Global fitting of a facial model to facial features for model–based video coding

P M Hillman  J M Hannah  P M Grant
University of Edinburgh  School of Engineering and Electronics
Sanderson Building, King’s Buildings, Mayfield Road, Edinburgh EH9 3JL, UK
John.Hannah@ed.ac.uk

Abstract

Finding faces, locating important facial features and model fitting are required in a range of applications including model based video coding. In this paper we present a technique for locating important facial features in head and shoulders colour images. A fast and effective method of detecting skin, while excluding eyes, using PCA transformed CIE-Lab colour space, is described. Candidate eye regions are then processed using region growing and eigenimages using the distance from feature space approach. False matches are eliminated using facial geometry constraints. Nose and mouth features are also located using eigen images in search regions derived from the eye positions. Further facial characteristics are identified and automated fitting of a global head model is achieved using a genetic algorithm based approach. Results from applying our technique to a large database of facial images (XM2VTS) are presented.

1 Introduction

Detecting faces and locating their important facial features is a well researched field [1]. Our principal interest is to locate a face in a head and shoulders video sequence and subsequently fit a face model for use in a model based or hybrid video coder. In our application, largely frontal views of a single face are likely to be most common but idealised pose cannot be assumed.

2 Our Approach

Our approach to model fitting is outlined in Fig. 1. We use important facial features (eyes, mouth, nose and jaw position) to approximately fit the model to the face. The first step is an analysis of the colour input image to produce a segmentation into probable skin regions based on pixel colour alone. This will produce a mask which has 'holes' which correspond to most facial features, such as nostrils, eyes and eyebrows. These holes are all treated as eye region candidates. The next step is to analyse the skin segmentation to find these candidate regions using a modified max-tree algorithm. Principal Components Analysis is then used to search the candidate eye regions for eye matches. The final step performs geometrical analysis to find the best match for the eye positions based on the results of the PCA template matching and subject to geometrical constraints. Given the position of the eyes, the area likely to contain the nose and mouth is now searched using PCA template matching, again subject to geometrical constraints. The position of the mouth can be used to estimate the approximate position of the jaw. A Genetic Algorithm is then applied to find the optimal transform of the model in order to fit to the located facial features.

2.1 Skin colour segmentation

Many approaches to detecting skin in colour images have been proposed [1], including thresholding in colour spaces such as normalised RGB or HSI or using Gaussian Mixture Models (GMMs). For example, Mensur and Müller [2] use a GMM to produce a skin tone probability map. Park et al [3] threshold the GMM probability to produce a crisp segmentation. Lee and Yoo [4] use an elliptical boundary as a threshold for segmentation. However, our experiments suggest that these techniques often classify eyes as being skin and can be computationally intensive.

Our skin detector classifies pixels as being skin coloured if they lie within a bounding box in CIE-Lab colour space. We selected CIE-Lab space, after testing a number of alternative colour spaces (including normalised ones), as it achieved the best results. The limits of the bounding box were derived automatically by taking facial skin samples from a ‘training set’ of 70 images with varied skin tones and lighting conditions. These images were taken from a wide range of sources. The distribution of the pixels sampled from these images is shown in Fig. 2. This distribution can be seen to be skewed with respect to the axes, thus the conventional rectangular bounding box does not fit the distribution well.

Principal Components Analysis was therefore used to derive a new linear space, by finding the required transformation matrix $M$. For our training set this was found to be:

$$M = \begin{bmatrix} 0.836141 & 0.307689 & 0.454089 \\ -0.376524 & -0.280037 & 0.883068 \\ -0.398871 & 0.909344 & 0.118298 \end{bmatrix}$$ (1)
Colour face image

Skin detection

Candidate Eye Regions

PCA template finding

Geometrical analysis

Co-ordinates of eyes

PCA template finding

Geometrical analysis

Co-ordinates of mouth and nose

Jaw position estimation

Genetic fitting algorithm

fitted face model

Figure 1. Overview of our system

The limits \((p_{\text{min}}, p_{\text{max}})\) of the training set in this transformed space are used as the bounding box for thresholding.

The colour input image is segmented as follows: Each pixel \(p\) is processed in turn, converted into CIE-Lab space, and then into PCA space to find \(p' = Mp\). If the pixel falls within the limits \((p_{\text{min}}, p_{\text{max}})\), it is marked as skin (black), otherwise as non-skin (white). For our training set:

\[
p_{\text{min}} = [5.2462, 15.4459, 31.3839] \quad (2)
\]

\[
p_{\text{max}} = [77.7485, 63.4912, 57.1265] \quad (3)
\]

2.2 Candidate image finding

The output of the skin detector is a binary skin image which is black where there is skin and white in all other places. The eyes as well as other features will appear as white holes within the black mask of the face. This image is now processed using region growing to produce a region image. All pixels in this region image are initially unset. Regions are grown from seed pixels iteratively. Initially, the top left pixel is used a seed. In subsequent passes, the region image is scanned for the first unset pixel. A region is grown from each seed, by adding all pixels in the skin image which are the same classification as the seed, and neighbour either the seed or a previously included pixel. Once all pixels in the region image have been assigned, the regions are sorted in order of size.

Regions which may correspond to eyes are then found using a modified max–tree approach, similar to that presented by Salembier et al [5]. A max–tree is built using this region image and small regions are merged into their parent nodes to remove noise. Eye candidate regions are those which are children of the largest region of skin in the skin image, as this region is assumed to contain the face. All pixels belonging to such regions are output in an eye candidate image. Pixels in this image are either zero (indicating that the pixel is not an eye candidate) or else are assigned a value equal to their corresponding region number.

2.3 PCA eye template matching

Candidate eye regions are examined for eye–like images using Turk and Pentland’s Distance From Feature Space approach [6]. The reference set was derived from 600 left and 600 right eye images, of size \(60 \times 30\) pixels, extracted from the XM2VTS Database [7]. A separate mean eye and PCA transform is derived for each eye. Only the first 25 eigenimages are used to build the PCA space.

The eye candidate image derived from the unknown input image is scanned for non–zero pixels \(p\). For each pixel \(p\), a score for each eye is derived by subtracting the mean eye from the \(60 \times 30\) block centred around the pixel. The resultant image is transformed into PCA space and the absolute magnitude \(m\) of the resultant vector is taken. The smaller this number, the more similar the area around pixel \(p\) is to an eye.
The PCA stage outputs for each eye the position of the best match and the corresponding value of $m$ for each region in the eye candidate image.

2.4 Geometrical Analysis

The position of the left and right eyes could be taken to be the position for which the best PCA template match is achieved in each case. However, the probability of a false match, especially if there is significant orientation or scale difference with respect to the mean eye, is too high. Facial geometry information is therefore used to discard ‘impossible’ eye pairs. The output lists from the template matching step are scanned and each left eye/right eye pair is examined in turn. Any pair which fails the following geometrical constraints is rejected:

- The magnitude of the angle of the line joining the left and right eyes must be less than 30°.
- The distance between the regions must be between 10 and 20% of the image width.
- Both eyes must lie within the central 50% of the image.

The unrejected pair with the smallest combined $m$ is taken as the position of the eyes.

3 Mouth and Nose template matching

3.1 Search area

To speed up the template matching for the nose and mouth, only a small area area of the image is searched. The nose and mouth are assumed to lie near the perpendicular bisector of the line between the eye matches, as shown in Fig. 4. This assumption is only broken if the head is looking too far to the left or right. The region is $0.7w$ wide and $1.5w$ high, where $w$ is the width between the eyes. These factors were found by analysing the eye, nose and mouth positions in 80 images from the XM2VTSDB database.

3.2 PCA template searching

The search area is searched for nose and mouth matches in the same way as the eye candidate regions. Mean nose and mouth images were generated from 127 images, and 25 vector eigenspaces calculated. The output from this stage is a list of the position and match score ($m$) of the best nose and mouth fit for each horizontal row of the search area.

3.3 Geometrical constraints

As for the eyes, geometrical constraints are used to eliminate false matches. Let $d_m$ and $d_n$ be the distances between the eyes (marked $e$ in Fig. 4) and the position of the mouth and nose fits respectively. If the ratio $d_m/d_n$ is in the range $(1.3, 2.6]$ the geometrical constraint is satisfied. All matches output by the PCA template searching that satisfy the constraint are scanned, and the pair with the lowest combined score is taken to be the position of the mouth and nose.

4 Jaw finding

The final information required by the model fitting stage of our process is some indication of the width of the head. We have chosen to use the position of the jaws. The position of the jaw is taken as the position of the non-face pixels nearest to the mouth on each side of the face which lie on a line which is parallel to the axis of the eyes and passes through the mouth.

5 Model–fitting

We have used Candide-3 as our face model [8]. The model is fitted by finding model parameters that minimise the difference between the located feature points and the position of the corresponding points in the facial model. The following parameters are used:
- Global scale
- rotation about x-axis
- rotation about y-axis
- rotation about z-axis
- displacement on x-axis
- displacement on y-axis
- eye separation distance (shape unit 5)
- nose vertical position (shape unit 8)
- mouth vertical position (shape unit 10)

Our automatic fitting also requires estimation of the position of the top of the head or the chin, which will allow the optimisation of the head height (shape unit 0). Currently this parameter is estimated and entered manually.

Optimisation of these parameters is achieved with a Genetic Algorithm. The 8 parameters are encoded into a chromosome. Random mutation and crossover are employed to find a set of parameters which minimise the summed error between the position of each feature in the model when orthogonally projected onto the image, and the located position of the feature. Since not all features match perfectly, a weighting system is employed. For example, the horizontal position of the mouth is often less accurate, so the x coordinate is weighted less heavily than the more accurate y co-ordinate. The eyes are also weighted more heavily than the nose, as these are normally a more reliable match.

6 Results

We have mainly tested our system on the XM2VTS database [7]. This is a database of head and shoulder colour images (720 × 576) of 295 subjects, containing several images of each, taken over a period of time. The range of subjects and their characteristics (89% white, 46% female, 35% with glasses, 13% with facial hair) provides a realistic test.

6.1 Eye matching

Fig. 5 shows the results for one of the test images used (Fig. 5(a)). Fig. 5(b) shows the output of the skin detector. Some parts of the hair have also been identified as skin, but most of the eyes have been detected as non-skin. Fig. 5(c) shows the eye candidate regions. Three pairs of matches passed the geometry test; the one with the lowest combined template match is marked in Fig. 5(d).

The eye-finding system was tested on 1080 images from the database. None of the images used to form the PCA eye space or the mean eye template were used in testing. Also, the original colour training space for skin detection was not derived from this database. The testing set did include different images of some of the same subjects that were used

Figure 5. Results for image 001.4.2 of the XM2VTS Database
in eigenimage template building but we do not believe that this significantly affects the results.

Overall, there were 952 correct matches (mean error in this case was less than 10% of eye width), giving a success rate of 88%. There were 1009 matches to the eye region, giving a success rate of 94%. Most of the failures occurred where there were severe reflections in glasses, where the subject was looking away from the camera, or where hair was occluding the eyes. Skin colour and the presence of facial hair did not significantly affect performance.

Since the XM2VTS database contains images with uniform backgrounds, the system was also tested on a variety of other images, with a success rate similar to that of the XM2VTS database. Only images which had extensive areas of skin-tones in the background, or which had very poor illumination, proved challenging.

In order to test the scale-sensitivity of the system, it was tested with scaled versions of the database. A variation in scale of up to ± 25% was found to have very little effect on the overall performance, which is more than sufficient for our intended application.

6.2 Nose and mouth matching

Fig. 6 shows a collection of matches to mouth and nose. We tested the performance of nose and mouth matching on 114 subjects which were not used in the formation of the nose and mouth mean images. The eye region was not correctly located in 5 of these images. Of the remaining 109 images, the nose was correctly located in every case, and both nose and mouth were located in 94% of the images. The seven failures were on subjects who had beards or moustaches. Employing an elliptical template allowed the correct location of five of the seven subjects, but decreased overall performance to 83%.

In many cases the mouth fit is not in the centre of the mouth area, but this rarely affects the model fitting step. Fig. 6(b) shows a case where the mouth and nose are fitted well, even though an eyebrow has been matched instead of an eye.

6.3 Model fitting

Fig. 7 shows examples of results of automated model fitting. In most cases, the automatic model fit is reasonably close to a manual fit. Occasionally the mouth or nose fit is too far displaced to one side, causing the model to rotate significantly. The presence of hair or large ears can cause the jaw finding to fail, resulting in the head size of the model being too large.

7 Conclusions and future work

This paper has presented a new system for automated fitting of models to facial images. Our CIE-Lab/PCA skin detector is capable of distinguishing between skin and eyes under normal lighting conditions and the template based

Figure 6. Results of feature location on images in the XM2VTS database
The feature finder is successful at locating the eyes, noses and mouths in most images in order to achieve model-based coding.

Future work will involve techniques for location of the head and chin to allow automatic finding of the head height. The location of further feature points (such as the corners of the eyes and the mouth) would allow more shape units to be fitted. An analysis–by–synthesis approach could then be used to fit the model more accurately to the face and to track it in subsequent frames.

8 Acknowledgements

This work is supported by EPSRC under grant number GR / M 57255. This research acknowledges the use of the Extended Multimodal Face Database and associated documentation [7].

References


