Journal of Zhejiang University SCIENCE A
ISSN 1009-3095 (Print); ISSN 1862-1775 (Online)
www.zju.edu.cn/jzus; www.springerlink.com
E-mail: jzus@zju.edu.cn

Multi-face detection based on downsampling and modified subtractive clustering for color images

KONG Wan-zeng†, ZHU Shan-an

(*School of Electrical Engineering, Zhejiang University, Hangzhou 310027, China*)

†E-mail: kwzzju@126.com

Received Jan. 9, 2006; revision accepted Aug. 2, 2006

Abstract: This paper presents a multi-face detection method for color images. The method is based on the assumption that faces are well separated from the background by skin color detection. These faces can be located by the proposed method which modifies the subtractive clustering. The modified clustering algorithm proposes a new definition of distance for multi-face detection, and its key parameters can be predetermined adaptively by statistical information of face objects in the image. Downsampling is employed to reduce the computation of clustering and speed up the process of the proposed method. The effectiveness of the proposed method is illustrated by three experiments.

Key words: Multi-face detection, Skin color, Modified subtractive clustering, Downsampling
doi:10.1631/jzus.2007.A0072 **Document code:** A **CLC number:** TP391

INTRODUCTION

Face detection has been widely used in fields such as security, multimedia retrieval, human computer interaction, etc. Therefore it becomes one of the most active research areas in computer science. Recently, approaches to face detection include neural network (Rowley *et al*., 1998), boosting (Viola, 2001; Viola and Jones, 2004), template matching (Kim *et al*., 2000) and skin color (Cai and Goshtasby, 1999; Wang and Yuan, 2001; Soriano *et al*., 2003), etc. The methods of neural network and boosting are based on the mechanism of learning. Therefore a great deal of faces should be collected for training. In template matching, several standard patterns of a face are stored to describe the face as a whole or the facial features separately. The correlations between an input image and the stored patterns are computed for detection. However, they are sensitive to variations of both pose and orientation. The method of skin color is a feature invariant approach which can work even when the pose, viewpoint, or lighting conditions vary; hence it becomes an important way in many applications from face detection to hand tracking.

Face detection in images always deals with multi-face rather than a single face. The aim of multi-face detection is to confirm the location and total number of faces in an image. In this paper, an approach based on downsampling and modified subtractive clustering is presented. In order to speed up the clustering algorithm, downsampling is used to reduce the number of pixels considerably. Aiming at face detection, a novel distance definition and adaptive parameter estimation are introduced in subtractive clustering in this paper. Thus, it is a fast and accurate method for multi-face detection. The proposed method in this paper consists of three parts: (1) downsampling images; (2) separating faces from the background with skin color detection; (3) applying modified subtractive clustering to locate faces.

IMAGE DOWNSAMPLING

Image downsampling which chooses a row (or column) from every T rows (or columns) in an image

is a way to condense images regardless of resolution decay. Let $f'(x,y)$ be the downsampled image. It can be formulated as

$$f'(x,y)=f(x,y)\sum_{i=0}^{+\infty}\sum_{j=0}^{+\infty}\delta(x-iT_x,y-jT_y), \quad (1)$$

where $f(x,y)$ is an original image and $\delta(x-iT_x, y-jT_y)$ is Dirac function with sampling intervals of column and row as T_x and T_y respectively. After downsampling, the total number of the pixels is reduced to $1/(T_x \cdot T_y)$ of the original.

SKIN COLOR DETECTION

Normalized RGB color model

It is very important to choose a suitable skin-color model which can be used to detect faces of different races, sexes and ages. Generally, colors of each pixel consist of R, G, B components with the range of [0, 1, 2, …, 255]. Consequently, the brightness value $I=R+G+B$. Research results (Kim and Kim, 1998) showed that human skin colors cluster in a small region in the RGB color space and differ more in brightness than in color. Therefore, the normalized RGB color space is considered to characterize human faces. Since the color information is very sensitive to the brightness value of each pixel, each color component can be normalized by the brightness value I as follows

$$r=R/I,\ g=G/I,\ b=B/I, \quad (2)$$

where $r+g+b=1$. Note that $b=B/I$ is no longer unique since $b=1-r-g$. Thus, the normalized RGB color space can be represented only with r and g.

HSV color model

It is commonly recognized that the HSV (hue, saturation and value) model (Gonzalez and Woods, 2003) is more similar to the human perception of color than the above normalized RGB model. The hue (H) is a measure of the spectral composition of a color and is represented as an angle varying from 0° to 360°. The saturation (S) refers to the purity of colors, which varies from 0 to 1. The darkness of a color is defined by the value (V), which also ranges from 0 to 1. The HSV color model can be converted from the RGB model using the following equations

$$H=\begin{cases}\theta, & B\le G,\\ 360^\circ-\theta, & B>G,\end{cases} \quad (3)$$

with

$$\theta=\arccos\left[\frac{0.5(R-G)+0.5(R-B)}{\sqrt{(R-G)^2+(R-B)(G-B)}}\right], \quad (4)$$

$$S=1-\frac{3\cdot\min(R,G,B)}{R+G+B}, \quad (5)$$

$$V=\frac{1}{3}(R+G+B). \quad (6)$$

It is assumed that the RGB values have been normalized to the range [0, 1] and that the angle θ is measured with respect to the red axis of the HSV space. Hue can be normalized to the range [0, 1] by dividing it by 360° with all values resulting from Eq.(3). The other two HSV components are already in the range [0, 1] if the given RGB values are normalized.

Color model selection

Though the normalized RGB color model can reduce the influence of the intensity, it is still sensitive to the intensity because of the saturation without separation. Therefore, in this paper, both normalized RGB model and normalized HSV model are used to detect faces. For an experiment, 14892 facial pixels of Asian race were selected to test the range of distribution in normalized RGB space and HSV space. The results are shown in Fig.1 and Fig.2, respectively.

Based on the test, the parameters are chosen as follows

$$0.4\le r\le 0.6,\ 0.22\le g\le 0.33,\ r>g>(1-r)/2, \quad (7)$$

$$0\le H\le 0.2,\ 0.3\le S\le 0.7,\ 0.22\le V\le 0.80. \quad (8)$$

Conditions (7) and (8) can be used to detect skin color pixels. Therefore, a binary image can be obtained from the color image which contains human faces through skin color detection. The binary image $f_b(x,y)$ can be defined as

$$f_b(x,y)=\begin{cases}1, & \text{if pixel is skin color},\\ 0, & \text{else}.\end{cases} \quad (9)$$

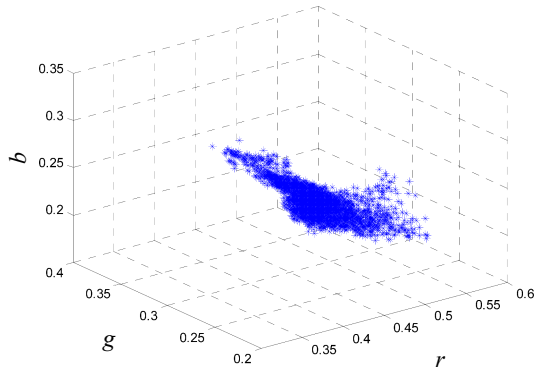


Fig.1 Skin color clustering in normalized RGB space

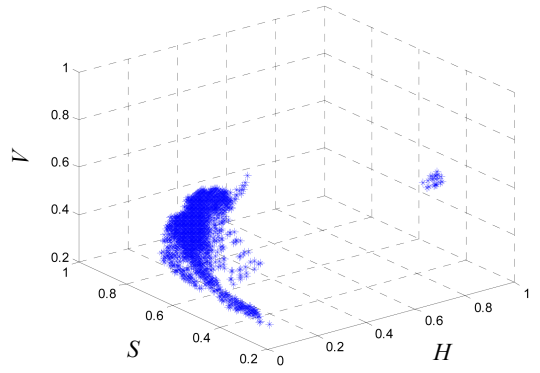


Fig.2 Skin color clustering in normalized HSV space

ADAPTIVE SUBTRACTIVE CLUSTERING IN FACE DETECTION

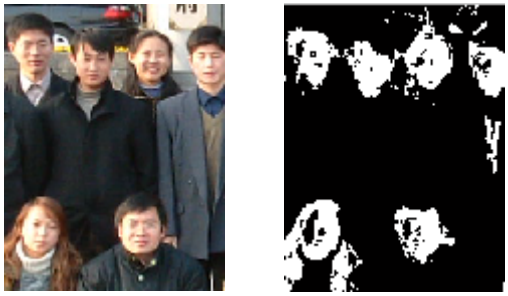
Modified subtractive clustering

Generally, no prior information tells us the positions and the number of faces in an image. Consider an original image as shown in Fig.3a which contains six objects (faces) to be located. After skin color detection, a binary image can be obtained. Though the regions of faces are separated from the background as shown in Fig.3b, it is still difficult for a computer to confirm the location and the number of faces. Therefore, the method of subtractive clustering (Chiu, 1994) is employed to solve this problem.

Define M and N as the height and width of the image, respectively. The data to be clustered are a 2D data collection

$$\{(x_1,y_1), (x_2,y_2), \ldots, (x_n,y_n)\},$$

where n is the number of skin color pixels, $f_b(x_i,y_i)=1$,



(a) (b)

Fig.3 Experiment of skin color detection
(a) Original image; (b) Binary image

$x_i \in [1, 2, \ldots, N]$ and $y_i \in [1, 2, \ldots, M]$. Without loss of generality, we assume that the data points have been normalized in each dimension so that their coordinate ranges in each dimension are equal, i.e., the data points are bounded by a hypercube. The normalized data collection is

$$\{(x_1',y_1'),(x_2',y_2'),\ldots,(x_n',y_n')\},$$

where $x_i'=x_i/N$ and $y_i'=y_i/M$.

Let $\boldsymbol{d}_i$ be a data point denoted as $\boldsymbol{d}_i=[x_i' \ y_i']^{\mathrm{T}}$, then the distance can be defined by

$$||\boldsymbol{d}_i-\boldsymbol{d}_j||=[(\boldsymbol{d}_i-\boldsymbol{d}_j)^{\mathrm{T}}\boldsymbol{A}(\boldsymbol{d}_i-\boldsymbol{d}_j)]^{1/2}, \quad (10)$$

where the matrix $\boldsymbol{A}$ determines the types of distance.

Each data point is considered as a potential cluster center and a measure of the potential of the data point $\boldsymbol{d}_i$ proposed by Chiu (1994) is defined as

$$P_i=\sum_{j=1}^{n}\exp(-4||\boldsymbol{d}_i-\boldsymbol{d}_j||^2/r_a^2), \quad (11)$$

where the matrix $\boldsymbol{A}$ in distance is $\boldsymbol{A}=\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}$ and r_a is an effective radius defining a neighborhood. Data points outside this radius have little influence on the potential. In other words, the original algorithm takes the same effective radius for all dimensions. However, different objects have different shapes, and for a face, for instance, its height is about 1.2 times as long as its width. In this paper, considering the shape of faces, the effective radiuses with respect to each dimension should be different. Thus, r_{ax} and r_{ay} are the effective

radiuses of respective dimensions. The potential of a data point in this paper can be denoted as

$$P_i=\sum_{j=1}^{n}\exp(-4\,||\,\boldsymbol{d}_i-\boldsymbol{d}_j\,||^2/r_{ax}^2), \qquad (12)$$

where the matrix $\boldsymbol{A}$ in Eq.(10) is

$$\boldsymbol{A}=\begin{pmatrix}1 & 0\\ 0 & r_{ax}^2/r_{ay}^2\end{pmatrix}.$$

Each constant r_{ax} and r_{ay} is the effective radius defining a neighborhood corresponding to its dimension. A point with many neighboring data points will have a high potential value.

After the potential of every data point has been computed, we select the data point with the highest potential as the first cluster center, which will be considered as the location center of the first detected face. Let $\boldsymbol{d}_1^*$ be the location of the first cluster center and P_1^* be its potential value, the potential of each data point $\boldsymbol{d}_i$ can be revised by

$$P_i \Leftarrow P_i-P_1^*\exp(-4\,||\,\boldsymbol{d}_i-\boldsymbol{d}_1^*\,||^2/r_{bx}^2). \qquad (13)$$

The distance $||\,\boldsymbol{d}_i-\boldsymbol{d}_1^*\,||$ in Eq.(13) has the same definition as Eq.(10) with

$$\boldsymbol{A}=\begin{pmatrix}1 & 0\\ 0 & r_{bx}^2/r_{by}^2\end{pmatrix},$$

where r_{bx} and r_{by} have the same effect as the constants r_{ax} and r_{ay} which define the neighborhoods that will have measurable reductions in potential. Thus, a potential is subtracted from each data point as a function of its distance from the first cluster center. The data points near the first cluster center will have greatly reduced potential value, and therefore are unlikely to be selected as the next cluster center. In order to avoid closely spaced cluster centers, r_{bx} and r_{by} should be somewhat greater than r_{ax} and r_{ay} accordingly. A reasonable choice is r_{bx}=1.5r_{ax} and r_{by}=1.5r_{ay}.

When the potential of all data points have been revised according to Eq.(13), the remaining highest potential is selected as the second cluster center. We then further reduce the potential of each data point with the second cluster center. In general, after the kth cluster center has been obtained, the potential of each data point can be revised by

$$P_i \Leftarrow P_i-P_k^*\exp(-4\,||\,\boldsymbol{d}_i-\boldsymbol{d}_k^*\,||^2/r_{bx}^2), \qquad (14)$$

where $\boldsymbol{d}_k^*$ is the location of the kth cluster center and P_k^* is its potential value.

The process of acquiring new cluster center and revising potential is repeated until

$$P_k^* < \xi P_1^*, \qquad (15)$$

where ξ is a small fraction. This is the end condition of mountain clustering (Yager and Filev, 1994) in stead of original subtractive clustering, for it possesses a clearer physical meaning in face detection in images which will be introduced in the following section.

Parameter selection

By now, there are three parameters, i.e., r_{ax}, r_{ay} and ξ to be predetermined, which exert an important influence on the result of face detection. On one hand, if r_{ax} and r_{ay} are too small, it always results in finding an uncompleted object. On the other hand, if r_{ax} and r_{ay} are too large, a cluster may contain more than one object. Parameter ξ has an effect on the number of the objects to be clustered. In original subtractive clustering, these parameters are selected by experience. For example, r_a is suggested to be a value between 0.2 and 0.5, and ξ is usually between 0.15 and 0.50. However, the proposed values of the parameters do not always work well in face detection. How to estimate the values of the parameters adaptively? We assume that the statistics information of the objects (faces) which need to be detected in the image of a surveillance system can be obtained approximately as follows:

(1) Screen a single man walking frontally from far to the camera.

(2) Select S images evenly from a sequence of images.

(3) Find the deviations of the boundary of the object o^i to its center as σ_x^i and σ_y^i which are normalized in the range [0, 1], where $1\leq i\leq S$. Accordingly, $2\sigma_x^i$ and $2\sigma_y^i$ are the normalized weight and height of the object. Then σ_x and σ_y are two collections

where $\sigma_x^i \in \sigma_x$, and $\sigma_y^i \in \sigma_y$. The definition of data potential is based on Gaussian PDF which can be denoted as

$$P(x) = \exp[-(x-\mu)^2/(2\sigma^2)]. \quad (16)$$

Combining the Gaussian function with Eq.(12), we have the conclusion (Hadjili and Wertz, 2002)

$$E(\sigma_x) = \sqrt{r_{ax}^2/8}, \quad E(\sigma_y) = \sqrt{r_{ay}^2/8}. \quad (17)$$

According to Eq.(17), the effective radius of each dimension is computed as

$$r_{ax} = 2\sqrt{2}E(\sigma_x), \quad r_{ay} = 2\sqrt{2}E(\sigma_y). \quad (18)$$

The end condition of clustering depends on Eq.(15). Recall that P_1^* is the highest potential of the data which has the most neighboring data points while the center data possesses the least neighboring data points with the potential P_k^*. In other words, the larger the object, the higher the potential will be. Based on that, the parameter ξ can be calculated as

$$\xi = \beta \frac{\min(\sigma_x^i \sigma_y^i)}{\max(\sigma_x^i \sigma_y^i)}, \quad (19)$$

where β is a security coefficient and is selected as 0.8 in this paper.

Reducing the clustering data

The speed of the subtractive clustering mainly lies in the number of the data points. Downsampling is an effective approach to reduce the number of clustering data. There are two multi-face detection methods in this paper, each having the same detection procedures, but with different downsampling step orders.

Method A can be described as follows:

Step 1: downsample the original image;

Step 2: detect faces in the downsampled image with skin color;

Step 3: locate faces with the proposed subtractive clustering.

Method B can be described as follows:

Step 1: detect faces in a color image with skin color;

Step 2: downsample the binary image;

Step 3: locate faces with the proposed subtractive clustering.

Comparing Methods A with B, we find that Method A is faster, as it puts the downsampling as the first step and thus reduces the data both in skin color detection and clustering. Furthermore, the results of the two methods are the same that can be illustrated by the binding rule of composite mapping. Define

$$f:\mathbb{R}^2 \to \mathbb{R}^3, \ g:\mathbb{R}^3 \to \mathbb{R}^2, \ h:\mathbb{R}^2 \to \mathbb{R}^2,$$

where f is a mapping of color image, g is a mapping of skin color detection, h is a mapping of downsampling, $\mathbb{R}^2$ is the space of (x,y), and $\mathbb{R}^3$ is the space of (R,G,B).

Theorem 1 If $f:X \to Y$, $g:Y \to Z$, $h:Z \to A$, then $(h \circ g) \circ f = h \circ (g \circ f):X \to A$ holds.

$(h \circ g)$ takes downsampling as the first step while $(g \circ f)$ takes skin color detection as the first, but both of them have the same results on the basis of Theorem 1. Thus Method A is selected to reduce the computation for the detection.

EXPERIMENT

To illustrate the good performance of the proposed algorithm in this paper, three experiments were carried out. The first one involves the comparison of the two methods in which the downsampling steps are in different orders. The second one is to show the resolution-precision relationship of face detection in downsampled images. The final experiment shows the intuitionistic performance of the proposed algorithm.

Experiment 1 This experiment is performed on the color image shown in Fig.3a. The sampling intervals of the 2D image are selected as T_x=T_y=3. In other words, the downsampled image is reduced to 1/9. The statistics information of faces in the image can be found through the method introduced in the subsection of "Parameter selection" as

$$E(\sigma_x) = 0.1, \quad E(\sigma_y) = 0.11, \quad \xi = 0.33. \quad (20)$$

Based on Eqs.(17)~(19), the parameters of the proposed clustering are estimated as r_{ax}=0.2828, r_{ay}=0.3111, and ξ=0.33. Comparison of the two methods in face detection are shown in Table 1, where N is the number of faces, $\boldsymbol{C}_i$ ($1 \le i \le N$) are the centers of each face, and T is the cost time in detection. Table 1 shows that the two methods have different computation times. It justifies the conclusion described in the subsection of "Reducing the clustering data".

Experiment 2 Downsampling reduces the computation time but decays the resolution of the image. The larger the interval of downsampling, the worse the resolution of the image will be. Generally, decayed resolution of the image always results in finding deviated centers of the objects through the proposed method. Accordingly, the purpose of this experiment is to find a suitable downsampling interval which will reduce the computation but still retain a good performance of face detection. Nine sub-experiments are tested to verify the computation time-sampling interval relationship and the detection error-sampling interval relationship.

Select $T_x^{(i)} = T_y^{(i)} = i$ in Eq.(1), where i is from 1 to 9 in sequence of sub-experiments. When i is 1, i.e., in the first sub-experiment, it is an original image without any resolution decay. The measure of detection error of each sub-experiment with the original is a function of the deviation distance of all detected centers which is denoted as

$$E^{(i)} = \left[\frac{1}{N} \sum_{j=1}^{N} \langle (i \cdot \boldsymbol{C}_j^{(i)} - \boldsymbol{C}_j^{(1)}),\ (i \cdot \boldsymbol{C}_j^{(i)} - \boldsymbol{C}_j^{(1)}) \rangle \right]^{1/2}, \quad (21)$$

where $\boldsymbol{C}_j^{(i)}$ is the jth detected center in the ith sub-experiment and < , > is an operation of inner product. From the result shown in Fig.4, we find that the time cost is rapidly reduced when sampling interval varies from 1 to 2, and then reduced slowly as the sampling interval increases. From the result shown in Fig.5, the detection error increases when the sampling interval increases. Therefore, it is reasonable to choose

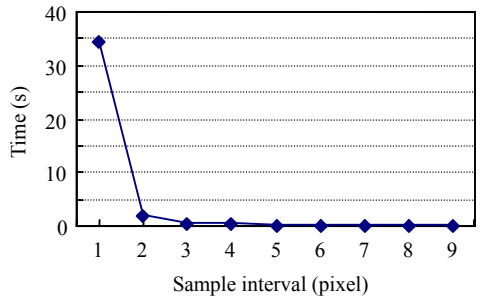


Fig.4 Relationship between cost time and sample interval

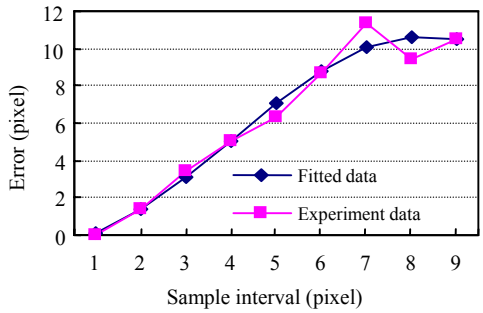


Fig.5 Relationship between detection error and sample interval

T_x=T_y=2 or 3 in order to obtain good performance while reducing a mass of computation.

Experiment 3 About 100 images were tested by the proposed algorithm in this experiment. Some faces are inclined and some have different scales. Parameters r_{ax}, r_{ay} and ξ are estimated according to the subsection of "Parameter selection", and sampling intervals T_x, T_y are both selected as 3. The proposed method can detect rotated faces with different sizes, as the skin color is not so sensitive to the orientation and scale. The results prove its effectiveness. Some results are shown in Fig.6.

CONCLUSION

This paper presents a multi-face detection method for color images. Because the parameter estimation and distance definition are both based on objects, the modified subtractive clustering gains

Table 1 Comparison of Method A and Method B

Method	N	$\boldsymbol{C}_1$	$\boldsymbol{C}_2$	$\boldsymbol{C}_3$	$\boldsymbol{C}_4$	$\boldsymbol{C}_5$	$\boldsymbol{C}_6$	T (s)
A	6	(16,42)	(70,11)	(16,11)	(67,41)	(20,61)	(20,26)	1.5120
B	6	(16,42)	(70,11)	(16,11)	(67,41)	(20,61)	(20,26)	2.2130

Fig.6 Some results of Experiment 3

quite good performance if faces can be well separated from the background. The restriction of the subtractive clustering used in image processing is dealt with by downsampling. Moreover, the order of downsampling in the proposed method is illustrated. It is important to choose a reasonable sampling interval to balance the precision and the computation in face detection. Furthermore, the proposed method can be applied to other object detection as well.

References

Cai, J., Goshtasby, A., 1999. Detecting human faces in color images. *Image and Vision Computing*, **18**(1):63-75. [doi:10.1016/S0262-8856(99)00006-2]

Chiu, S.L., 1994. Fuzzy model identification based on cluster estimation. *Intell. Fuzzy Syst.*, **2**:267-278.

Gonzalez, R.C., Woods, R.E., 2003. Digital Image Processing (2nd Ed.). Electronics Industry Press, Beijing, p.282-302 (in Chinese).

Hadjili, M.L., Wertz, V., 2002. Takagi-Sugeno fuzzy modeling incorporating input variables selection. *IEEE Trans. Fuzzy Syst.*, **10**(6):728-742. [doi:10.1109/TFUZZ.2002.805897]

Kim, S.H., Kim, H.G., 1998. Facial region detection using range color information. *IEICE Trans. Inform. Syst.*, **E81-D**(9):968-975.

Kim, H., Kang, W., Shin, J., Park, S., 2000. Face detection using template matching and ellipse fitting. *IEICE Trans. Inform. Syst.*, **E38-D**(11):2008-2011.

Rowley, H.A., Baluja, S., Kanade, T., 1998. Neural network-based face detection. *IEEE Trans. Pattern Analysis and Machine Intelligence*, **20**(1):23-38. [doi:10.1109/34.655647]

Soriano, M., Martinkauppi, B., Huovinen, S., Laaksonen, M., 2003. Adaptive skin color modeling using the skin locus for selecting training pixels. *Pattern Recognition*, **36**(3):681-690. [doi:10.1016/S0031-3203(02)00089-4]

Viola, P., 2001. Rapid Object Detection Using a Boosted Cascade of Simple Features. Proc. IEEE Conference on CVPR, p.511-518.

Viola, P., Jones, M.J., 2004. Robust real-time face detection. *International Journal of Computer Vision*, **57**(2):137-154. [doi:10.1023/B:VISI.0000013087.49260.fb]

Wang, Y.J., Yuan, B.Z., 2001. A novel approach for human face detection from color images under complex background. *Pattern Recognition*, **34**(10):1983-1992. [doi:10.1016/S0031-3203(00)00119-9]

Yager, R.R., Filev, D.P., 1994. Generation of fuzzy rules by mountain clustering. *Intell. Fuzzy Syst.*, **2**:209-219.