Estimating Face orientation from Robust Detection of Salient Facial Structures

Nicolas Gourier Daniela Hall James L. Crowley
Prima Project, GRAVIR Laboratory, INRIA Rhône-Alpes, Montbonnot, FRANCE

Abstract

This paper addresses the problem of estimating face orientation from automatic detection of salient facial structures using learned robust features. Face imagettes are detected using color and described using a weighted sum of locally normalized Gaussian receptive fields. Robust face features are learned by clustering the Gaussian derivative responses within a training set face imagettes. The most reliable clusters are identified and used as features for detecting salient facial structures. We have found that a single cluster is sufficient to provide a detector for salient facial structures that is robust to face orientation, illumination and identity. We describe how clusters are learned and which facial structures are detected. We show use of this detection to estimate facial orientation.

1. Introduction

We are interested in automatically determining which facial structures can be most reliably detected under variations in illumination, position, orientation and human identity. Our objective is to obtain a set of facial structures that can serve as landmarks for tracking and recognition of facial expressions. We employ a fast, pixel level, detection algorithm to isolate and normalize the face regions. Normalized face images are described by calculating a vector of scale-normalized Gaussian derivatives at each pixel. Salient facial structures are detected using linear combinations of these descriptors. Such functions are learned using K-means clustering of the Gaussian derivative responses obtained from a set of training images. The resulting clusters specify linear combinations of Gaussian derivatives that act as detection functions for facial features that remain salient under variations in pose, illumination and identity.

2. Approaches to Facial Structure Detection

Facial structure detection may be performed using global or local features. A popular method for global analysis of face images is to project a normalized image into a linear subspace determined using a technique such as principal components analysis (PCA) [12]. However, projection highly sensitive to small changes in face position and image scale, as well as partial occlusions and as a result has proved unusable in real systems. In general, global techniques such as projection to a principle components space tend to be sensitive to partial occlusions as well as changes in identity.

An alternative is to measure the relative position of salient anatomical facial structures such as eyes and lips [2]. The challenge is that such facial structures are difficult to detect in a general manner. Most authors rely on complex adhoc operations that tend to be highly sensitive to environmental conditions.

We define salient features as features that draw attention. Features isolated in a dense feature space are salient features [18]. Determining such local feature points can be performed by partitionning the face image into several regions, by using textons as in [8] or finding generic features [4, 10, 11]. Facial structures detection can be done using eigenfeatures [15], blobs [16] or saddle points and maxima of the luminance distribution [17]. But such descriptors are sensitive to illumination and provide an overabundance of points, which can lead to accumulation of errors. Interest points are not robust to pose, and are not well adapted to deformable objects such as the human face.

Our objective is to design descriptors that are robust to illumination, scale and orientation. Such a description can be obtained using Gaussian Derivatives, as well as Gabor Wavelets to describe the appearance of each local neighborhood.

Gabor wavelets provide a very general description function as presented in [3], [14], [7] and [9]. Unfortunately, normalized Gabor wavelets tend to be very expensive to compute.

Similar information can be obtained from a vector of Gaussian derivatives, with the advantage the very fast techniques exist for computing scale normalized Gausian derivatives [13]. We employ such a description to compose a detection function for salient facial features that is invariant to scale, orientation and illumination intensity.

Our approach is composed of several steps. First we employ a robust face tracker to detect and normalize the image of the face. This step, described in section 3, provides a substantial reduction in computation time. Scale normalized Gaussian Derivatives are then computed using a fast pyramid algorithm [13]. Weighted sums of Gaussian derivatives are then used to detect pixels that correspond to salient face regions. The weighting functions are learned by a process that selects combinations of Gaussian derivatives that correspond to the regions that can be detected in the faces of a maximum number of individuals see from a maximum number of viewing directions. This learning process is described in section 4. Face orientation is estimated from the relative positions of the salient regions, as described in section 5. Experimental results using the Pointing '04 face data base are provided in Section 6.

3. Face Image Normalization

We employ a robust video rate face tracker to provide an initial detection and normalization of a face region to a face imagette. Our tracker uses pixel level detection of skin colored regions using a Bayesian estimation of the probability that a pixel corresponds to skin based on its chrominance [6]. This process is described in this section.

3.1. Pixel Level Detection and Tracking using Skin Chrominance

To detect the face, we first detect skin regions in the image using a probabilistic detection of skin chrominance. We compute chrominance by normalizing the red and green components of the RGB color vector by the intensity (R+G+B). Normalizing intensity removes the variations due to angle between the local surface normal and the illumination source. Photons reflected from skin will exhibit a precise value of (r,g) that is determined by the skin pigment and the illumination spectrum. The conditional probability densities for the (r,g) vector for skin regions and for all the image can easily be estimated by histograms. Bayes rule shows that the ratio of these histograms provides a lookup table that maps (r,g) to the conditional probability of skin $p(Pixel \in Skin|r,g)$.

$$p(Pix \in Skin|r,g) = \frac{p(r,g|Pix \in Skin)p(Pix \in Skin)}{p(r,g)} \tag{1}$$

Face position and surface extent are estimated using moments and tracked using a zeroth order Kalman Filter. The tracking process provides a region of interest (ROI) that permits processing to be focused on the face region. Tracking reduces computational cost and improves resistance to distraction by background clutter.

In each image, the skin probability image is calculated within the predicted ROI by table lookup as described above. Pixels within the ROI are then multiplied by a Gaussian predicted by tracking. This step, inspired by robust statistical techniques, improves robustness to background clutter [6].

Both the tracking process and face normalization are based on moments. The first moment (center of gravity) provides a robust estimate of face position, while the second moment provides a measure of the width, height and slant of the face. The first and second moments of the face are used to normalize the face position and orientation, as well as the size of the imagette that represents the face.

We estimate first and second moments with the following formulas (2):

$$\begin{aligned}
&\mu_x = \tfrac{1}{S}\sum p_{skin}(x,y)\cdot x\cdot G(x,y,\vec{\mu},C),\\
&\mu_y = \tfrac{1}{S}\sum p_{skin}(x,y)\cdot y\cdot G(x,y,\vec{\mu},C),\\
&\sigma_x^2 = \tfrac{1}{S}\sum p_{skin}(x,y)(x-\mu_x)^2 G(x,y,\vec{\mu},C),\\
&\sigma_y^2 = \tfrac{1}{S}\sum p_{skin}(x,y)(y-\mu_y)^2 G(x,y,\vec{\mu},C),\\
&\sigma_{xy} = \tfrac{1}{S}\sum p_{skin}(x,y)(y-\mu_y)(x-\mu_x) G(x,y,\vec{\mu},C),
\end{aligned} \tag{2}$$

where $S = \sum p_{skin}(x,y)\cdot G(x,y,\vec{\mu},C)$

3.2. Performance of the Face Tracker

To initialize our face tracker, we employ either the user's selection on the frame, or a generic ratio histogram. The choice of the number of histogram cells used to form the lookup table for skin detection is an important parameter. Histograms with too few cells will not properly discriminate skin from similar colored surfaces such as wood. Inversely, using too many cells renders the process overly sensitive to minor variations in illumination spectrum as well as skin blemishes. We have empirically observed that (r,g) histograms on the order of ranges 32x32 cells provides a good compromise for face detection. A more thorough analysis is provided by [1].

The face tracker has been carefully optimized to run at real time, and can process 384x288 pixel images at video-rate on a 800 MHz Pentium processor. Eye detection rate on representative video sequences can be seen in table 1 and Figure 1. In this case, an error occurs when the computed ellipse does not contain an eye visible in the image.

An important property for a face tracker is jitter. Jitter is measured as the square of the difference in position and size of the detected pixels of the face when the subject is not moving. We have calculated the variance of the moments of the position and size of the detected face region

Table 1. *Eye detection rate*

Sequence	Number of images	Eye Detection rate
A	500	99,9 %
B	700	99,8 %
C	580	94,2 %
D	300	93,1 %

A : Head slow translation
B : Head fast translation
C : Head zoom and inclination in the plane
D : Head pitch and yaw

Table 2. *Stability of the position and the size of the detected face*

Pose	Front	Half-profile	Profile
X Center	0,31 %	1,13 %	3,23 %
Y Center	0,64 %	1,05 %	1,58 %
Width	0,55 %	1,08 %	1,38 %
Height	0,64 %	1,14 %	1,38 %

on sequences of 20 seconds taken when the subject's head has a certain pose and is not moving. Results are shown in Table 2. We observe that many of the errors occur when the subject is in profile. In this case, detection of the neck can modify the detected region.

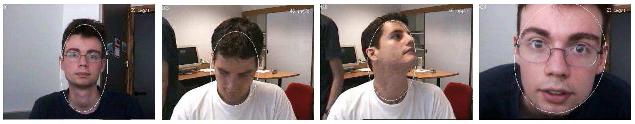

Figure 1. Example of face tracking. First and second moments provides an ellipse which delimits the face in the image

3.3. Normalized Face Imagette

The process described above provides a gray scale (intensity) imagette of the face that is normalized in position and size. Intensity, computed as sum of R+G+B, provides stable salient features based on facial structures. Normalizing the moments of the face imagette allows us to restrict processing to a fixed set of positions and scales, thus reducing computation time, as well as providing a fixed number of operations for each face.

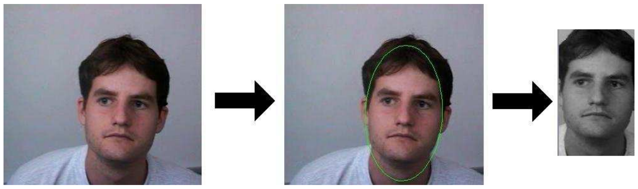

Figure 2. Face Image Normalization

4. Generic Face Features Selection

In this section, we search for facial features robust to changes in illumination, pose and identity. We show how to describe an image with receptive fields, then how to automatically learn facial features with clustering and finally determine salient regions of a face.

4.1. Normalized Receptive Fields

Gaussian derivatives provide a feature vector for local appearance that can be made invariant. We use a five dimensional feature vector computed at each pixel by computing the convolution with the first derivative of a Gaussian in x and y direction (G_x, G_y) and the second derivatives (G_{xx}, G_{xy} and G_{yy}). We use grey-level image of the face to be robust to chrominance variations of lights (sun, neon lights,...). We do not use the zeroth order Gaussian derivative in order to remain robust to changes in illumination intensity. Derivatives of higher order have been found to contribute little information for detection [5].

The feature vector (G_x,G_y,G_{xx},G_{xy},G_{yy}) describes the local appearance of a neighboorhood and is determined using Gaussian derivatives that are normalized to the characteristic scale at each pixel. An example of feature vector of a pixel can be seen in Figure 3. The characteristic scale at each pixel is determined with the local maximum of the Laplacian as function of scale (the scale parameter of the Gaussian), as proposed in [21]. The normalization of face image into an imagette allows us to reduce the range in which the characteristic scale is searched. Two neighboorhoods similar in appearance are close in the feature space. We use a fast, pyramid based, process for determining scale normalized gaussian derivatives [13].

4.2. Learning robust feature detectors by clustering

The vector of Gaussian derivatives form a feature space. In order to provide a distance metric, Gaussian derivative vectors may be normalized by their variance taken from a sample set. The vectors of Gaussian derivatives from face imagettes taken from a variety of viewing angle form clouds

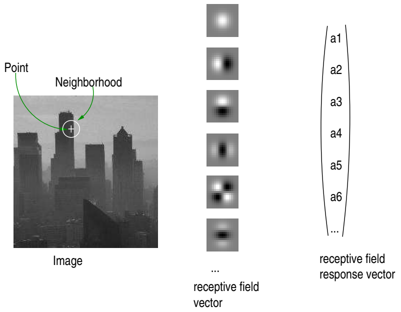

Figure 3. Appearance based feature vector

in this feature space. Each such cloud corresponds to a linear combination of Gaussian derivatives that do not change as viewing angle changes. Such clouds may be detected using K-means clustering.

A clustering algorithm, such as K-means can be used to determine a local description of appearance for specific facial structures. The center of gravity of a cluster can be used to determine the coefficients for a linear classifier. The mass of the cloud provides the basis for the determining the suitability of the cluster. Ideally we want clusters that have a low mass in each image (i.e. that correspond to a few specific facial structures), but a high overall mass in a set of training images taken from different viewpoints. Gaussian derivative vectors that satisfy both criteria are ideal for robust facial structure detection. Our experiments have shown that such clusters are sufficiently robust that even a single cluster can provide a robust detector for salient facial structures.

4.3. Robust Facial Structures

Applying clustering to the feature vectors for multiple images from several faces provide appearance clusters for background, hair and different skin regions as well as salient facial structures. For each pixel, we determine the most likely cluster using a sum of squared difference from the cluster center. The squared difference of each Gaussian derivative is normalized by the overall variance of that derivative. The sum of the squares of the normalized distance provide a similarity metric.

The process of robust facial structures detection is shown in Figure 4. Each pixel is assigned to the most likely cluster as defined by the smallest normalized distance. If the normalized distance is greater than a threshold, the pixel is assigned to a ”background” class. Adjacent pixels in the same class are detected by connectivity analysis and grouped to form image regions. These image regions correspond to salient facial structures such as the eyes, nose, mouth and chin.

Detection using this method can give rise to a number of small spurious detected regions. These can be eliminated by using a bounding box of the region as defined by the connected components algorithm. Regions with a small bounding box are eliminated. The remaining regions correspond to salient facial structures. Connected components analysis also provides some geometrical information about the detected regions. The first and second moments of the connected components provides information about position and extent. This information can be reprojected to the original image.

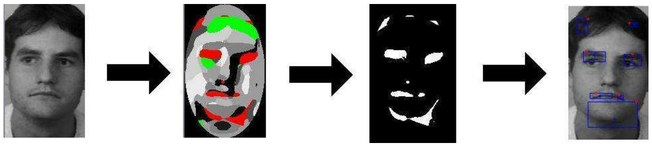

Figure 4. Robust facial structures detection process

Face image normalization,
Mapping : Regions in red and green are considered as salient robust facial structures and reprojected into a binary map,
Connected components analysis

5. Pose Estimation

Head orientation, or pose, is determined by 2 angles, the vertical angle α_v and the horizontal angle α_h. A dense sampling of appearance space, such as provided by the Pointing ’04 database, makes it possible estimate these angles by image classification. A more precise estimate requires geometric calculation based on the relative image positions of salient image structures. These two methods may be used in a complementary manner, with the coarse estimate obtained by classification used to initialize a refined calculation based on image position of salient facial structures.

In this section we discuss how to compute this more refined calculation based on the relative image positions of salient facial structures.

5.1. Detecting Eyes

The position of salient facial structures using the method described above can vary with respect to image pose, as illustrated in figure 5. Even for a particular viewing angle, a particular robustly detected salient facial structure may occur at different relative positions for different subjects. For example, figure 6 shows the eye positions detected for several people. Our conclusion is that the simple position of robust facial structures is not sufficient to allow direct structure identification.

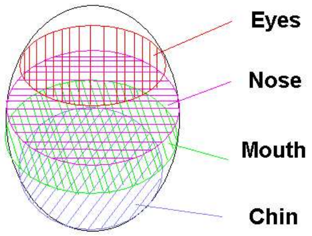

Figure 5. Facial structures position variation for one person when the pose is changing

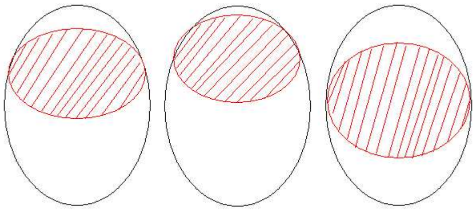

Figure 6. Eyes postion variation for 3 subjects when the pose is changing

We use a bayesian classifier to identify detected regions corresponding to particular salient facial structures. We estimate the probability that a bounding box contains a particular facial structure. Eyes have been found to be the most salient of facial structures (see Section 6.3 for details). Furthermore, knowing their position in the face provides strong geometric constraints for searching for other facial structures. Thus our first step is to identify the bounding boxes that correspond to the eyes.

There are three possible configurations of detected eye regions for bounding boxes that contain eyes :
1) One bounding box for each eye
2) One bounding box for both eyes
3) One bounding box for one eye. This situation appears when the face is turned so that only one eye is visible from the camera.
Giving a configuration, we compute the probability that each bounding box corresponds to an eye. Configurations are tested in the order of the three configurations listed above. The bounding box containing eyes are selected with a winner-takes-all process using Bayes rule. Eyes are identified and their position is given by the center of gravity of the bounding box in configurations 1 and 3, and extremities of the bounding box in configuration 2.

5.2. Computing Head Pose

In this section we discuss how to compute α_h and α_v using robust facial structures. Because of facial symmetry, horizontal pose can be estimated with positions of both eyes with regard to the face. The trigonometric computation of the horizontal angle is shown figure 7. We obtain the following equations (3):

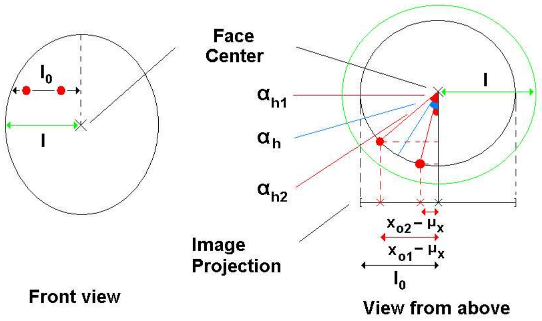

Figure 7. Horizontal pose computation

$$x_o^1 - \mu_x = l_0 sin(\alpha_h^1)$$

$$x_o^2 - \mu_x = l_0 sin(\alpha_h^2)$$

$$\alpha_h = \frac{(\alpha_h^1 + \alpha_h^2)}{2} \tag{3}$$

The relative position of eyes is not sufficient to estimate the vertical pose α_v. Because of the variation of eye position in the face when $\alpha_v = \alpha_h = 0$ due to the subjects, specifying eyes position for $\alpha_v = 0$ is a difficult task. This problem can be bypassed by calibrating an eyes position at $\alpha_v = 0$ for each subject, but the system becomes dependent on identity. We must use positions of other robust facial structures to estimate the vertical pose. But as with eyes, positions of salient structures in the face varies with identity. Furthermore, even for human eyes, a 15 degrees difference in vertical orientation is not apparent. A solution would be to consider distances between other facial structures.

6. Experimental Results and Discussion

6.1. Training data

The choice of a good database is essential for proper learning. To detect salient facial structures that are robust under changing conditions, we used two front images of 15 subjects to learn feature vectors. Subject were 20 to 40

Table 3. *Recall / Precision in % when changing the head orientation in the training process*
Results obtained with a detection threshold of 0.25

Images	Frontal	Near-Frontal	All
Person 1	36,7 / 30,1	40,9 / 4,8	31,1 / 24,6
Person 2	34 / 35,4	35,6 / 4,1	35,2 / 6,7

years old. Five subjects have facial hair and seven people wear glasses. Non-frontal images can introduce noise in the data, because some facial structures have different appearances in different poses. As an example, the experiments from two subjects are shown in Table 3. Front pose provides more generic appearance for salient facial structures, which remain robust on multiple poses after learning, whereas profile images provides appearance for salient structures in profile, but not for front.

To remain robust to changes in identity, we have used images from 15 different people from the Pointing '04 database. These subjects may be grouped into two classes:

- Class A, in which the face is typical with regard to people in the database. In the Pointing '04 data, 73% of the subjects have white skin, European facial structure and no beard.
- Class B, in which the face is atypical in some respect. Examples of atypical faces from the database include those who wear glasses, have a beard or have a slightly different skin pigment. In the Pointing '04 data, 27% of the subjects have darker skin, oriental facial structures or a beard.

We have observed the following results from the learning process:

- The clustering C(A) is performed only with faces from class A.
  - Regions obtained for facial structures of a subject a ∈ A are significant and robust.
  - Regions obtained for facial structures of a subject b ∈ B are less significant and more noisy.
- The clustering C(A+a) is done with people belonging to class A and a new subject (a) also belonging to class A. Regions obtained for facial structures of the subject (a) are less robust than those obtained with the previous clustering C(A), indicating that robustness decreases as we add subjects.
- The clustering C(A+b) is done with people belonging to class A and a single subject (b) belonging to class B. Regions obtained for facial structures for the subject (b) are less noisy and more salient than those obtained with the previous clustering C(A).

Table 4. *Results obtained with 30 front images and a detection threshold of 0.4*

Number K	2	3	5	7
Recall	11,7 %	22,7 %	70,7 %	30,7 %
Precision	2,3 %	13,1 %	18,2 %	21,5 %

Number K	10	15	20
Recall	40,2 %	12,2 %	6,1 %
Precision	47,7 %	57,3 %	11,7 %

These observations can be explained in the following way. The clustering C(A) performed with "common" faces provides better results on subjects of class A, than subjects of class B. Therefore the clustering C(A) is not well adapted for subjects of class B. We must then use other people in our learning process to remain robust to changes in identity.

Adding a new subject a from ∈ A in the clustering does not bring much additional information, even on the subject (a). Furthermore, it can lead to a degradation of robustness and more noise, because the method becomes specialized for people from the class A, degrading independence to identity. The method may be said to "overfit" the training data and lose generality.

Adding a new subject b ∈ B provides better detection of facial structures on (b), whose appearance differs from those of class A. Clustering C(A+b) adapts to the image of the face of (b) without becoming specialized. Furthermore, salient facial structures are more often detected with C(A+b) than with the clustering C(A).

6.2. Influence of the number of clusters

The clustering step gathers feature vectors into K clusters. This step is an important part in the learning process. Therefore, the choice of the number of clusters K is crucial. If K is too small, appearance clusters won't be discriminative enough to detect salient facial structures. If K is too big, regions will be too small and too unstable in the image. During our experiments, we tested several K and obtained good results with K = 10. Resulting images with different number of clusters can be seen in Figure 8.

To measure the recall and the precision for each different K, we have employed a 10x15 grid on the normalized imagette of the face (see Table 4). Cases in the grid are manually labelled as follows : 1 if the case contain a facial salient structure, 0 otherwise. During the tests, a case of the grid gets the value 1 if the ratio of the number of salient cluster pixels in the case over the total number of pixels in the case exceeds a fixed threshold. This threshold is called

Table 5. *Recall/Precision with regard to detection threshold*

Detection Threshold	0,1	0,25	0,4
Recall	46,3 %	34 %	23,2 %
Precision	22,2 %	25,4 %	26,9 %
Detection Threshold	0,5	0,66	0,75
Recall	17,9 %	10,3 %	7,5 %
Precision	27 %	27 %	27,3 %

the detection threshold (see Table 5).

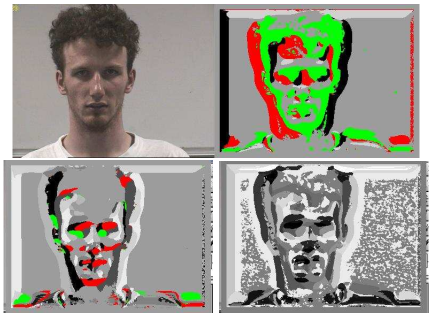

Figure 8. Influence of the number of clusters
Regions in red and green are considered as salient robust facial structures
Top left image is the original image
Top right image is obtained with 5 clusters, which are not discriminative enough
Bottom left image is obtained with 10 clusters
Bottom right image is obtained with 15 clusters. Regions are too small to be relevant

6.3. Facial structure detection performance

Tests have been made with representative people under changing lighting and pose conditions. The pose is determined by 2 angles (h,v), which vary from -90 degrees to +90 degrees. Each set contains 93 images of the same person at different poses. The Pointing ’04 database includes faces with glasses as well as a variety of skin pigments. We have calculated the detection rate for each structure for four representative faces (see Table 6).

With an average detection rate of 97 %, eyes are the most often detected facial structure. Eye appearance does not vary as much as the other facial structures because of their spherical shape and thus eyes can be detected under several points of view. Glasses have little effect on eye detection. The salience of a mouth improves when it is surrounded by a beard and thus mouth detection is slightly better than eye detecion for bearded subjects.

Table 6. *Facial structure positive detection rate*

Structure	Person 1	Person 2	Person 3	Person 4
Eyes	99 %	97 %	98 %	95 %
Nose	70 %	82 %	61 %	82 %
Mouth	85 %	90 %	95 %	85 %
Chin	84 %	88 %	91 %	84 %
Specificity	-	Glasses	Beard	Matt skin

For 63% of the observed errors, the head pitch is inferior to -30 degrees, indicating that the subject is looking down. This situation represents only 29% of all poses. Indeed, in this situation, eyes are no more visible in the image, but only eyebrows. Therefore, we have trained our algorithm on images on which subjects’ head pitch is inferior to -30 degrees. In this case, the resulting clusters are less discriminating and provides lower detection rate on face images. As a consequence, some facial structures, such as chin and eyes, are less salient. Eye detection is 59% less efficient with the algorithm trained with images of people looking down. The nose has the worst detection average rate with 74%. It does not have as many symmetry properties as eyes and its appearance can suffer many variations. That is why the nose is less often detected than other facial structures.

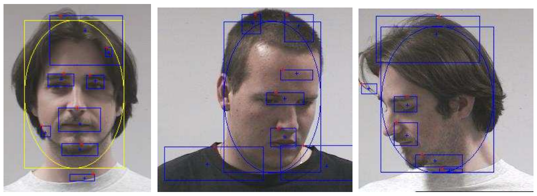

Figure 9. Examples of facial structures detection [19]

6.4. Influence of the size of the face imagette

To show the importance of the face image normalization step, we have measured eye detection rates with different sizes of the face imagettes. Results of these experiments can be seen in table 7. Tests have been made on a sequence of 500 images in which the subject moves but has both eyes remaining visible on the screen. The head size changes from 50x50 to 20x20 pixels in the sequences.

The last size, 50x50 pixels, corresponds to face image analysis without normalization, as the face in the sequence

Table 7. *Eyes detection rate with regard to the size of the imagette in pixels*

Size	120x200	120x120	60x100	50x50
Detection	98,2 %	97,8 %	94,2 %	1,8 %

has a maximal size of 50x50 pixels. We can see how the normalizing the first and second moments of the imagette enhances the detection rate. This provides the ability to deal with 20x20 pixels images of the head, such as panoramic or wide-angle public cameras images. When this operation is not performed, regions will be more imprecise and may not be found. Increasing the size of the normalized face image increases the accuracy of structure detection in the original image of the face. For our experiments, the face imagette has a size of 60x100 pixels.

6.5. Pose Estimation

Due to difficulties in estimating the vertical pose, we have only estimated the horizontal pose. Absolute difference between the real and the estimated horizontal angle α_h has been computed for all 1395 images of fifteen subjects in the Pointing'04 database. Mean absolute error in degrees of the horizontal angle for each pose is represented figure 10.

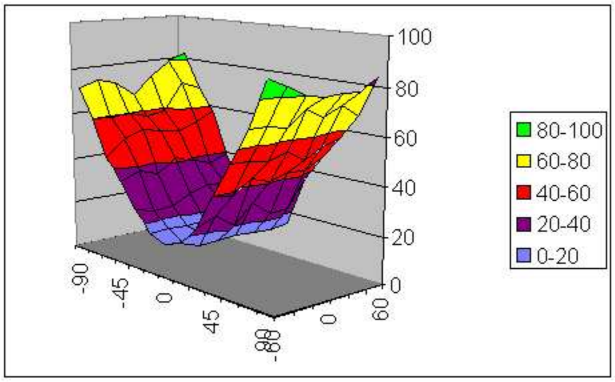

Figure 10. Mean absolute error in degrees of horizontal angle at each pose

Mean error of the horizontal angle does not vary much as vertical pose changes. When $|\alpha_h| <= 45^{\mathrm{o}}$, mean error drops to 15 to 5 degrees. But at $|\alpha_h| > 45^{\mathrm{o}}$, mean error can reach 90 degrees. There are several explanations for this observation.

First, in the case only one eye is visible at the screen, which roughly corresponds to $|\alpha_h| >\approx 45^{\mathrm{o}}$, estimating the horizontal pose becomes difficult, because the other eye cannot be seen. As we need two position for eyes to estimate α_h, the computation of the horizontal pose can not be made accurately. Furthermore, the fact that people do not have the same distance between eyes makes the prediction of the position of the other eye inaccurate and computing the horizontal angle is even more difficult.

Another problem is the neck detection. The user's neck can be detected or not as part of the face because of its chrominance. In profile, detecting the neck disrupts face orientation estimation and can modify the slant angle of the face in the image plan. As the horizontal pose estimation relies on the face estimation, the neck also yields a false estimation for the horizontal pose.

An additional problem is caused by hair. Hair are the part of the face that vary the most with regard to identity, degrading invariance to identity. When the user is in profile, hair is more visible in the image. Finally, A mesh of hair is sometimes detected as an eye, and this detection provides a false estimation for the horizontal pose.

Conclusions

We have proposed a new approach to detect salient facial structures in a manner that is robust to changes in viewing angle, illumination and identity. The imagette containing the face is normalized in scale and orientation using moments provided by a face tracker. Each pixel in the face image is associated with an appearance cluster. One particular cluster stands for salient robust face structures which are: eyes, nose, mouth, chin. We have tried to extract and exploit a maximum of information provided by a single image of a face and to limit the loss of generality.

Detected regions can be delimited with rectangles in the image. Identifying facial structures using positions relative to the face image is difficult because multiple variations of structures are possible.

These variations are due to changing orientation, facial expression of emotion and especially identity. A Bayesian classifier is used to identify the regions. Eyes have been found to be the most salient of the facial structures. They can be used to obtain a coarse estimation of the horizontal pose, but are not sufficient to compute vertical pose. Because of variations in the structures in the face with regard to the identity and the pose, vertical pose is difficult to accurately estimate.

Mean error for the horizontal pose does not vary with vertical angle. Error reaches 5 to 15 degrees when $|\alpha_h| <= 45^{\mathrm{o}}$, but increases when $|\alpha_h| > 45^{\mathrm{o}}$. This is due to the fact the horizontal angle is hard to estimate with only one eye visible on the image and that the neck detection disrupts the face estimation. Hair can also be misclassified as eyes. All these observations tend to show that the robustness to identity is the most difficult criteria to respect.

References

[1] M. Storring "Computer Vision and Human Skin Color" Doctoral Thesis, Alborg University, June, 2004.

[2] M. Zobel, A. Gebhard, D. Paulus, J. Denzler, H. Niemann "Robust Facial Feature Localization by Coupled Features" *4th International Conference on Automatic Face and Gesture Recognition*, March 28-30/2000 Grenoble, France pp. 2-7.

[3] D. Hall, J.L. Crowley "Computation of Generic Features for Object Classifi cation" *ScaleSpace*, 2003.

[4] S. J. McKenna, S. Gong, R. P. Wurtz, J. Tanner, D. Banin "Tracking Facial Feature Points with Gabor Wavelets and Shape Models" *1st Int. Conf. on Audio- and Videobased Biometric Person Authentication*, Lecture Notes in Computer Science 1997.

[5] D. Hall "Viewpoint Independant Object Recognition from Local Appearence" *PhD Thesis*, 2001.

[6] K. Schwerdt, J.L. Crowley "Robust face Tracking using Color" *Automatic face and Gesture Recognition*, pp 90-95, 2000.

[7] L. Wiskott, J-M. Fellous, N. Kruger, C. von der Malsburg "Face Recognition by Elastic Bunch Graph Matching" *Pattern Analysis and Machine Intelligence*, Vol 19 pp 775-779, 1997.

[8] J. Malik, S. Belongie, T. Leung, J. Shi "Contour and Texture Analysis for Image Segmentation" *IJCV*, Vol 43 pp 7-27, 2001.

[9] V. Kruger, G. Sommer "Gabor Wavelets Networks for Object Representation and Face Recognition" *Deutsche Arbeitsgemeinschaft fur Mustererkennung*, Vol 22, Sept. 13-15, 2000.

[10] C. Schmid "Constructing Models for Content-Based Image Retrieval" *Computer Vision and Pattern Recognition*, 2001.

[11] D. Lisin, E. Risemann, A. Hanson "Extracting Salient Image Features for Reliable matching using Outlier Detection Techniques" *Computer Vision Systems Third International Conference*, pp 481-491, 2003.

[12] M. Turk, A. Pentland "Eigenfaces for Recognition" *Cognitive Neuroscience*, Vol 3(1), pp 71-96, 1991.

[13] J.L. Crowley, O. Riff "Fast Computation of Scale Normalised Receptive Fields" *International Conference Scale Space*, June 2003, Island of Skye.

[14] Y. Wu, K. Toyama "Wide-Range, Person- and Illumination-Insensitive Head Orientation Estimation" *4th International Conference on Automatic Face and Gesture Recognition*, March 2000 Grenoble, France pp. 183-188.

[15] A.C, Varchmin, R. Rae, H. Ritter "Image based Recognition of gaze Direction using Adaptative methods" *International Gesture Workshop*, Springer, 1997, pp. 245-257.

[16] A.J. Howell, H. Buxton "Active Vision Techniques for Visually Mediated Interaction" *Image and Vision Computing 20(12)*, 2002, pp.861-871.

[17] T. Otsuka, J. Ohya "Real-time Estimation of Head Motion Using weak Perspective Epipolar Geometry" *Proceedings of WACV'98*, October 1998, Princeton.

[18] K.N. Walker, T.F. Cootes, C.J Taylor "Automatically Building Appearance Models from Image Sequences using Salient Features" *IVC(20), No. 5-6*, 15 April 2002, pp. 435-440.

[19] A.M. Martinez and R. Benavente "The AR Face Database" *CVC Technical Report n.24*, June 1998.

[20] G.J. Klinker, S.A. Shafer, T. Kanade "A Physical Approach to Color Image Understanding" *IJCV 1990*

[21] T. Lindeberg "Feature Detection with Automatic Scale Selection" *IJCV 1998 (30) Number 2*, pp. 79-116.