5.0 HIGH GRADIENT COMPONENT ANALYSIS

Facial motion produces transient darkened skin-color lines or edges perpendicular to the motion direction of the activated muscle. Those darkened lines or edges are called furrows or wrinkles. The facial action produces the motion position, shape, length and gray-value changes of these furrows in the face image and which are strongly associated with different meanings of the facial expression. Extracting (segmenting) and realizing the motion of those high gradient components (*i.e.*, furrows) may provide a very important source for recognizing facial expressions.

5.1 High Gradient Component Detection in the Spatial Domain

The shapes of furrows in the face image contain horizontal, vertical and/or diagonal directions of lines or arched curves (Figure 32). To extract the high gradient component at pixel (x,y) from the face image (in the spatial domain) I at time t, we use horizontal, vertical and diagonal line or edge detectors, L_x, L_y, L_{xy}, respectively.

$$\frac{\partial I(x,y,t)}{\partial x} = L_x \otimes I(x,y,t) = D_x(x,y,t) \tag{5-1}$$

$$\frac{\partial I(x,y,t)}{\partial y} = L_y \otimes I(x,y,t) = D_y(x,y,t) \tag{5-2}$$

$$\frac{\partial I(x,y,t)}{\partial (xy)} = L_{xy} \otimes I(x,y,t) = D_{xy}(x,y,t) \tag{5-3}$$

and

$$D(x,y,t) = \left\{ D_x(x,y,t), D_y(x,y,t), D_{xy}(x,y,t) \right\} \tag{5-4}$$

where $I(x,y,t)$ is the image gray value at position (x,y) and frame t, and $\otimes$ denotes convolution. $D_x(x,y,t)$, $D_y(x,y,t)$ and $D_{xy}(x,y,t)$ are gradient intensities of the high gradient

components at pixel (x,y) in the horizontal, vertical and diagonal directions, respectively. $D(x,y,t)$ is used as the general term of the gradient intensities including $D_x(x,y,t)$, $D_y(x,y,t)$ and $D_{xy}(x,y,t)$.

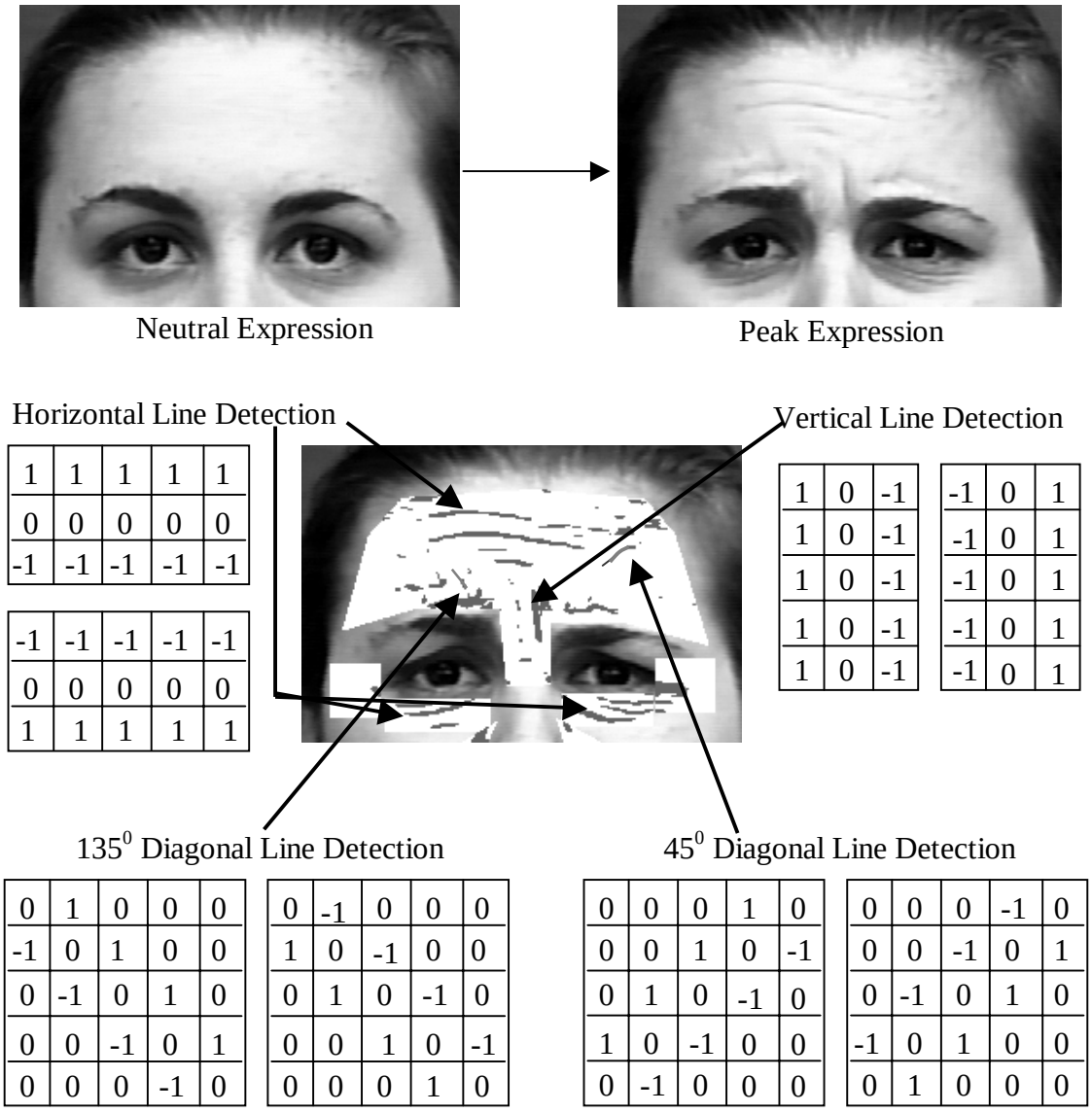
Before normalization of each 417 x 385-pixel image using affine transformation by three feature points (the medial canthus of both eyes and the uppermost point on the philtrum), a 5 x 5 Gaussian filter is used to smooth the image. For the upper face expression (Figure 32.a), a 3 x 5 (row x column) horizontal- and a 5 x 3 vertical-line detectors are used to detect horizontal lines (*i.e.*, high gradient components in the vertical direction) and vertical lines in the forehead region, around the eye region, and between brows or eyes for AU4, AU1+4 and AU1+2 expressions. Two 5 x 5 diagonal-line detectors are used to detect 45-degree and 135-degree diagonal furrows at the forehead region during AU1+4 expression. For the lower face expressions, two 5 x 5 diagonal-line detectors are used to detect 45-degree and 135-degree diagonal lines along the nasolabial furrow (Figure 32.b). Two 3 x 3 edge detectors are used to detect high gradient components around the lips and on the chin region by thresholding their magnitudes (Figure 32.b and 32.c). The components filtered out (or thresholded out) by the line or edge detectors are set to a value of zero.

5.2 High Gradient Component Detection in the Spatio-Temporal Domain

To verify the detected high gradient component $D(x,y,t)$ which is produced by transient skin or feature deformations and not a permanent characteristic of the individual's face (Figure 33), it is necessary to consider its the temporal gradient.

$$\frac{\partial D(x,y,t)}{\partial t} = D(x,y,t) - D(x,y,t-1) = \Gamma_t(x,y) \qquad (5\text{-}5)$$

where $D(x,y,t)$ and $D(x,y,t\text{-}1)$ are the gradient intensities of the detected high gradient components at pixel (x,y) and frames t and t-1 in the spatial domain, respectively. $\Gamma_t(x,y)$ is the temporal gradient intensity between gradient intensities at frames t and t-1 at pixel



1	1	1	1	1
0	0	0	0	0
-1	-1	-1	-1	-1

-1	-1	-1	-1	-1
0	0	0	0	0
1	1	1	1	1

1	0	-1
1	0	-1
1	0	-1
1	0	-1
1	0	-1

-1	0	1
-1	0	1
-1	0	1
-1	0	1
-1	0	1

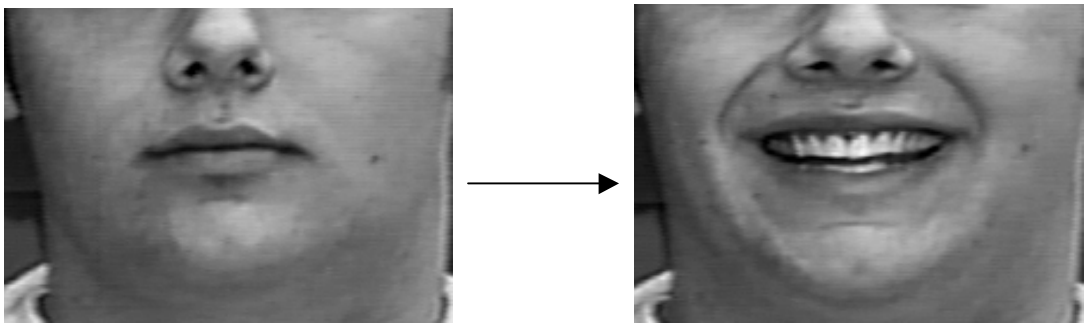
0	1	0	0	0
-1	0	1	0	0
0	-1	0	1	0
0	0	-1	0	1
0	0	0	-1	0

0	-1	0	0	0
1	0	-1	0	0
0	1	0	-1	0
0	0	1	0	-1
0	0	0	1	0

0	0	0	1	0
0	0	1	0	-1
0	1	0	-1	0
1	0	-1	0	0
0	-1	0	0	0

0	0	0	-1	0
0	0	-1	0	1
0	-1	0	1	0
-1	0	1	0	0
0	1	0	0	0

Figure 32.a High gradient component (furrow) detection for the forehead and eye regions.



Neutral Expression

Peak Expression

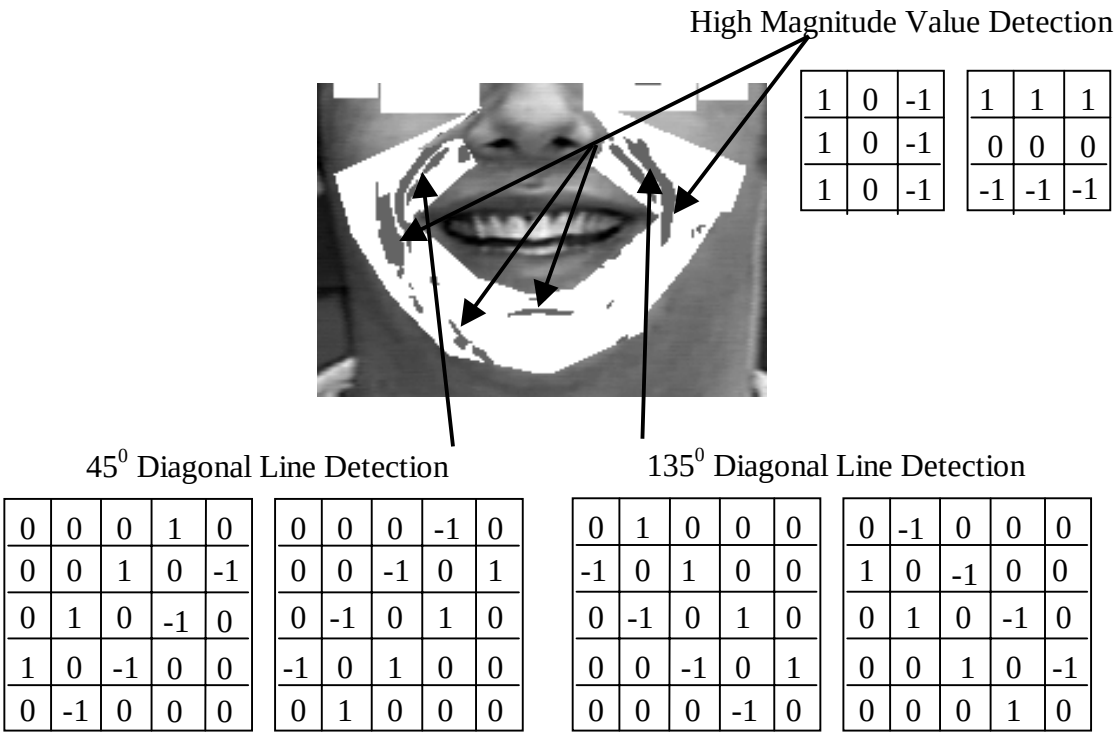


Figure 32.b High gradient component (furrow) detection for the mouth, cheek, and chin regions.

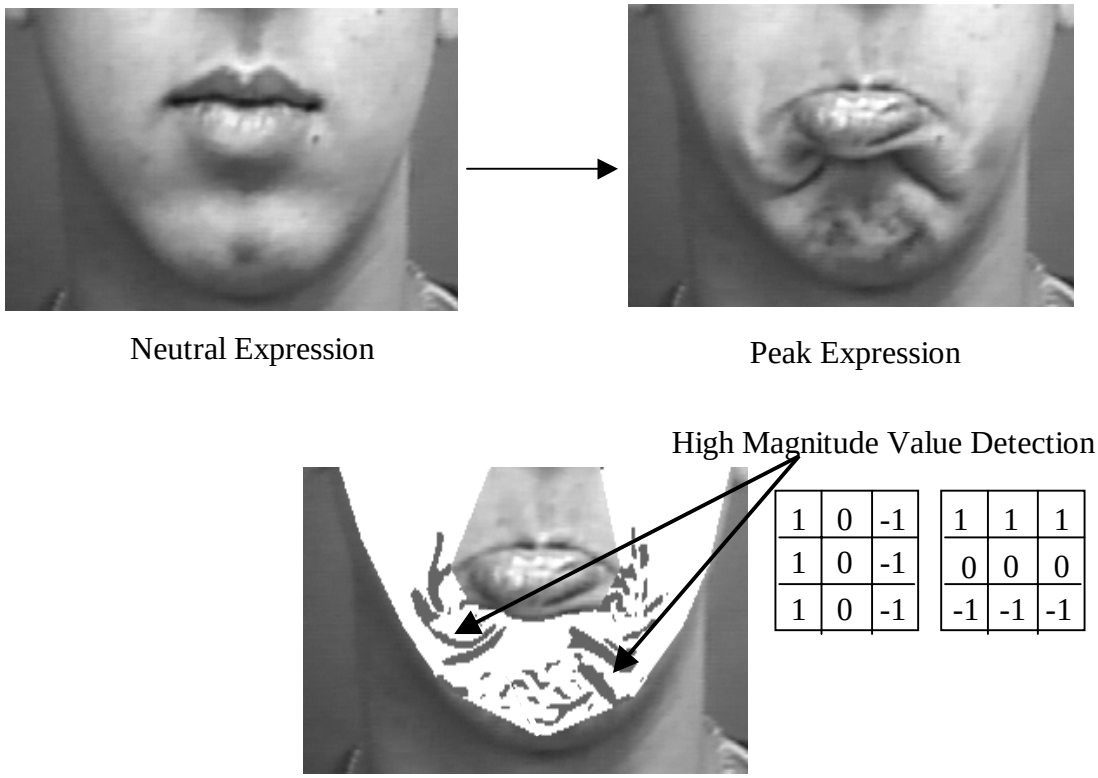


Figure 32.c High gradient component (furrow) detection for the chin region.

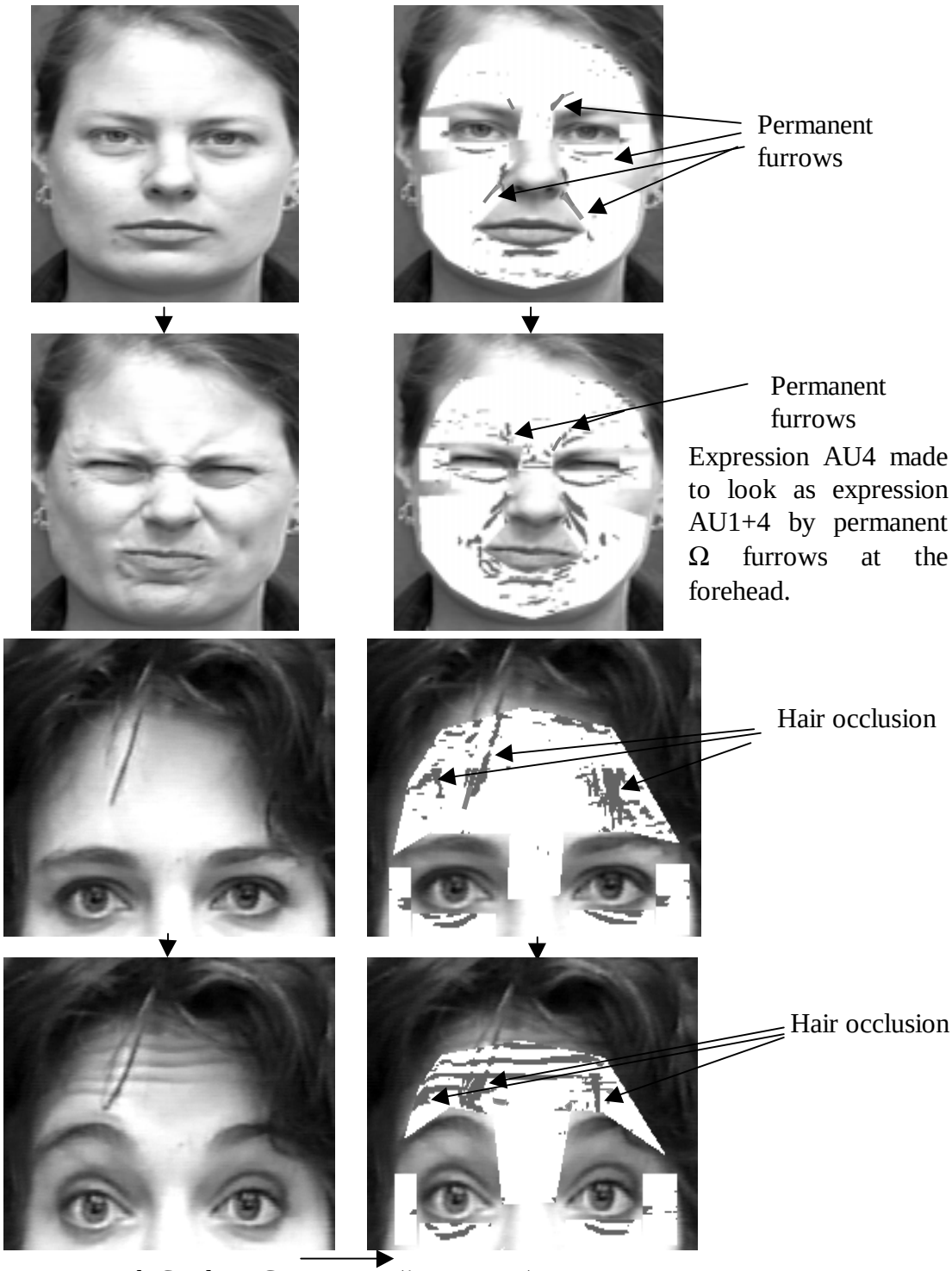


Figure 33 Permanent furrows or hair occlusion.

(x,y) considering the spatio-temporal domain. Equation (5-5) is the method used for tracking high gradient components such as the motion lines or edges in the spatial and temporal domains.

In our current study, the gradient intensity of each detected high gradient component $D(x,y,t)$ at the current frame t and pixel (x,y) is compared with corresponding points within a 3 x 3 region of the first frame for each sequence.

$$\frac{\partial D^0(x,y,t)}{\partial t} = D(x,y,t) - D(x-\Delta x, y-\Delta y, 0) = \Gamma_t^0(x,y) \qquad (5\text{-}6)$$

where

$$-1 \le \Delta x \le 1 \quad \text{and} \quad -1 \le \Delta y \le 1$$

A 3 x 3 region is used in order to avoid the error of geometrical correspondence since affine transformation works well for close (but not exact) geometrical correspondence. If the absolute value of the difference in gradient intensity between these points is higher than the threshold value, it is considered a valid high gradient component produced by facial expression. All other high gradient components are ignored. In the former case, the high gradient component (pixel) is assigned a value of 1. In the latter case, the pixels are assigned a value of 0. An example of the procedure for extracting high gradient components in the spatio-temporal domain for the upper facial expression is shown in Figure 34. A gray value of 0 corresponds to black and 255 to white. Using this procedure, we also can remove the hair blocking the forehead region (Figure 35).

5.3 Morphology and Connected Component Labeling

In order to use line and edge detectors, the threshold is employed to segment the higher gradient components for the foreground furrows and the lower gradient components for the background. If the given threshold is too high, then it will filter out more significant components. If the threshold is given too low, then it will include more high gradient components with unnecessary noise (Figure 36). For line or edge detection,

it is very difficult to give a constant or even dynamic threshold to satisfy all conditions, especially in dealing with images containing the conditions of rigid and non-rigid motion such as images of facial expression whose gray values vary in lighting, ages and individuals. Younger subjects, especially infants, show smoother furrowing than older ones, and initial expressions show weaker furrowing than that of peak expressions for each sequence (Figure 37). To overcome this difficulty, further low level image processing is needed.

Because we do not want to lose any useful information of high gradient components, we give a low threshold for each furrow detection processing sequence (Figure 36). An erosion morphological transformation is used to eliminate the piece regions or the very short lines, thin the lines, or smooth the line boundary (Figure 38.a). The eliminated (either one or several) pixels are assumed as noise and introduced might be because of the low threshold. A dilation morphological transformation is then used to connect two end-to-end close but separated lines (Figure 38.a). Finally, we implement a connected components labeling (CC labeling) algorithm based on Haralick and Shapiro's [43] to label each cluster of connected high gradient components (Figure 38.b). This algorithm is based on the 8-connected component to link components of its 8 neighbors at the binary image (1 is the high gradient component, 0 is the background) and includes two processes: top-down and bottom-up processes with 4 steps (APPENDIX). If the number of detected high gradient components for each connected component cluster is less than 6 pixels, or the horizontal to vertical ratio for the horizontal line detection or the vertical to horizontal ratio for the vertical line detection is less than 5, then this cluster is assumed to be noise, not furrow, and will be removed (Figure 38.b).

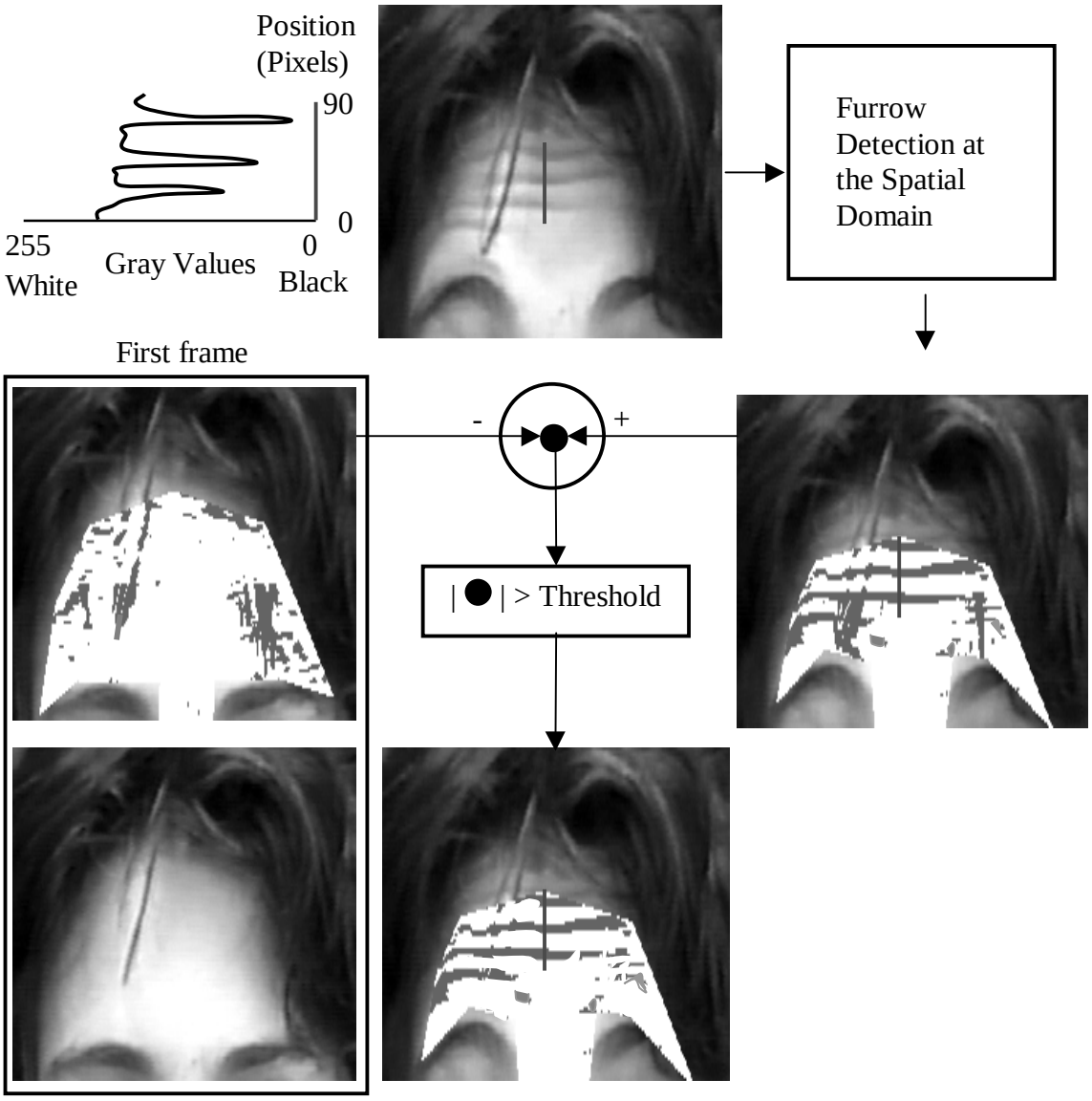


Figure 34 The procedure of the high gradient component analysis in the spatio-temporal domain, which can reduce the effect of the permanent high gradient components (furrows) and hair occlusion for the upper facial expression.

Figure 35 (a) Original gray value images. (b) High gradient component (furrow) detection in the spatial domain. (c) High gradient component analysis in the spatio-temporal domain.

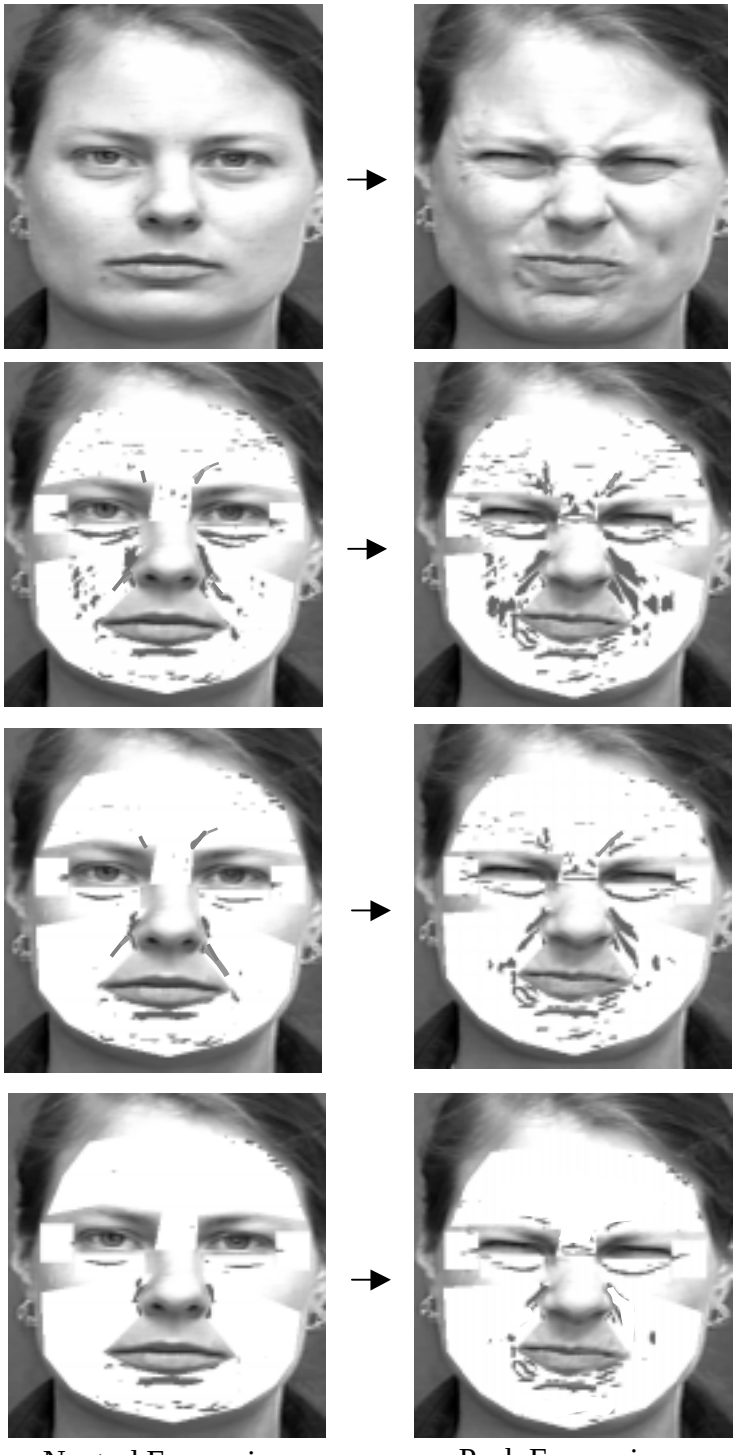
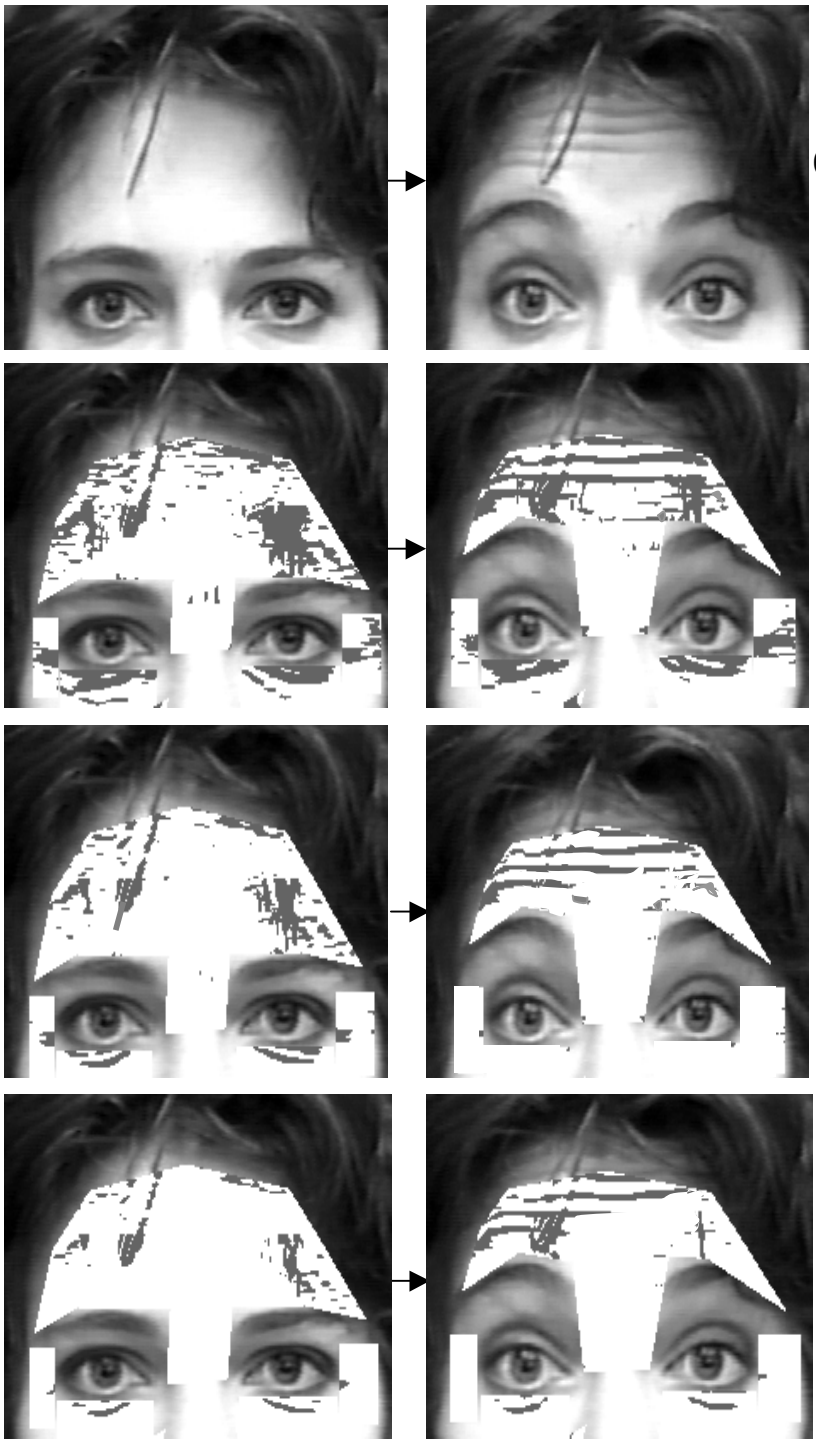


Figure 36.a High gradient component detection with different constant threshold values.



(a) Gray value image

(b) High gradient component (furrow) detection with a very low threshold, which will include too many furrow or noise information.

(c) High gradient component (furrow) detection with a low threshold, which was used in our study.

(d) High gradient component (furrow) detection with a high threshold, which will lose some furrow information.

Neutral Expression Peak Expression

Figure 36.b High gradient component detection with different constant threshold values.

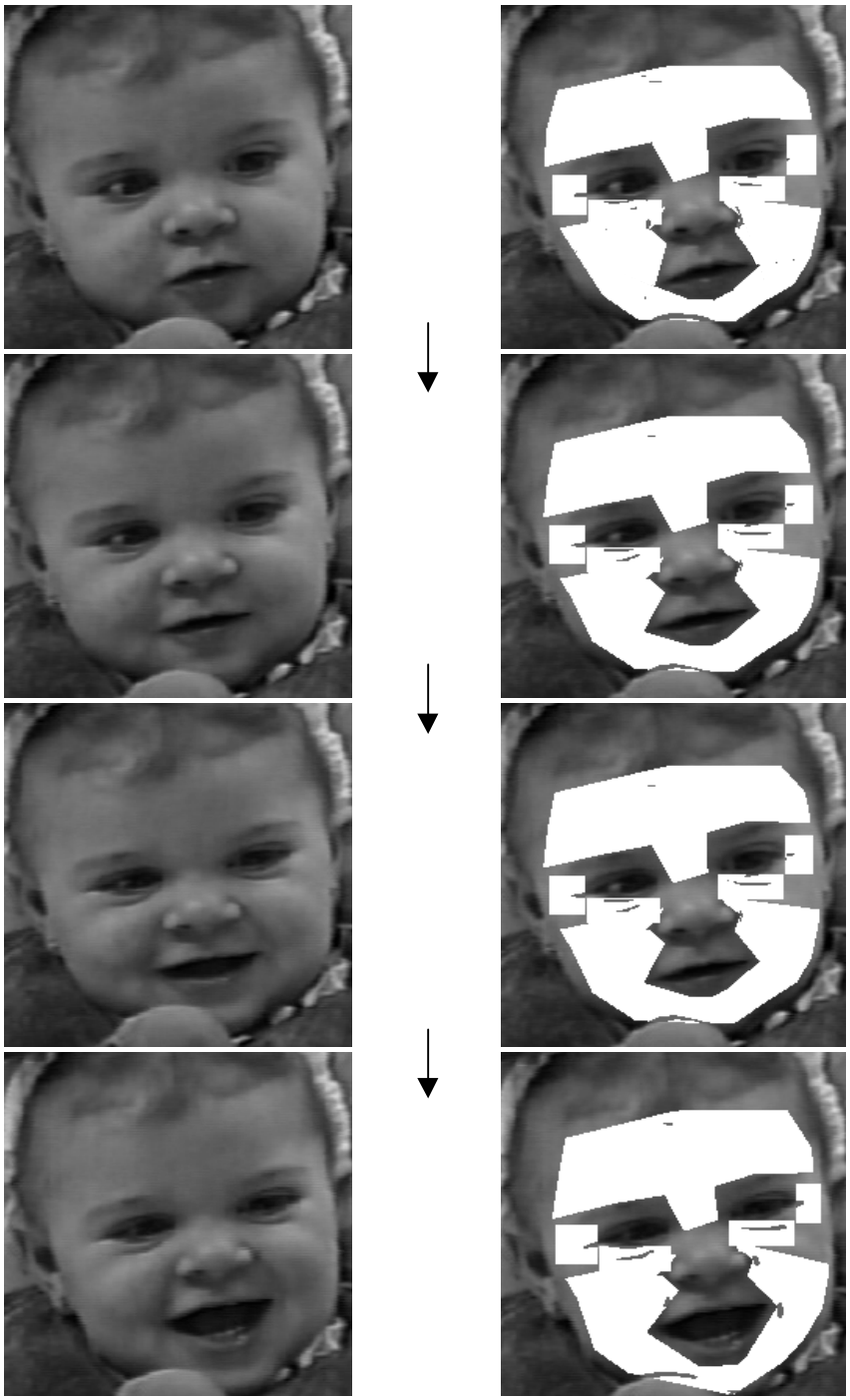


Figure 37.a Younger subjects, especially infants (Figure 37.a), show smoother furrowing than older ones (Figure 37.b), and initial expressions show weaker furrowing than that of peak expressions for each sequence.

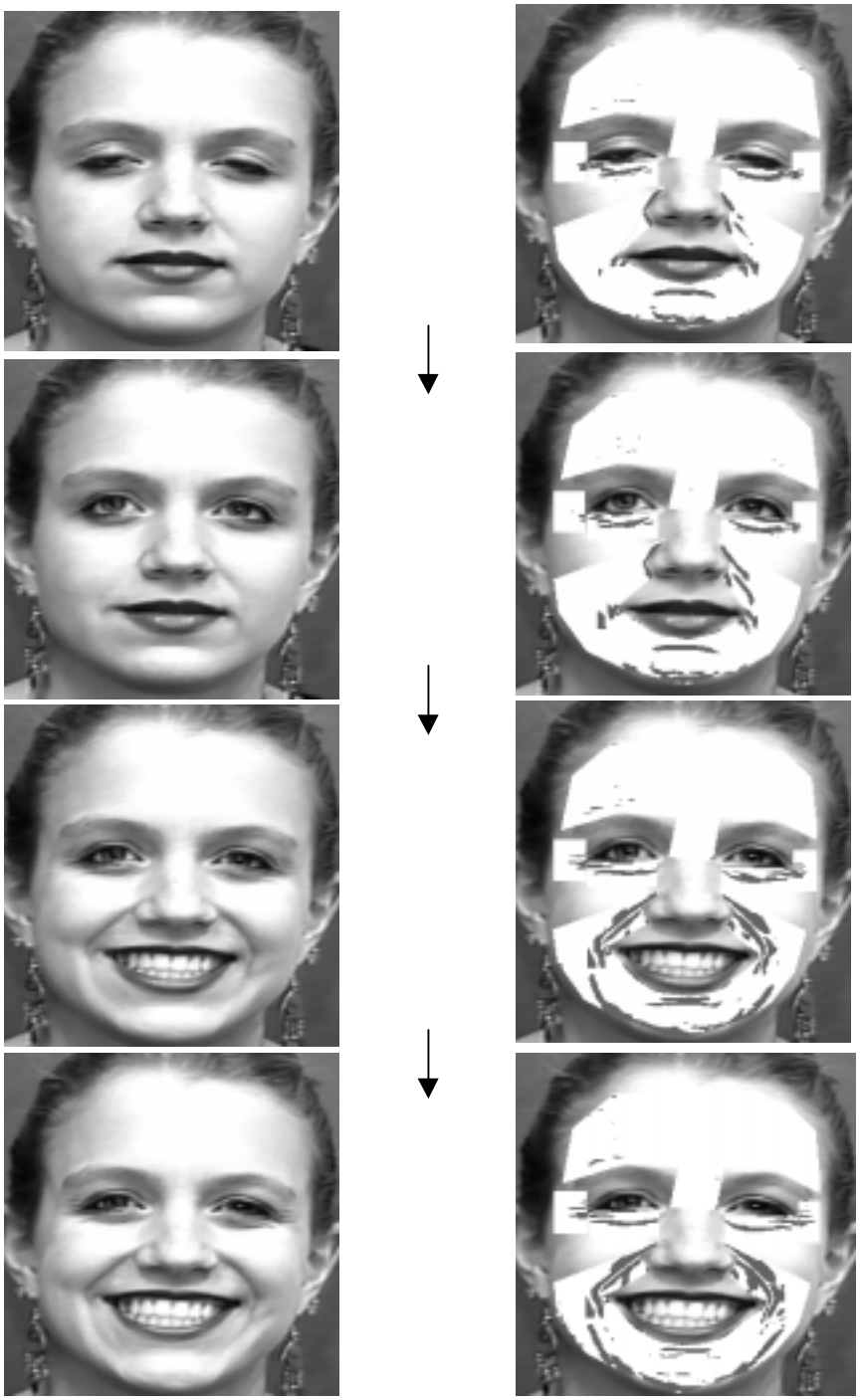
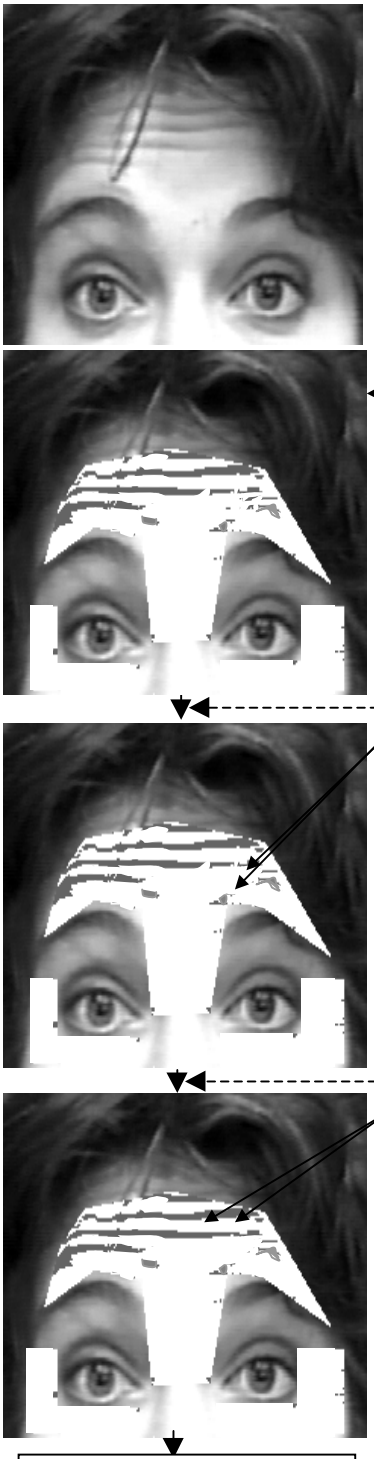


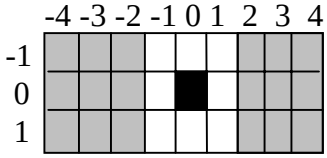
Figure 37.b Younger subjects, especially infants (Figure 37.a), show smoother furrowing than older ones (Figure 37.b), and initial expressions show weaker furrowing than that of peak expressions for each sequence.



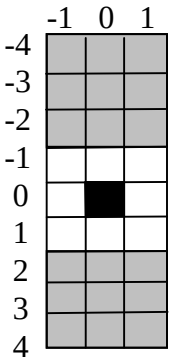
(a) Continue processing following Figure 34 & 35 (high gradient component analysis).

Template for erosion or dilation processing

Horizontal line template



Vertical line template



High gradient component

(b) Erosion processing – eliminating the small piece causing noise or the very short lines, thinning the lines, or smoothing the line boundary:
Consider the 8-connected component of the grey color region at each template, if there is no three-pixel connection of the horizontal (or vertical) direction at both regions after detection by the horizontal (or vertical) line template, then the center black pixel will be removed.

(c) Dilation processing – linking the broken lines:
Consider the 8-connected component of the grey color region at each template, if there is three-pixel connection of the horizontal (or vertical) direction at both regions after detection by the horizontal (or vertical) line template, then the center black pixel will be connected with both grey regions.

Figure 38.b: Connected Component Labeling.

Figure 38.a Delete redundant high gradient components using morphological transformation including erosion and dilation processings.

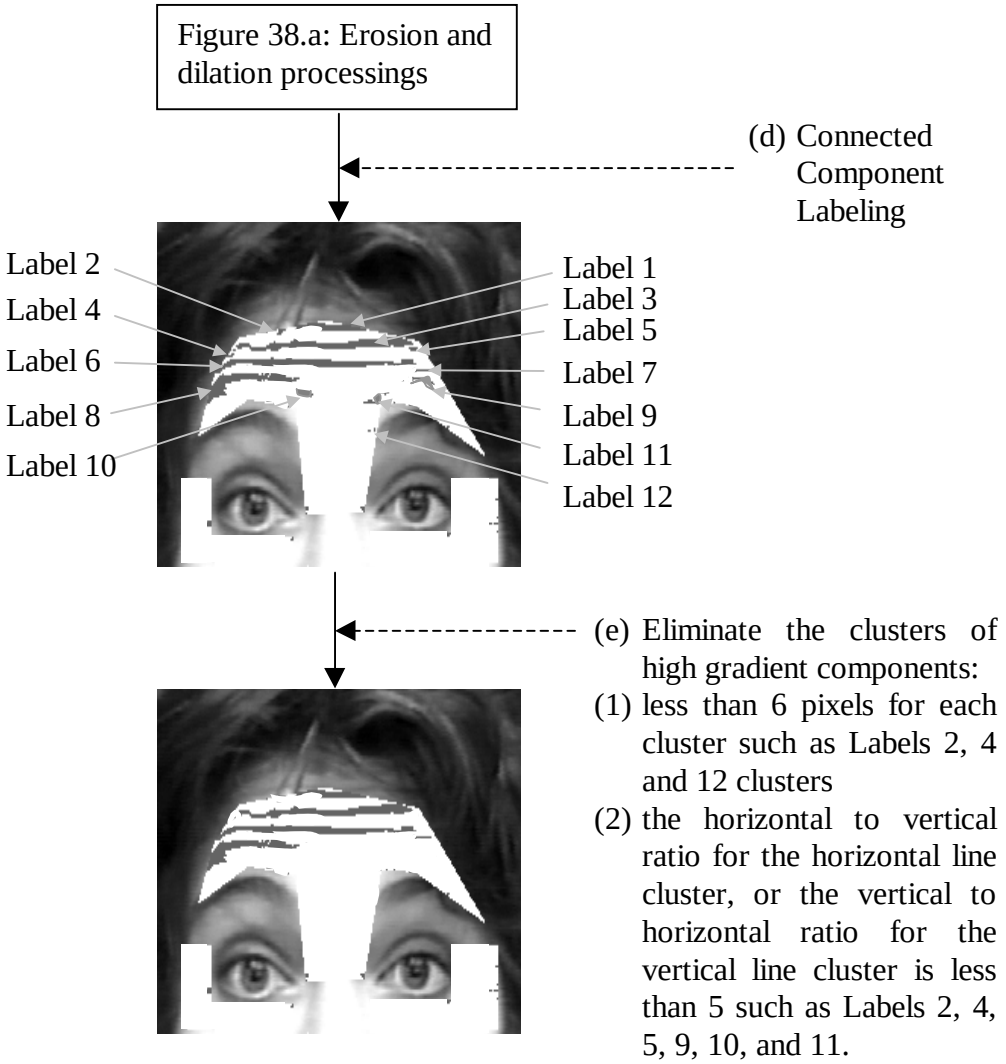


Figure 38.b Delete the redundant high gradient components using the connected component labeling algorithm.

5.4 Analysis of High Gradient Component Detection Problems

We have tried to use 3 x 3 Canny [19], Sobel [43], Prewitt [43] and wavelet-based [52] edge detectors to detect the high gradient components on the face image. We can not find any difference in the results.

Some high gradient components, such as the horizontal lines along the furrows at the forehead, are wide. If we use high gradient component detectors with small sizes, such as 3 x 3, to extract this line from vertical directions, which are perpendicular to this line or furrow direction, then this wide line will be detected into two lines (Figure 39). One way to solve this problem is to use line detectors with large sizes such as 3 x 5 or 5 x 7 in order to match the width and the length of the detected lines or furrows (Figure 39). In our approach, the 3 x 5, 5 x 3, and 5 x 5 horizontal, vertical, and diagonal line detectors, respectively, are adequate for most of the facial expression images. The larger size of detector requires more computation time.

The high gradient components (such as furrows) move in consecutive frames, so we need to consider the motion furrows in the spatio-temporal domain to ensure the detected high gradient components are produced by transient skin or feature deformations, and are not a permanent characteristic of the individual's face. Equation (5-5) is a way for tracking the motion high gradient component in the spatio-temporal domain. It is not accurate because of introducing the zeroing result: it is based on the tracking of individual pixels, which sometimes appear or disappear because of lighting or deformation of skin movement, or do not have any movement between consecutive frames because of a high sampling rate for an image sequence. To overcome above weakness, equation (5-6) can give a more reliable result by assuming the first (neutral expression) frame to be the background, then the foreground components. The motion of high gradient components can be extracted easily by subtraction instead of tracking processing from the remaining frames of each sequence.

According to equations (5-5) and (5-6), an interesting approach will be demonstrated by directly subtracting the gray values instead of gradient components: considering the

(a) Gray value image

(b) Using 3 x 3 horizontal line detectors

1	1	1
0	0	0
-1	-1	-1

-1	-1	-1
0	0	0
1	1	1

(c) Using 3 x 5 horizontal line detectors

1	1	1	1	1
0	0	0	0	0
-1	-1	-1	-1	-1

-1	-1	-1	-1	-1
0	0	0	0	0
1	1	1	1	1

Figure 39 Horizontal line (furrow) detection using different sizes of detectors. (b) If the size of the detector is too small compared with the width and length of the line (furrow), then each line will be extracted to two lines. (c) It is necessary to adjust the size of the line detector to match the width and length of the line in order to obtain the correct result.

motion effect in the temporal domain and ignoring the gradient effect in the spatial domain.

$$\frac{\partial I(x,y,t)}{\partial t} = I(x,y,t) - I(x,y,t-1) = \tilde{\Gamma}_t(x,y) \tag{5-7}$$

or

$$\frac{\partial I^0(x,y,t)}{\partial t} = I(x,y,t) - I(x-\Delta x, y-\Delta y, 0) = \tilde{\Gamma}_t^0(x,y) \tag{5-8}$$

where

$$-1 \le \Delta x \le 1 \quad \text{and} \quad -1 \le \Delta y \le 1$$

Next, a threshold is given to remove unnecessary gray-value components at which the absolute gray-value differences between the two images are below the threshold. This threshold process can compensate for the lacking of without considering the gradient factor in the spatial domain. According to our experiments, it works well to extract the teeth when mouth is opening. Since the gray values of the target (teeth) are obviously different from the background (skin, lips, and tongue), it is easy to segment the foreground from background using a simple threshold process (Figure 40). It is very difficult to extract the furrows exactly by using a simple thresholding process of different gray values between two face images, because the gray values of the entire face image are affected by lighting and are different across the ages and individuals of subjects. Since we define and can observe furrows, which are constituted from the motion high gradient components on face image sequence, it is necessary to extract the motion furrow by considering the spatial and temporal gradient components at the mean time.

5.5 Data Quantization and Conversion for the Recognition System

After the motion and high gradient components (furrows) are extracted in the spatio-temporal domain, the high gradient pixels are assigned a value 1 and other background components are assigned a value 0 for each facial expression sequence, we want to summarize the high gradient components of many pixels into a low-dimensional vector for each face image as an input to the recognition system. The forehead (upper face) and lower face regions of each normalized face image are divided into 16 and 16 blocks (Figure 41). The mean number of high gradient components in each block is calculated by dividing the number of pixels having a value of 1 by the total number of pixels in the block. The positional variance of high gradient components in each block is calculated as the sum of variances in the row and column directions. The mean number and positional variance per block discussed here are simply abbreviated as mean-variance for brevity. These give 32 parameters for 16 blocks in each of the upper and lower face regions. For

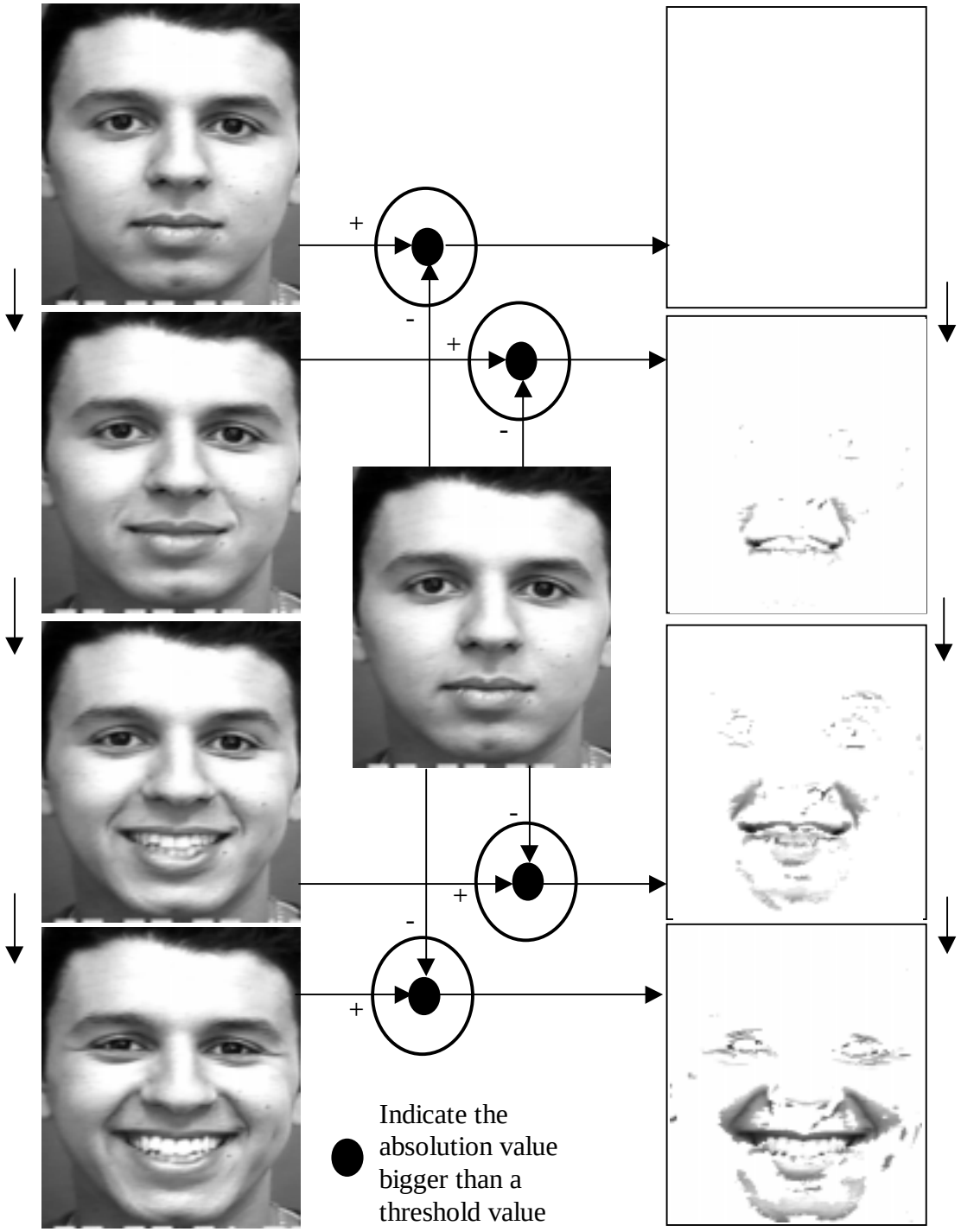
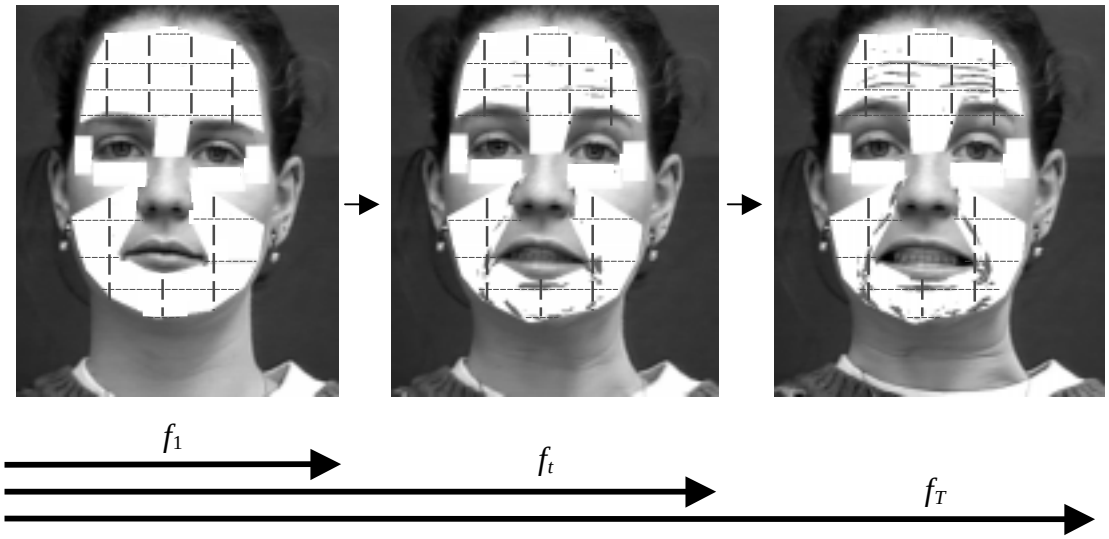


Figure 40 Teeth can be extracted directly from the subtraction of the gray value image at the current frame to that at first frame for each image sequence whose absolute value is larger than a constant threshold.



Mean-Variance vector $f_t = (m_{t,1}, m_{t,2}, \ldots, m_{t,j}, \ldots, m_{t,i}, \sigma_{t,1}, \sigma_{t,2}, \ldots, \sigma_{t,j}, \ldots, \sigma_{t,i})$

where $m_{t,j}$ is the mean number of high gradient components at the block j and frame t.

$\sigma_{t,j}$ is the positional variance of high gradient components at the block j and frame t.

and i = 16 (blocks) for the upper face region.

i = 16 (blocks) for the lower face region.

Mean-Variance vector sequence $f = (f_1, f_2, \ldots f_t, \ldots, f_T)$

where T is the length of this image sequence.

Figure 41 Mean-Variance vector of the high gradient component analysis in the spatio-temporal domain for input to the Hidden Markov Model.

Table 7 Sample symbol sequences for three upper facial expressions and six lower facial expressions under consideration.

AUs	**Motion Furrow Detection: Upper Facial Expressions (Symbol Sequence)**
0	0 0 0 0 *0* 0 0 0 0 *0* 0 0 0 0 *0* 0
4	0 0 0 0 *4* 4 10 10 10 *10*
1+4	0 0 0 12 *12* 12 12 12 6 *6* 6
1+2	0 0 0 0 *0* 12 9 9 3 *11* 11
AUs	**Motion Furrow Detection: Lower Facial Expressions (Symbol Sequence)**
12 (6+12+25)	10 10 10 10 *10* 10 1 1 1 *5* 5 5 5 5 *5* 5 5 5
9+17 (17+23+24)	10 10 10 10 *10* 10 12 12 12 *12* 12 4 4 4 *4* 4 4 4 4 *4* 4

upper and lower facial expression recognitions, these mean and variance values are concatenated to form a 32-dimensional mean-variance vector for each region in a frame. Table 7 shows sample symbol sequences for nine facial expressions under consideration. Such symbol sequences are used as inputs to the HMMs of upper facial expressions and lower facial expressions, respectively, for automatic recognition.

5.6 Expression Intensity Estimation

The sum of squared difference (SSD) criterion is employed to find a close estimation of furrow expression intensity. The furrow detection on each 417 x 385 image gives a 32-dimensional mean-variance vector for the upper facial expressions and similarly another 32-dimensional mean-variance vector for the lower facial expressions.

In the training data, the furrow expression intensity of individual frames of each facial expression sequence with length varying from 9 to 47 frames has been quantified by experts: from neutral expression (expression intensity: 0.0) to peak expression (expression intensity: 1.0). The corresponding mean-variance vector of each training frame has also been extracted. The Euclidean distance between the mean-variance vector of the testing frame to the mean-variance vector of the individual frame in the chosen training sequence is a measure of how close are their expression intensity values. After recognizing an input furrow expression sequence, the furrow expression intensity of an individual frame in the sequence is estimated by finding the best match of the furrow expression intensity from a training frame based on the shortest distance in the mean-variance space (Figure 42).

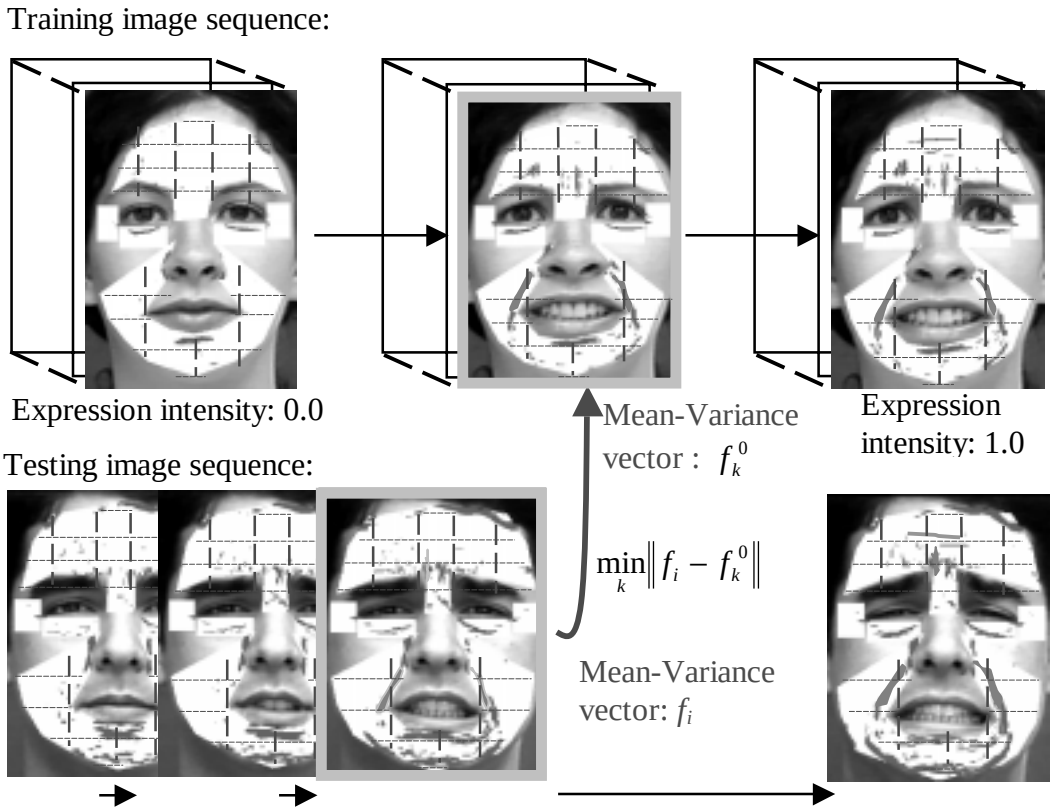


Figure 42 Furrow expression intensity matching by measuring the minimum value (distance) of the sum of squared differences (SSD) between the mean-variance vector of the known training image and that of the testing image. Each mean-variance vector of the training image corresponds to a given expression intensity value.