Objective Saliency and Facial Models for Face Detection in Super-Resolution

Oh,Seong-Gi
The graduate school of
Advanced image Science,
Multimedia & Film
Chung-Ang University
Seoul, Korea
osg7412@cau.ac.kr

Kim,Tae-Yong*
The graduate school of
Advanced image Science,
Multimedia & Film
Chung-Ang University
Seoul, Korea
tykim@cau.ac.kr

Jang,Chang-Young
The graduate school of
Advanced image Science,
Multimedia & Film
Chung-Ang University
Seoul, Korea
codemist@cau.ac.kr

James Rwigema
The graduate school of
Advanced image Science,
Multimedia & Film
Chung-Ang University
Seoul, Korea
Rwigema1@cau.ac.kr

***Abstract*—This paper suggests the solution to fast processing in high resolution image. Viola-Johns face detector known as Adaboost takes long time because of too much pixel information. So we build up the method to predicting facial ROI in down-sampled image based on GMM(Gaussian Mixture Model). The color space used to find the ROI is the Color Opponent Space constituted by O1, O2, O3 instead of R, G, B for intensity-invariant processing, which performs better than HSV space. The 3 types of also invariance facial models and the Saliency based color segmentation are our critical idea to better face detection performance. As a result, it performed better detection speed than Adaboost and ROI ratio than GMM using HSV color space.**

Keywords: Image processing, Machine learning, Math

I. INTRODUCTION

Viola-Jones detection[1] based on the Adaboost Classifier[1] provided by Open-cv is too slow for real-time processing of high-resolution images. The basic Open-cv Viola-Jones face detector [1] based on the Haar-like feature, the average speed was 3909ms per a frame on the 4k (3840 x 2160) image (Intel i5-4460 3.20GHz CPU). To reduce calculations required for face detection in a high-resolution image, we estimated the position of the face region in the degraded image. Human skin color has special distributions in the color space because of types of race, specification of lighting, direction of lights and etc. So it needs a difference prediction between original region and averaged color of facial region in degraded image. In this paper, we propose facial-model for face position estimation in more various case by applying above methods and estimate facial position more accurately in down-sampled image using modified GMM (Gaussian Mixture Model)[2] and frequency-tuned Saliency[3].

I. BACKGROUND

4k as noted high-resolution, an average of 25ms is required for the pixel by pixel addition command based on OpenCV 3.4. It takes too long time for standardized film framerate 23.976. For this quality of real-time processing, it requires less than 41.71 ms per frame.

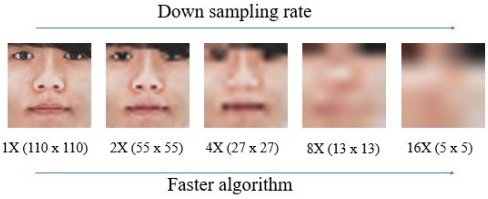

Figure 1 : How down-sampling rate works in a face image.

As shown in the Figure 1, it is difficult to estimate the original state as the size becomes smaller, but the number of pixels to be calculated decreases proportionally to square root . So we use followings to estimate position of faces in degraded data.

A. *Viola-Jones object detection*

Viola-Jones detection is a method that finding special object in an image using Harr-like features. This detector as knows as Boosting and Ada-boost trained by complexed weak characters about target object. Cheng and Lakemond introduced the idea of minimum face size to estimate the face location in the down-sampled image [4].

B. *Skin Color segmentation[5]*

Our method requires facial region estimation of faces from degraded image.

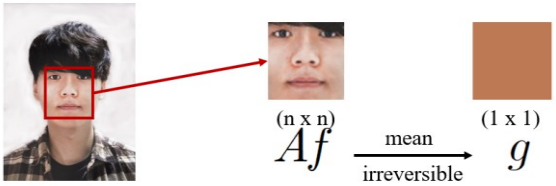

Figure 2: The irreversibility of mean function.

Figure 2 is an output of distorted input where down-sampling is n, f is an original image, g is a mean. This is equal to the mean of the input and the calculation is irreversible because there are no matrices that extend the dimension of the matrix A^{-1}. In this case, skin color segmentation is one of the best.

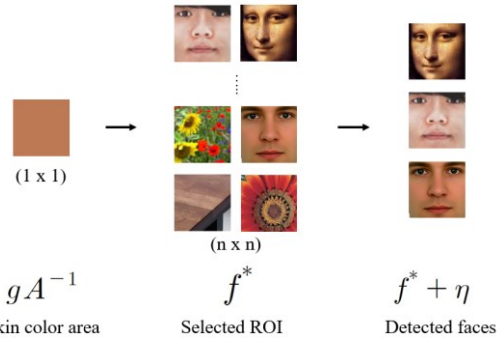

Figure 3 Overview of proposing method

Figure 3 shows the skin color range can be specified as the ROI for the regions. An averaged pixel value gA^{-1} can

contains the other information figured in above. So, it needs more steps for detecting faces only.

C. *Color space[6]*

In order to predict facial regions on the averaged pixel, we chose Color Opponent Space to select Invariant Color space for illumination change[6]. Intensity-invariant color spaces include Hue (HSV) and O1 + O2 (Opponent). Paper [6] compared the differences between the two and compared the performance according to the number of descriptors detected. Hue for Light Intensity, Opponent for Color-shift, arrangement change and viewpoint change. Color Opponent Space shown the best result in tests considering the intensity rather than Hue. RGB to Color Opposite transform is noted as eq.1

$$\begin{pmatrix} O1 \\ O2 \\ O3 \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 \\ \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & \frac{-2}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \end{pmatrix} \begin{pmatrix} R \\ G \\ B \end{pmatrix}$$

eq. 1

*O*1 presents the difference between *R* and *G*, O2 is RG and B and O3 is the intensity.

D. *Resizing Method*

In the paper[5] introduces the minimum face size in an image for reducing size. They set the resolution of face regions for detecting as smaller as detectable. This down-sampling was performed using grid-sampling. If the skin color is detected in the grid, its surroundings are designated as ROI. This is not a suitable method that causes a false negative when the non-skin part of the face (eye, nostril, etc.) on the grid is laid on the grid.

II. FACIAL COLOR MODELING

A. *Training to get common standard deviation*

Since we have introduced a Gaussian distance to specify the range of skin color, we use a lot of training datasets to obtain a universal sigma to obtain an estimate of the standard deviation (σ) and use it as the basis for estimation.

B. *Facial model*

We suggest a facial model for estimate faces in degraded image. Degraded region includes the average of original pixels only. So, we have to cosider following problems that images of a person's face can be confused by lighting condition.

Figure 4 : Results of averaged face region.

Figure above shows 5 different mean colors of one person's face after 5 different luminatic specifications. Assuming that there are k variations, the detector trained by the datasets that include the lighting condition and others (races, skin tones, hairs, glasses and etc.). The gaussian texture for each facial model is mixed to the detector can catch all cases. k is number of face models. Our selected k is the 3 for best performance.

C. *Objectiv color Saliency*

An arbitrary estimate of the irreversible restoration equation is needed. Trained data are used to estimate which class the sample belongs to. In this case, the Gaussian Mixture Model is widely used. Mixtures of distances defined as eq. 2

$$P(x) = W_1 P_1(x) + W_2 P_2(x) + \ldots + W_k P_k(x)$$

eq. 2

W is weighting value of each distances.

$$W_1, W_2, \ldots, W_k$$
$$\sum_{i=1}^{k} W_i = 1 \quad \text{and} \quad W_i \geq 0$$

eq. 3

And P(x) is gaussian distance between x and μ, where σ is the standard deviation, x the sample's value and μ the mean.

$$P_i(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_i-\mu_i}{\sigma}\right)^2}$$

eq. 4

In our method, it needs to mix for including all range of facial models. We set the k as the number of facial models, x is the samples and u the central color of each models.

D. *Estimation of vairations*

We set the k = 3 as initial value after 3 facial models.

- Model 1. Whole shadows on the face.

- Model 2. Partial shadows on the face.

- Model 3. No shadows on the face.

Each case is just abstractive models because an averaged image of facial models is the flexible information as status of training data. This k that the number of models will be the number of Gaussians in GMM since its characteristics. The mean color of each classes is set as target color q. And σ is estimated by getting standard deviation of samples.

Weighting factor W of each model is determined by maximizing total energy noted as

$$E_{total} = E_1 + E_2 + ... + E_k$$

eq. 5

Where n is the number of samples, the energy of a model defined as

$$E_i = W_i P(C\,;\,C_i^*) = W_i \sum_{j=0}^{n} \frac{1}{\sigma_i \sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{D(C_i, C_i^*)}{\sigma_i}\right)^2}$$

eq. 6

C* is the mean color vector and C the color vector of all samples and W the weighting factor that can be changed by the training sets. The color difference between samples and trained data converted form of RGB to Opponent space is

$$D(C, C^*) = \| \overrightarrow{C^* - C} \|$$

$$where \quad \overrightarrow{C} = \begin{pmatrix} O_1 \\ O_2 \\ O_3 \end{pmatrix} \qquad \overrightarrow{C^*} = \begin{pmatrix} O_1^* \\ O_2^* \\ O_3^* \end{pmatrix}$$

eq. 7

III. Objective Color Saliency

We combined saliency to show energy differences in each region. Saliency equation in the 2D image is denoted by

$$S(x,y) = \| I_u - I_{whc} \|_2 = \sum_{I_{i,j} \in U} \| I_{x,y} - I_{i,j} \|_2$$

eq. 8

Where U is a set of pixels around $I_{x,y}$. Since Saliency is the energy that containing differences with neighborhoods, the energy between color of a pixel and facial models can be written as

$$S(x,y) = \sum_{C_i^* \in U} \left| \overrightarrow{C_{x,y} - C_i^*} \right|_2$$

eq. 9

And U is a set contains all facial model. This Saliency means the energy of color difference. The sigle pixel's distance with mixtures can be rewritten using Saliency energy instead of numerical distance,

$$E_{total} = \sum_{i=1}^{k} W_i P_i(C\,;\,C_i^*) = \sum_{i=1}^{k} W_i \frac{1}{\sigma_i \sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}\left| \overrightarrow{C - C_i^*} \right|_2}$$

eq. 10

This equation can be expanded as the definition of Color Opponent Space.

$$P(C\,;\,C^*) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}|O_1 - O_1^*|_2} e^{-\frac{1}{2\sigma^2}|O_2 - O_2^*|_2}$$

eq. 11

It means the exponent of an error between samples and models is constructed multiplication that absolute values of each color channels. The reason why we don't use O3 channel is avoiding the intensity channel. So Objective color Saliency is can be computed simply like eq.11 and defined as

$$S_{obj}(x,y) = \sum_{O_1^*, O_2^*, \sigma_*^2 \in U} \frac{1}{\sigma_* \sqrt{2\pi}} e^{-\frac{1}{2\sigma_*^2}\left| \overrightarrow{O_1(x,y) - O_1^*} \right|_2} e^{-\frac{1}{2\sigma_*^2}\left| \overrightarrow{O_2(x,y) - O_2^*} \right|_2}$$

eq. 12

IV. Performance Evaluation

The evalouation of proposed is tested without GPU computing likes CUDA. Tested CPU is Intel i5-4460 3.20GHz CPU and we used the OpenCV 3.4.1 Viola-Jones object detector. Its critical variable is preprocessing because of except for preprocessing, all of test environment is the same.

A. *Comparing with other method*

Figure 5 : Graphs for comparison with the method [5] and effects of facial ROI ACCURACY BY DOWN-SAMPLING RATE (%)

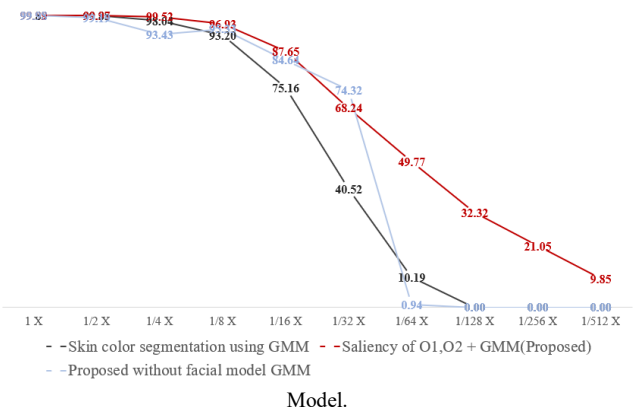

Figure 5 shows the accuracy of each 3 color segmentation methods as down-sampling rate. Our Proposed method performs the best of them because of using Color Opponent Space instead of HSV. The definition of used accuracy is

$$\frac{1 - (ROI_{1X} - ROI_{nX})}{ROI_{1X}} * 100(percentage)$$

eq. 13

It means the ROI ratio between 1X(original image) and nX(down-sampled image as $r=2^n$)

B. *Accuracy/ time ratio as down-sampling rate*

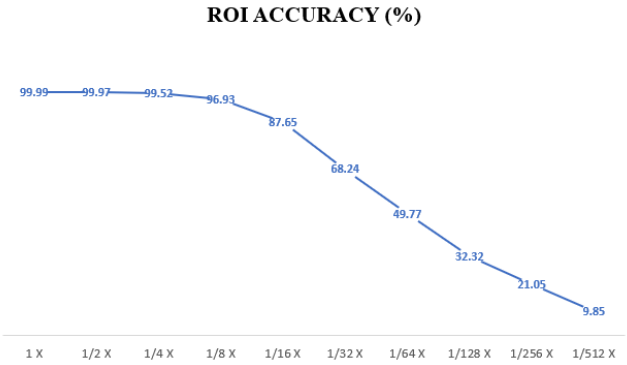

Figure 6 : Graph about the accuracy as down-sampling rate

Figure 6 shows the single accuracy noted in Equation above of proposed method. Its accuracy is going far worse from 16X rate.

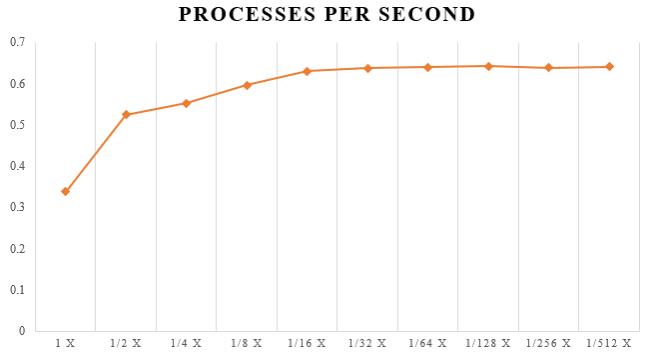

Figure 7 : Graph about the number of processes per second

Figure 7 shows the processing speed of proposed method as down-sampling rate. Since the 16X, there are almost no changed because of we not changed the parameter about minimum face size of Viola-Jones detector. We defined the accuracy for find best value of down-sampling rate as

$$\frac{ROI\ accuracy}{processing\ time}$$

eq. 14

Equation 14 means higher accuracy is better, less processing time is better. So it can be the one of methods to find the best value about down-sampling rate. Belowing Figure 8 shows 8 is the best score among other numbers.

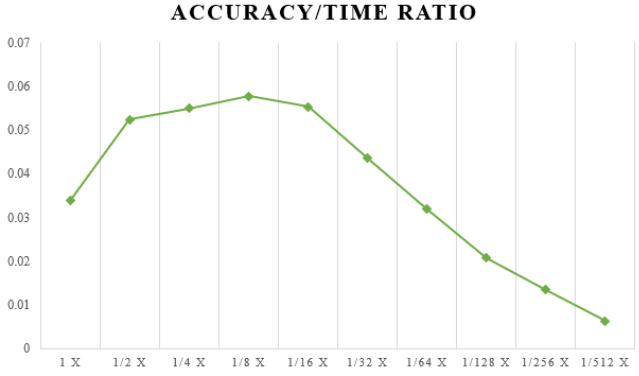

Figure 8: Graph about the score of result shown as accuracy/time ratio

C. *Detection time per a image*

We tested using a video has 20 more dancer per image set by 4k resolution and 60 frames per second spec. The number of frames for a step of testing is 100 and each frames are randomly selected among 11,600 frames.

Method	Hit Rate	False Positive Rate	Time
Viola-Jones	79.71%	14.83%	4399ms
Proposed method	77.24%	2.88%	1812ms

Figure 9 : Table for comparison with Viola-Jones detector.

Figure 9 shows the result about T-P/Ground-truth rate(Hit Rate), F-P rate and processing time. As a result, Vioal-Jones method show better HitRate, but Proposed method show less F/P rate and 2.43 times faster processing time per a image for best condition that down- sampling rate is 8.

AKNOWLEDGEMENT

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (NRF2018R1D1A1B07044286) and BK21 plus.

V. REFERENCES

[1] Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on* (Vol. 1, pp. I-I). IEEE.

[2] Zivkovic, Z. (2004, August). Improved adaptive Gaussian mixture model for background subtraction. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*(Vol. 2, pp. 28-31). IEEE.

[3] Achanta, R., Hemami, S., Estrada, F., & Susstrunk, S. (2009, June). Frequency-tuned salient region detection. In *Computer vision and pattern recognition, 2009. cvpr 2009. ieee conference on* (pp. 1597-1604). IEEE.

[4] Cheng, Xin, et al. "Efficient real-time face detection for high resolution surveillance applications." *Signal Processing and Communication Systems (ICSPCS), 2012 6th International Conference on*. IEEE, 2012.

[5] Yang, M. H., & Ahuja, N. (1998, December). Gaussian mixture model for human skin color and its applications in image and video databases. In *Storage and retrieval for image and video databases VII* (Vol. 3656, pp. 458-467). International Society for Optics and Photonics

[6] Danelljan, M., Shahbaz Khan, F., Felsberg, M., & Van de Weijer, J. (2014). Adaptive color attributes for real-time visual tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1090-1097).