Multi-modal Beauty Camera System

簡鈺晴 葛力維 李承宗
National Yang Ming Chiao Tung University

***Abstract* – Digital portrait enhancement involves a complex balance between noise reduction and feature preservation. This project presents a Multi-Modal Beauty Camera system constructed using fundamental signal processing techniques, while conducting a rigorous comparison against a modern Deep Learning framework (MediaPipe). The proposed mathematical pipeline analyzes imagery across three modalities: Illumination (YCbCr/Histogram Equalization), Structure (Sobel), and Texture (Gabor Filters). A specialized morphological algorithm is developed for precise blemish detection. To validate the efficacy of this pure mathematical approach, we contrast its performance with a MediaPipe-based pipeline, which utilizes 468 facial landmarks for masking. The final skin smoothing in both approaches is handled by a Guided Filter. The results demonstrate that while MediaPipe offers superior robustness in extreme angles, the proposed signal processing method achieves competitive natural aesthetics with greater interpretability and lower dependency on pre-trained models.**

Index Terms - Signal Processing, Multi-Modal Analysis

I. Introduction

This report is for Multi-model Image Processing class final project.

A. Motivation

1) *Acne and Blemish Removal:* From a signal processing perspective, acne shares similar high-frequency characteristics with essential facial features like eyelashes or pores. A naive low-pass filter (like Gaussian Blur) destroys both. This necessitates a selective approach: we must mathematically distinguish between "noise" (acne) and "detail" (features). This motivates the use of Mathematical Morphology (Top-Hat Transform) and Statistical Clustering to pinpoint imperfections for targeted healing (Inpainting) rather than global smoothing.
2) *Whitening and Tone Uniformity: Simply increasing global brightness often leads to a "washed-out" image where facial 3D structure is lost. The motivation here is to enhance the luminance (Y channel) and adjust the chrominance (Cb/Cr channels) to achieve a translucent, radiant aesthetic without compromising the natural shadows that define facial geometry. This requires precise color space manipulation rather than simple filter overlays.*
3) *Brightness and Illumination Correction:* Before any feature detection or smoothing can occur, the signal must be normalized. This motivates the implementation of Gamma Correction and Histogram Equalization as the first stage of our pipeline. By mathematically redistributing the pixel intensity probabilities.

II. Method

A. RGB to YUV

$$\begin{pmatrix} Y \\ C_b \\ C_r \end{pmatrix} = \begin{pmatrix} 0.299 & 0.587 & 0.114 \\ -0.169 & -0.334 & 0.500 \\ 0.500 & -0.419 & -0.081 \end{pmatrix} \begin{pmatrix} R \\ G \\ B \end{pmatrix} + \begin{pmatrix} 0 \\ 128 \\ 128 \end{pmatrix} \quad (1)$$

Based on (1), Separating image information into "Luminance" (Y) and "Chrominance" (Cb, Cr) allows for more efficient processing and transmission.

- Y (Luminance): Composed of a weighted sum of R, G, and B. The weights are 0.299, 0.587, and 0.114, respectively. This reflects the different luminance sensitivities of the human eye to different colors of light (especially green), representing the total energy of the visible light spectrum.
- Cb/Cr (Chrominance):
  1) Cb: Blue-difference component, reflecting the difference between the blue component and luminance.
  2) Cr: Red-difference component, reflecting the difference between the red component and luminance.

  Constants in the formula (e.g., +128) are used to shift the values to a standard digital signal range.

B. Gamma Correction

In digital image processing, the relationship between the actual light intensity recorded by a camera sensor and the perceived brightness by the human eye is not linear. Gamma correction bridges this gap by remapping pixel values to match human visual perception or to correct exposure issues.

$$V_{out} = V_{in}{}^{\gamma} \quad (2)$$

- V_{in}: The input pixel intensity (normalized to the range [0,1]).
- V_{out}: The output pixel intensity.
- A: A constant (usually 1 in simple correction).
- γ (Gamma): The exponent that defines the curvature of the mapping.

We empirically select a Gamma value of $\gamma < 1.0$ (specifically $\gamma = 0.9$), which expands the dynamic range of low-intensity pixels (shadows) while compressing high-intensity pixels.

C. Histogram Equalization

To further standardize the illumination and maximize the dynamic range of the input signal, we employ Global Histogram Equalization (HE). This process re-distributes the pixel intensities r of the input image to achieve a uniform distribution, relying on two fundamental statistical metrics: the Probability Density Function (PDF) and the Cumulative Distribution Function (CDF).

1) *Probability Density Function (PDF)*

The first step involves analyzing the frequency distribution of pixel intensities. Let the input image have L discrete gray levels (typically L=256) denoted by r_k , where k=0,1,…, L-1.It represents the normalized frequency at which each gray level appears in the image. It is mathematically defined as:

$$P(r_k) = \frac{n_k}{MN} \quad (3)$$

- n_k is the number of pixels with intensity r_k.
- MN is the total number of pixels in the image.

The PDF essentially quantifies the frequency of occurrence for specific pixel values, revealing whether the image is predominantly dark, bright, or low-contrast.

2) *Cumulative Distribution Function (CDF)*

While the PDF describes individual frequencies, the transformation function required to "flatten" the histogram is derived from the Cumulative. It is obtained by summing the PDF values:

$$s_k = T(r_k) = (L-1)\sum_{j=0}^{k} P_r(r_j) = \frac{L-1}{MN}\sum_{j=0}^{k} n_j \quad (4)$$

In the context of Histogram Equalization, the CDF serves as the transformation function (lookup table). By mapping the original intensities r_k to new intensities s_k using the CDF (scaled by the maximum intensity L−1), the system ensures that the output histogram approximates a uniform distribution, thereby maximizing global contrast and enhancing the visibility of facial structures.

D. Gaussian Blur

It is a low-pass filter to suppress noise while preserving the geometric structure of the face. While blurring an image before extracting edges may seem counter-intuitive, it is a fundamental prerequisite for robust feature extraction. The behavior of the Gaussian filter is governed by the standard deviation, σ .

$$h(x,y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \quad (5)$$

- $\sigma = 0$*: No filtering; noise remains.*
- $\sigma = 5.0$**:** Excessive blurring; structural edges are lost.
- $\sigma = 1.0$ **(Selected):** This value represents an optimal trade-off. It provides a "mild" blur sufficient to eliminate pixel-level noise and skin texture, yet it is narrow enough to maintain the sharpness of significant edges required for the protection mask.

E. Sobal

Image sharpening and edge detection are fundamentally based on differentiation. To identify the structural boundaries of the face (such as the jawline, nose bridge, and eyes), we need to detect regions where pixel intensity changes drastically. The Sobel Operator is a first-order derivative operator. Unlike second-order derivatives (e.g., Laplacian) which detect zero-crossings, the Sobel operator approximates the gradient of the image intensity function. This characteristic allows it to produce slightly thicker, more robust edge responses, which is advantageous for creating a "Protection Mask" that ensures complete coverage of facial contours.

1) Horizontal Gradient (G_x)**:** Convolved with kernel K_x, this detects changes in the horizontal direction, effectively highlighting vertical edges.

$$G_x = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}$$

2) Vertical Gradient **(G_y):** Convolved with kernel K_y, this detects changes in the vertical direction, effectively highlighting horizontal edges.

$$G_y = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix}$$

After computing the partial derivatives G_x and G_y for every pixel, combine them to determine the overall edge strength, which defined as the Gradient Magnitude (G):

$$M(x,y) = mag(\nabla f) = \sqrt{g_x^2 + g_y^2} \quad (6)$$

The magnitude value represents the rate of change at each pixel. A higher value indicates a steeper "slope" in intensity, corresponding to a sharper edge.

F. Gabor Filter

Unlike simple edge detectors, Gabor filters perform joint analysis in both the spatial and frequency domains, mimicking the receptive fields of simple cells in the mammalian primary visual cortex (V1).

Gabor kernel is defined as the product of a Sinusoidal Plane Wave and a Gaussian Kernel. x and y will change by angle to coordinate rotation.

$$x' = xcos(\theta) + ysin(\theta) \quad (7)$$

$$y' = -xsin(\theta) + ycos(\theta) \quad (8)$$

And combine (7) and (8) into gabor filter.

$$G(x,y) = \exp\left(\frac{-x'^2 + \gamma^2 y'^2}{2\sigma^2}\right) * cos(\frac{2\pi x'}{\lambda}) \quad (9)$$

- The Sinusoid: Responsible for detecting specific frequencies (repetitive patterns or "stripes").
- The Gaussian: Acts as a localization window, restricting the filter's view to a local neighborhood. Visually, the kernel resembles a "zebra crossing" pattern that fades out towards the edges. This structure allows the filter to act as a band-pass filter.

And Facial features possess specific geometric orientations. To capture all relevant features, we construct a "Filter Bank" comprising kernels at four distinct orientations (θ):

$$\theta \in \{0°, 45°, 90°, 135°\}$$

This multi-directional approach is critical for comprehensive facial analysis:

- 0° (Horizontal): Optimizes detection of horizontally aligned features such as eyebrows, eyes, and lips.
- 90° (Vertical): Optimizes detection of the nose bridge and facial outline.
- 45° / 135° (Diagonal): Captures angular features such as the jawline and nasolabial folds.

The result if is high-frequency signals are correspond to dense textures like hair, brows, lashes, otherwise, low-frequency signals are correspond to smooth, homogeneous regions which maybe is skin cheeks.

G. Guided Filter

The Guided Filter is derived from a local linear model. The fundamental hypothesis is that the filtering output q is a linear transformation of the guidance image I within a local window ω_k centered at pixel k. This assumption ensures that the output image inherits the structural edges of the guidance image while smoothing out details (noise/texture). And the relationship is expressed as:

$$q_i = a_k I_i + b_k \quad (10)$$

- q_i (Output): The pixel intensity of the final smoothed image.
- I_i (Guidance): The pixel intensity of the reference image (typically the original image), which contains the structural information we wish to preserve.
- a_k (Slope/Gradient): The linear coefficient that determines the fidelity of edge preservation.
- b_k (Intercept/Bias): The constant term that determines the local base luminance level.

To determine the linear coefficients (a_k, b_k), the algorithm minimizes the squared difference between the output q and the input image p. The analytic solution is derived using basic statistical moments:

- Covariance (Cov_{Ip}): This measures the correlation between the guidance image I and the input image p. It quantifies how similar the intensity trends are between the structure reference and the raw input.
- Variance (Var_I): This measures the variance of the guidance image I within the local window. It serves as a metric for the local "activity" or "roughness" of the image. The coefficient a_k is calculated as:

$$a_k = \frac{\text{Cov}_{Ip}}{\text{Var}_I + \epsilon} (11)$$

By utilizing this statistical dependency, the Guided Filter achieves what simple Gaussian blurring cannot: it selectively smooths low-variance skin textures while rigorously protecting high-variance facial features.

H. False Positive Reduction via Chrominance Statistics

A significant challenge in blemish detection is the semantic ambiguity of "redness." While acne is characterized by inflammation (high redness), desirable facial features such as lips (mucosal tissue) and ears (vascular tissue) also exhibit strong red characteristics. Relying solely on morphological shape (Top-Hat transform) may fail when the lips are textured or when the ear structure is complex, potentially leading to the algorithm identifying these features as clusters of acne and blurring them out.

To address this, we utilize the Cr (Chrominance-Red) component from the YCbCr color space previously computed in the pre-processing stage. Since lighting conditions and skin tones vary between images, a fixed pixel value threshold is unreliable. Instead, we implement an Adaptive Statistical Threshold based on the global distribution of the Cr channel.

We calculate the global Mean (μ_{Cr}) and Standard Deviation (σ_{Cr}) of the Cr channel pixels. We then identify the "High Redness" regions—which correspond to lips and ears—using the following condition:

$$M_{lips}(x,y) = \begin{cases} 1 & \text{if } Cr(x,y) > \mu_{Cr} + \lambda \cdot \sigma_{Cr} \\ 0 & \text{otherwise} \end{cases} \quad (12)$$

- μ_{Cr} represents the average redness of the face.
- σ_{Cr} represents the contrast or spread of the redness.
- λ is a sensitivity coefficient (typically set between 0.6 and 2.0).

By incorporating the high-chrominance mask defined in Equation (12) into the global protection mask, the system's robustness is significantly enhanced. This integration creates a comprehensive barrier that prevents the Guided Filter from erroneously smoothing naturally red features, thereby ensuring that the lips and ears retain their structural and color fidelity.

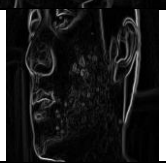
III. RESULT

A. *Acne and Blemish Removal:*

1) Face Mask

Table I presents the intermediate results obtained following the luminance extraction (as described in Section II.A) and the subsequent structural analysis. By applying Gaussian Smoothing followed by the Sobel Operator, the system successfully extracts the primary facial contours and high-gradient features.

TABLE I
USE GAUSSIAN AND SOBEL TO DETECT FACE

Original image	After Gaussian and Sobel
	
	

As observed in Table I, while this process effectively delineates the major edges, it also inadvertently highlights high-contrast skin blemishes. Furthermore, the Sobel operator alone is insufficient for capturing complex, repetitive patterns such as hair and eyebrows. To address this limitation and distinguish meaningful texture from random noise, we introduce Gabor Filter (as described in Section II.F) Banks for frequency-based analysis. The integration of these textural features with the structural edges is demonstrated in Table II. And we can get the whole facial features.

TABLE II
DETECT FACE MASK

Original image	After Process
	
	

To establish a benchmark for our proposed method, we incorporated the MediaPipe framework, a Python-based deep learning package, for facial feature detection. Instead of relying on standard OpenCV detection algorithms, we utilized MediaPipe's 468 facial landmarks to define specific Regions of Interest (ROIs). By calculating the convex hull of these landmarks, we generated binary

TABLE III
USE MEDIAPIPE TO DETECT FACE

Original image	After Process
	
	

polygonal masks to protect critical features such as the eyes and mouth.

A comparative analysis of the results in Table II and Table III reveals a distinct trade-off between the two methodologies. The proposed signal processing approach (Table II) demonstrates superior sensitivity to fine details, successfully capturing intricate textural features. However, this high sensitivity comes with a cost: the algorithm is more susceptible to retaining high-frequency noise within the protection mask.

In contrast, the MediaPipe-based approach (Table III) generates geometrically cleaner masks with fewer noise artifacts. However, it exhibits limitations in adaptability and precision regarding specific facial topographies. Notably, the MediaPipe model demonstrates segmentation errors, such as the incomplete coverage of the male subject's nasal bridge. Consequently, these unprotected areas are subjected to smoothing, resulting in a loss of structural definition that our mathematical approach successfully preserved.

2) *Ance* Mask*:* Following the establishment of the feature protection mask, the secondary objective is the precise identification and removal of acne.

I. *Traditional Method*

Acne detection is local erythema (redness), which manifests as localized high-intensity peaks within the Cr channel. To isolate these specific anomalies from the varying underlying skin tone, we applied a Morphological Top-Hat Transformation. This non-linear operation calculates the difference between the original Cr signal and its morphologically 'opened' version. This process effectively suppresses the low-frequency background skin color while extracting small, bright local features (acne) that are smaller than the defined structuring element.

TABLE IV
DETECT RED CHANNEL AND USE TOP-HAT TRANSFORMATION

Original image	After Process
	

As demonstrated in Table IV, while the algorithm successfully highlights acne, it suffers from significant false positives. Structural features with high contrast or pigmentation—specifically the lips and hair—are erroneously classified as blemishes. This necessitates a secondary filtration step.

To rectify the misclassification of lips, we implemented the Chrominance Statistical Exclusion method detailed in Section II.H. This method relies on the high intensity of the Cr (red chrominance) channel to distinguish mucosal tissue (lips) from skin inflammation.

TABLE V
GET ACHE MASK

Original image	After Process
	
	

For Table V, the method proved highly effective for female. Due to naturally higher lip pigmentation (or the presence of cosmetics), the lips exhibited a distinct Cr signature, allowing the algorithm to successfully detect and exclude them from the acne mask. But for male Subject, the detection was more challenging. The male subject's lips exhibited lower color saturation and contrast relative to the surrounding skin. Consequently, the statistical threshold from Section II.H struggled to distinctively separate the lips from the face, indicating a need for lower thresholds or morphological refinement in low-contrast scenarios.

To isolate facial features characterized by extreme contrast, such as hair and eyes, we employed a luminance-based segmentation approach. By detecting regions of low intensity (representing hair and pupils) and high intensity (representing the sclera), we effectively differentiated these anatomical structures from the surrounding mid-tone skin. The specific results of this segmentation are presented in Table V. Subsequently, these regions were integrated with the primary protection mask to ensure comprehensive feature preservation, and the final composite result is visualized in the figure below.

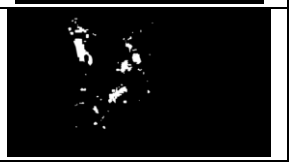
TABLE VI
THE RESULT MASK

Original image	After Process
	
	

II. *Mediapipe Method*

For the comparative analysis using the MediaPipe framework, the acne mask was generated by analyzing the Cr (red chrominance) channel within the segmented facial region. Specifically, we established a baseline skin tone and classified any pixels with Cr intensity significantly higher than this standard as blemishes.

TABLE VII
THE RESULT MASK

Original image	After Process
	
	

Ultimately, while the traditional computational approach lacks the semantic context awareness of MediaPipe, it exhibits a distinct advantage in edge fidelity. Our results indicate that although MediaPipe offers more stable general detection, our algorithm provides tighter control over high-frequency details, resulting in sharper transitions at specific facial boundaries.

B. Whitening and Tone Uniformity

In this section, we apply Gamma Correction (detailed in Section II.B) to enhance the luminance of underexposed facial regions. The results of this process are visualized in the figure below.

TABLE VIII
USE WHITENING METHOD

Original image	After Process
	

C. *Brightness and Illumination Correction*

To finalize the enhancement, we superimpose the protection mask onto the processed image. This step effectively reintegrates high-frequency details (such as facial features) into the smoothed background, thereby improving the global visual fidelity.

TABLE IX
USE BRIGHTNESS METHOD

Original image	After Process
	
	

IV. DIFFICULTIES WE MET

This project proved to be significantly more complex than anticipated. We initially underestimated the depth of the underlying signal processing principles, which required a substantial investment of time to fully grasp and implement from scratch. Technically, we encountered issues with over-smoothing, where essential facial feature signals were sometimes inadvertently suppressed or removed during the filtering process. Furthermore, due to time constraints, we were unable to deploy the system on large-scale public datasets or conduct extensive quantitative benchmarking against other state-of-the-art models.

ACKNOWLEDGMENT

Here have the whole work here: multi_modol_final_project

REFERENCES

[1] 許志仲, Lecture3-Image Filtering, e3@NYCU
[2] 許志仲, lecture4-Image resotration, e3@NYCU
[3] 許志仲, lecture6-Frequency analysis, e3@NYCU