S_k = S \tag{3}$$

It is known that the aspect of a "perfect" face is close to the "golden ratio" [9]. Therefore, we divide regions which deviate from this value into sub-regions so that the ratio between height and width approaches the ideal whilst the maximum compactness constraint is obeyed. If the width of the bounding box is greater than the height the splitting procedure operates from top to bottom, otherwise it propagates horizontally. An illustrative example is a head-and-shoulder image, commonly found in news anchorperson shots, where the head is merged with the clothes due to similar color (e.g. a yellow jacket). In this case, region partitioning will segregate the head area of the body.

Assuming that the human face has an elliptical shape, for each compact region the best-fit ellipse is calculated based on moments [10]. The orientation angle of the ellipse is given by:

$$\theta = \frac{1}{2}\arctan\left(\frac{2\mu_{1,1}}{\mu_{2,0} - \mu_{0,2}}\right) \tag{4}$$

where $\mu_{i,j}$ denotes the central moments of the connected component. If the orientation of the major axis falls in the range $\theta \in \left[-45^o, 45^o\right]$ the region is selected for the classification stage, otherwise it is rejected. Selected regions are extracted, reverse tilted if $\theta > 0$, rescaled to *64x64* pixels, histogram equalized and passed to the principal component analysis module for final classification.

### 2.4. Principal Component Analysis

Using a collection of test images, we construct a face space for discriminating the remaining candidate regions. Given a set of vectors {x}, where each vector is an image with rows in lexicographic ordering, the basis vectors can be computed by solving the eigenvalue problem:

$$\Lambda = P^T \Sigma P \tag{5}$$

where Σ is the covariance matrix of {x} and P the eigenvectors matrix. The extracted regions are normalized, rescaled and then arranged into a vector x. The principal component vector is obtained by projecting x into the face space spanned by the eigenfaces:

$$y = P_M^T (x - \bar{x}) \tag{6}$$

where $\bar{x}$ is the mean image of the training set {x}.

The measure of "faceness" of the input sample relies on the reconstruction error, expressed as the difference between the input image and its reconstruction using only *N* eigenvectors corresponding to the highest eigenvalues [11][12]. This is termed the distance from face space (DFFS):

$$\varepsilon^2 = \| x - \bar{x} \|^2 \tag{7}$$

The distance between the projected input image and the mean face image in the feature space is given by the norm of the principal component vector. Since the variance of a principal component vector $y_i$ is given by its associated eigenvalue $\lambda_i$, the squared Mahalanobis distance measure $d^2$ gives a measure of the difference between the projection of the test image and the mean face image of the training set {x}:

$$d^2 = \sum_{i=1}^{M} \frac{y_i}{\lambda_i} \tag{8}$$

[1] Hole-filling is carried out for each candidate region, prior to checking its compactness.

where $y_i$ are the projection coefficients and $\lambda_i$ are the associated eigenvalues. Therefore $d^2$ can be expressed as the distance in face space (DIFS).

Given these two distances a combined error criterion is calculated:

$$e = d^2 + c\varepsilon^2 \quad (9)$$

with c chosen to be a suitable constant value. Because PCA is sensitive to scale and rotation, a Gaussian multiresolution pyramid could be used to reduce the number of missed positive candidates, to the detriment of computation time.

### 2.5. Extension To Video Shots

When dealing with video shots, the above algorithm processes every 12th frame in the shot (every I-picture from the MPEG-1 bitstream). A confidence measure for each shot for the presence/absence of face is calculated as the average confidence value of each processed frame. The operation of the entire algorithm with the resulting face candidates is depicted in Figure 2.


Figure 2: Illustration of face detection in operation

## 3. FACE SEQUENCE CLUSTERING

Given a set of faces extracted from video frames, our next experiment addressed the problem of combining these individual face samples into a number of disjoint clusters.

The topic of incremental face sequence matching has been addressed in the literature [13-16]. In [13] a semi-automatic annotation system is first employed for training. For this, face sequences are extracted and aggregated using the underlying structure of drama video sequences. The face sets are then clustered using an eigenface-based method. The performances of different face sequence matching techniques are evaluated on actual drama content in [14]. A novel method for incremental learning is described in [16] where unsupervised recognition is accomplished by constructing a graph in which similar views are chained in the image space depending on local similarity measures.

Our clustering process is based on the individual PCA approach. In this approach, an eigenspace is constructed for each subject in the training set. In contrast to universal PCA where the space represents inter-variations of subjects and also intra-variations across different views of the same subject, in individual PCA each subject has a characteristic eigenspace. Using the dissimilarity measures to each individual eigenspace correlated with spatio-temporal information, we attempt to identify new face sequences.

The processing chain starts by constructing N individual eigenspaces for extracted and manually classified faces corresponding to specific anchorpersons commonly observed in news programs. These face spaces will play the role of reference or primary databases when comparing new acquired candidates. Within the news programmes in our test corpus, only a small number of newscasters usually appear so we considered it useful to form databases of these characters. Each extracted face candidate is a 64 × 64 grayscale image thus each data sample maps to a point in a 64 × 64 dimensional space. For each considered anchorperson the PCA algorithm described in section 2.4 is applied resulting in N eigenspaces. Given these initial eigenspaces, our approach proceeds as outlined below.

### 3.1. Face Sequence Extraction

For each continuous video shot a sequence of faces corresponding to the same person are extracted. In this context, an important factor is the location of the extracted candidate in the frame. Generally in news video, the person's position doesn't change drastically during an anchorperson shot or interview. We use a simple tracking method between adjacent I-frames, which examines the overlapping of the bounding boxes of the extracted faces. If the faces occur almost at the same position and more than 30% of the area overlaps they are considered to belong to the same person.

Potential face sequences are likely to exhibit similar dissimilarity values to the reference databases. If the number of candidates in a shot is higher than 15 we analyze the DFFS dissimilarity curve in order to establish the occurrence of a new face sequence. We have chosen this lower limit based on the assumption that an interview or any other significant part of a news program should not be shorter than 6 seconds.

### 3.2. New Face Sequence Classification

If the variance of the dissimilarity measures across a shot satisfies our constraints, these samples are regarded as a new sequence and they form a new face (character) database upon which PCA is performed. Since determining the eigenvectors is computationally intensive, the power method [17] is used to efficiently calculate the dominant eigenvectors. Typical dissimilarity curves are illustrated in Figure 3.

From Figure 3 it can be observed that between shot boundary changes the dissimilarity to each database remains relatively constant (discounting noise). For an anchorperson shot from the reference databases the DFFS to the correct face class exhibits an accentuated variation, whereas this is not the case for the other eigenspaces. This reflects the changes in expression and pose of the individual in that class.

If the dissimilarity measures fall within suitable thresholds for the duration of the shot and the number of samples exceeds a minimal value, then a face cluster is established. Mathematically, the condition is:

$$\frac{1}{M-1}\sum_{i=1}^{M-1}\left|\delta_{i,j}-\delta_{i+1,j}\right| < T_j \qquad (10)$$

where $\delta_{i,j}$ represents the distance of sample *i* to eigenspace *j*, M denotes the number of samples across the shot and $T_j$ a threshold for database *j*.

## 4. RESULTS

The results of the face detection algorithm on a number of selected Irish news broadcasts from the Físchlár [18] news archive are summarized in Table 1, whereas the face clustering results are presented in Table 2.

In calculating the ground truth used to evaluate the results obtained for our face detection algorithm we manually inspected each shot in the test corpus and recorded the number of faces present. The results presented in Table 1 are very encouraging with an average precision value of 71.50%.

In calculating the ground truth used to evaluate the results obtained in our face clustering experiments, we considered a "real" face sequence as a shot where a person appears for at least 6 seconds without significant occlusions. This scenario is commonly found in outdoor reporting/studio interviews. From Table 2, it can be seen that the number of face sequences identified is higher than the actual sequences appearing during the broadcast. The reason for this is that consistently false sequences are also detected as real face classes because of successful tracking and low PCA error residual. It should be noted however, that the misclassification percentage is quite low. The number of newly formed face classes is not high considering that new sporadic faces are common in news broadcasts which typically contain a mixture of indoor (studio) and outdoor (outside broadcast) segments.

From a computational point of view, for a 30 minutes news programme the face detection algorithm takes approx. 20 minutes (60% real time processing) on a Dual Pentium III 700 MHz running Red Hat Linux 7.3, whereas the face clustering algorithm is much less computationally expensive, requiring only a few minutes execution time. A screenshot of the GUI used for face clustering is presented in Figure 4.

| News ID | Extracted candidates | False alarms | Precision (%) |
|---|---|---|---|
| 06 Sept. 02 | 2254 | 386 | 82 |
| 12 Sept. 02 | 2501 | 523 | 79 |
| 15 Sept. 02 | 1611 | 473 | 70 |
| 17 Sept. 02 | 1574 | 510 | 67 |
| 19 Sept. 02 | 1192 | 385 | 67 |
| 22 Sept. 02 | 1303 | 489 | 62 |
| 23 Sept. 02 | 1408 | 412 | 70 |
| 24 Sept. 02 | 2715 | 677 | 75 |

Table 1: Face detection results

| News Id | No. candidates | No. face sequences | | |
|---|---|---|---|---|
| | | Real sequnces | Detected sequences | Misclassified faces (in sequences) |
| 06 Sept. | 2254 | 18 | 27 | 17 |
| 12 Sept. | 2501 | 21 | 35 | 21 |
| 15 Sept. | 1161 | 20 | 28 | 15 |
| 17 Sept. | 1574 | 17 | 32 | 14 |
| 19 Sept. | 1192 | 26 | 30 | 19 |
| 22 Sept. | 1303 | 23 | 25 | 11 |
| 23 Sept. | 1408 | 15 | 28 | 23 |
| 24 Sept. | 2715 | 28 | 37 | 12 |

Table 2: Face clustering results

## 5. CONCLUSIONS AND FUTURE WORK

In this paper, we presented an iterative colour-based face detection and clustering method for video sequences. The clustering process facilitates automatically detecting new anchor presenters not in the reference

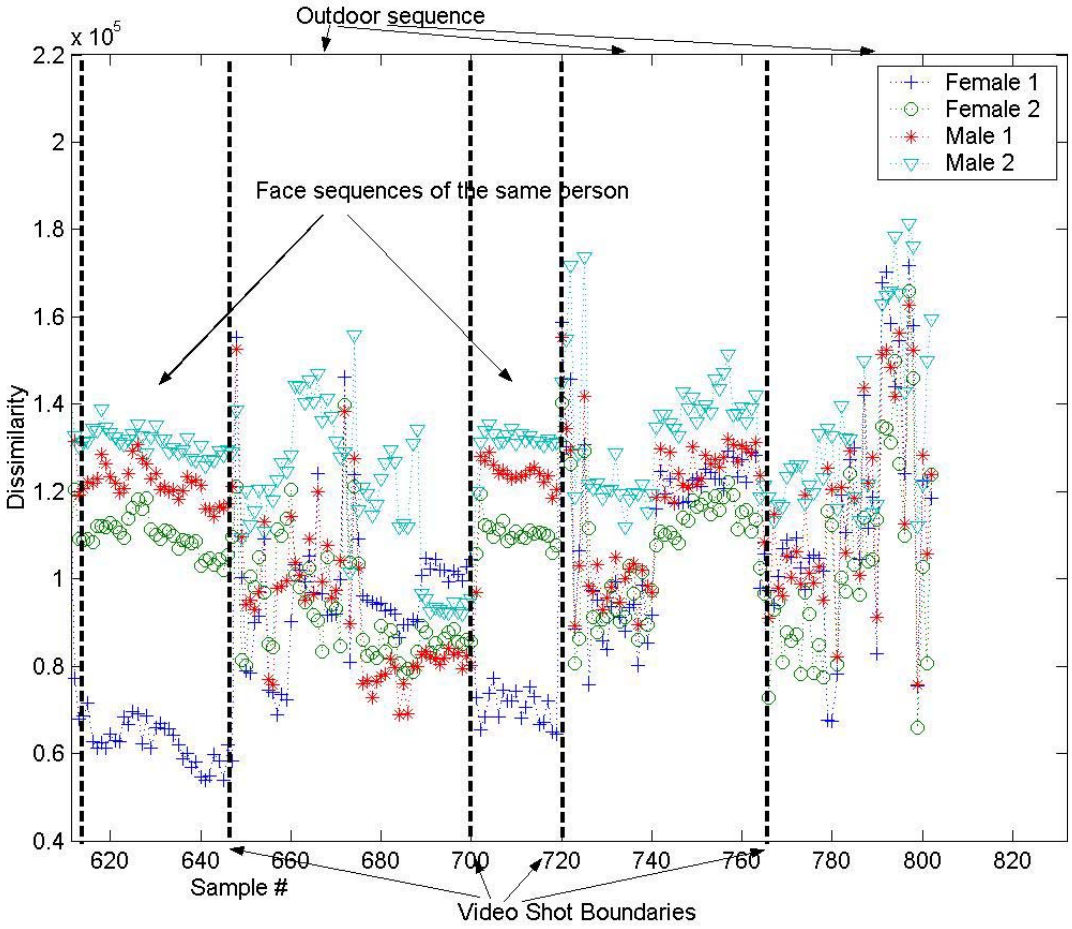


Figure 3: Face dissimilarity curves relative to four face databases across a number of shots

face databases, but also enables the detection of interview or report segments (e.g. coverage of parliamentary proceedings in our application). This provides higher level information to multimodal indexing tools by characterizing the content in terms of human presence and reoccurrence, extending the capabilities of our existing news story segmentation scheme [19].

The face detection technique could be significantly improved using facial feature extraction for face verification, and outliers could be eliminated exploiting the temporal redundancy within consequent video frames using techniques such as Kalman filtering. We are aware that out approach cannot deal with significant changes in ambient settings due to the fact that the underlying PCA is sensitive to scale, rotation and changing lighting conditions.

Our preliminary experiments on face clustering, are encouraging and we intend to investigate the possibility of improving our classification by modelling the distribution of face sequences in the face space and by introducing unsupervised clustering techniques such as Kohonen maps.

## 6. REFERENCES

[1] M-H. Yang, D.J. Kriegman, N. Ahuja . "Detecting Faces in Images: A Survey", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(1), pp. 34-58, 2002.

[2] A. Albiol, C.A. Bouman, E. J. Delp, "Face Detection for Pseudo-Semantic Labelling in Video Databases", *IEEE International Conference on Image Processing*, vol. 3, pp. 607-611, Oct. 1999.

[3] M. Soriano, B. Martinkauppi, S. Huovinen, "Skin Detection in Video Under Changing Illumination Conditions", *IEEE International Conference on Pattern Recognition,* pp. 839-842, 2000.

[4] M.-H. Yang and N. Ahuja, "Detecting Human Faces in Colour Images" *Proc. IEEE International Conference on Image Processing*, vol. 1, pp. 127-239, Oct. 1998.

[5] N. Herodotu, K. N. Plantaniotis, A. N. Venetsanopoulos, "Automatic Location and Tracking of the Facial Region in Colour Video Sequences", *Signal Processing: Image Communication*, 14 , pp. 359-388, 1998.

[6] K. Sobottka, I. Pittas, "A Novel Method for Automatic Face Segmentation, Facial Feature Extraction and Tracking", *Signal Processing: Image Communication,* 12*,* pp.263-281, 1998.

[7] J.D. Brand, J.S.D. Mason, M. Roach, "A Comparative Assesment of Three Approaches to Pixel-level Human Skin Detection", *Proc. International Conference on Pattern Recognition,* vol.1, pp. 1056-1059, 2000.

[8] M. J. Jones, J.M. Rehg, "Statistical Color Models with Application to Skin Detection", Compaq, Cambridge Research Laboratory, Technical Report, CRL-98-11.

[9] L.G. Frakas, I.R. Munro, "Anthropometric Facial Proportions in Medicine", Thomas Books, Springfield, IL, 1987.

[10] A.K. Jain, "Fundamentals of digital image processing", Prentice-Hall, NJ, 1989.

[11] ] M. Turk, A. Pentland, "Eigenfaces for Recognition", *Journal of Cognitive Neuroscience* (3), pp. 71-86, 1991.

[12] M.Moghaddam, A. Pentland, "Probabilistic Visual Learning for Object Representation*", IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7), pp. 696-710, 1997.

[13] S. Satoh, "Towards Actor/Actress Identification in Drama Videos", *Proc. ACM Multimedia*, 1999.

[14] S.Satoh and N. Katayama, "Comparative Evaluation of Face Sequence Matching for Content-based Video Access", *Proc. International Conference on Automatic Face and Gesture Recognition*, pp. 163-168, 2000.

[15] A. Fitzgibbon, A. Zisserman, "On Affine Invariant Clustering and Automatic Cast Listing in Movies", *Proc. European Conference on Computer Vision,* vol. 3, pp. 304-320, 2002.

[16] B. Raytchev, H. Murase, "Unsupervised Face Recognition by Associative Chaining", *Pattern Recognition 36*, pp.245-257, 2003.

[17] R.A. Horn, C.R. Johnson, "Matrix Analysis", Cambridge University Press, 1985.

[18] A. Smeaton, N. Murphy,N. O'Connor, S. Marlow, H. Lee, K. Mc Donald, P. Browne and J.Ye, "The Físchlár Digital Video System: A Digital Library of Broadcast TV Programmes", JCDL 2001 - ACM+IEEE Joint Conference on Digital Libraries, Roanoke, VA , 24-28 June 2001.

[19] N. O'Connor, C. Czirjek, S. Deasy, S. Marlow, N. Murphy, A. Smeaton, "News Story Segmentation in the Fischlar Video Indexing System", *Proc.IEEE International Conference on Image Processing*, vol. 3, pp. 418-421, October 2001.

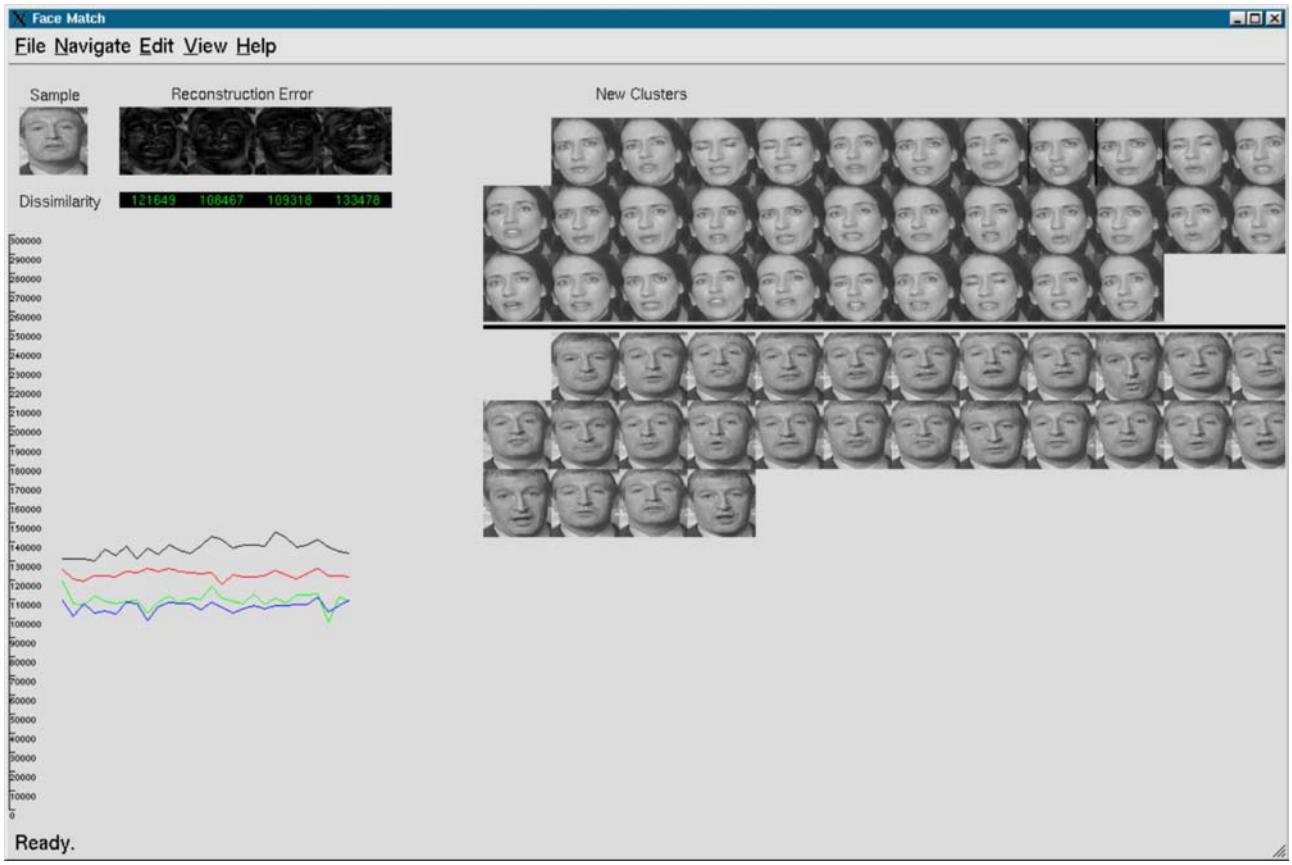


Figure 4: Face matching GUI