13

Building a Visual Front-end for Audio-Visual Automatic Speech Recognition in Vehicle Environments

Robert Hursig and Jane Zhang
California Polytechnic State University, USA

1. Introduction

Automatic speech recognition (ASR) holds the promise of providing a natural, efficient, and safer means for communication between humans and computers and can profoundly change the way we live. Since its invention in the 1950s, ASR has witnessed considerable research activities and in recent years is finding its way into practical applications as evidenced by more and more consumer devices such as PDAs and mobile phones adding ASR features. While mainstream ASR has focused almost exclusively on the acoustic signal, the performance of these systems degrades considerably in the real-world in the presence of noise. One way to overcome this limitation is to supplement the acoustic speech with a visual signal that remains unaffected in an audibly noisy environment, yielding what is known as audio-visual automatic speech recognition (AVASR).

While previous research demonstrated that the visual modality is a viable tool for identifying speech [1-4], the visual information has yet to become utilized in mainstream AVASR. Despite years of research attention, there has been limited success in creating a system that can reliably detect lips in unconstrained imagery. Existing systems employ methods such as snake and active shape models [5,6], Markov Random Field (MRF) techniques [7], and multi-class, shape-guided fuzzy c-means (FCM) clustering algorithm [8], to detect and locate lips within an image. While the results are commendable, the extensive calculations demanded by these methods are significant. Moreover, a majority of existing lip localization techniques focused on lip parameter extraction within controlled environments with ample image resolution. Within the unconstrained visual environment, AVASR systems must compete with constantly changing lighting conditions and background clutter as well as subject movement in three dimensions. The difficulty of accurately and reliably detecting and tracking lips in unconstrained imagery is a major obstacle in the development of a practical AVASR system in the real world.

In this work we directly address the unconstrained imagery in the development of the visual front end of a practical AVASR system. Generally, the in-car audio-visual environment can be considered as a worst-case scenario for AVASR. Background noise and mechanical vibrations from traveling vehicles severely decreases operational signal-to-noise ratios for audio processing. Several products such as Ford Motor Company's Sync® and BMW's high-end Voice Command System use strictly audio information to recognize user

requests. However these systems notably suffer from user voice dependence and background noise such as open windows or ambient noise from highway speeds. Likewise, the visual environment inside a car is also challenging, imposing rapidly changing lighting conditions, moving faces within the vehicle, and constantly changing background clutter. In this work, algorithms were developed based on training and test datasets drawn from the AVICAR database [9] that was collected in such an environment. This database contains audio-visual recordings of 50 male and 50 female participants with varying ethnicities, constantly changing lighting conditions and cluttered background within a moving automobile. Video and image resolution for this database is 240-by-360 pixels, height-by-width.

The goal of this work is to develop a robust image lip localization algorithm designed as a visual front end of an AVASR system in vehicle environments. First, we address one essential first step – accurately and reliably locate the face in a moving car. In this work, both color and spatial information are exploited to find a face in a given image. A novel Bhattacharyya-based face detection algorithm is used to compare candidate regions of interest with a unique illumination-dependent face model probability distribution function approximation. In the subsequent step, a lip-specific Gabor filter based feature space is then utilized to extract facial features and locate lips within the frame. In both modules, extensive training and test sets from the AVICAR database will be used to justify design decisions and performance.

2. Face detection

Accurate face detection plays a critical role in successful lip localization and subsequent interpretation of the spoken words through extractred lip parameters. The relatively small size and constantly changing shape of lips does not realistically allow for feasible direct lip detection. Coupled with the difficulties introduced by an unconstrained operational environment, a robust, computationally efficient face detection algorithm is desirable to precede lip localization itself. Many facial recognition methods exist, such as the popular face detector proposed by Viola and Jones in 2001 [10]. However this and many other detectors requires only the intensity component of an image without taking full advantage of the inherent color information which is readily available in most images. In addition, they tend to break down in imagery with complex background such as the database in this study. We believe color could be used as a far more efficient criteria that could drastically reduce the search area and simplify the face detection process. The following sections offer a fast and noise-resistant face detection algorithm by which skin is first classified in an appropriate color space and then subsequently classified as a face or non-face.

2.1 Skin classification via sHSV color space

To determine the optimal color space for efficient skin and face detection, various color spaces have been examined, such as RGB, nrgb, YcbCr, YIQ, and HSV in [11]. Manaully drawn lip masks were constructed from a database of over 400 images that were subsequently used to develop statistical models of Lip, Non-lip, and Skin classes. Histograms were generated for each class and color space and, when applicable, the Gaussian approximations are calculated. Fig.1 shows the approximated Gaussian distributions for each of the three components in five color spaces. Each color spaces'

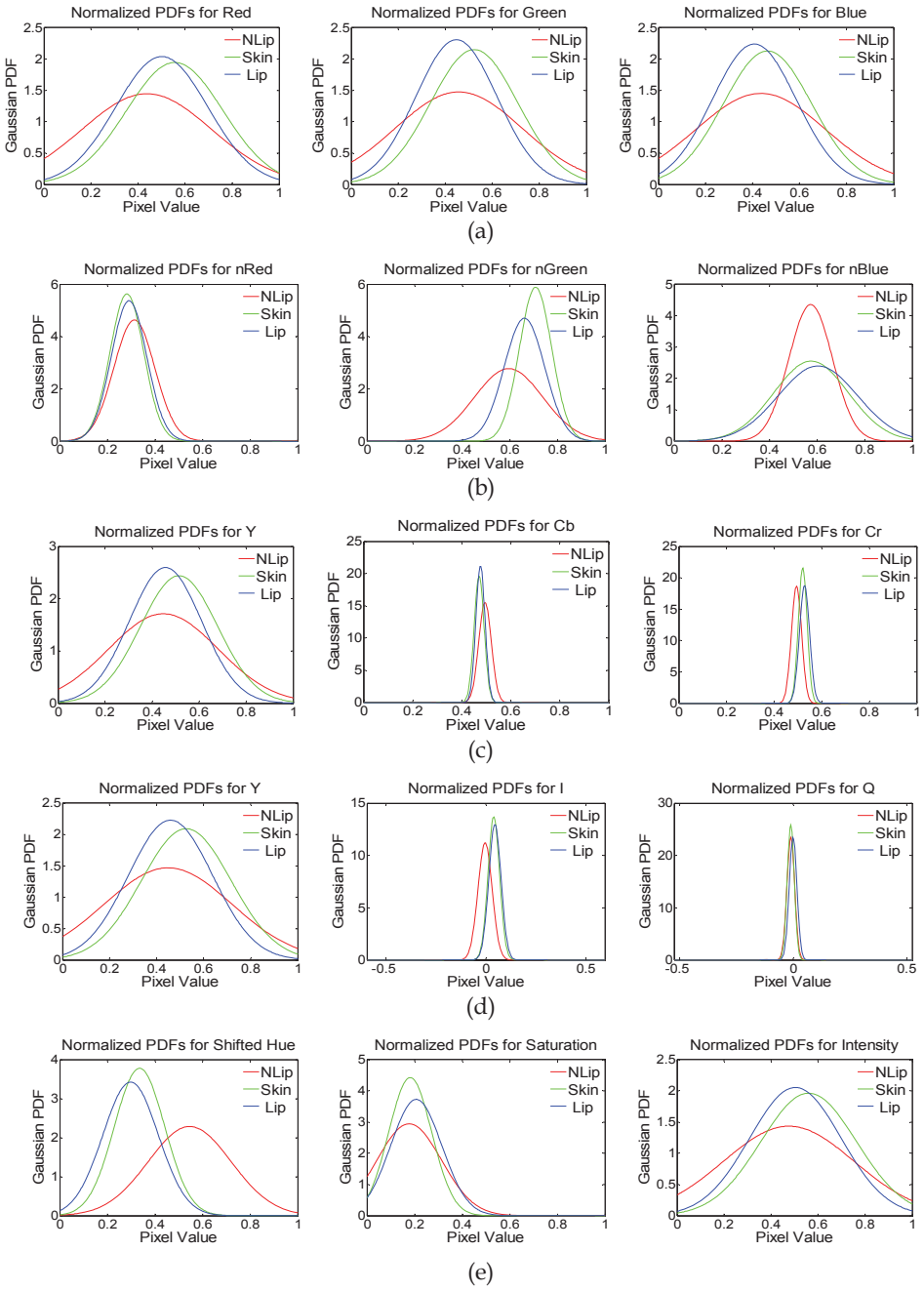


Fig. 1. Gaussian distributions for (a) RGB, (b) NRGB, (c) YCbCr, (d) YIQ, and (e) sHSI color spaces.

components are then compared among the three regions to determine the minimum correlation between non-lip, skin and lip regions. Referencing Fig. 1, while the low variances of the (r,g,b), (Cb, Cr), and (I,Q) components provide a relatively uniform representation for the given region, the non-lip, skin, and lip regions are highly correlated (demonstrated by the overlap seen in the Guassian distributions). Therefore, these components are poor classifiers for discerning lips and skin from that of the background. Additionally, (R,G,B), (Y_{cb}) and (Y_{IQ}) show high correlation between the regions, resulting in a similarly poor classifier. The hue component, on the other hand, provides the maximum separation between skin and non-skin regions and, therefore, is the strongest classifier. Since face images in the database were taken under varying illumination conditions and for different skin tones, hue also provides an illumination-independent and race-independent component, making it ideal for simple, uniform-color thresholding for skin classification. Because of the hue's color wheel effect, to simplify thresholding operations, the standard hue is shifted to the right by a value of 0.2 (72°), resulting in a shifted HSV color space, or sHSV, where region of interest (skin color) incurs no discontinuity. By deploying Bayesian classifier, optimal decision boundaries for the classification can be determined [12]. Fig. 2 illustrates the un-normalized posterior hue distributions, where shifted hue for skin class is approximated by N (0.34, 0.11^2) and the non-skin class by N (0.55, 0.17^2). Here the green lines represent the zero-dimensional decision boundaries that separate the skin and non-skin regions. Between these boundaries, from a shifted hue value of 0.052 to 0.325, the skin posterior distribution surpasses that of non-skin and will classify as a skin pixel.

Building upon the theoretical Bayes classifier, the final skin classification system adds robustness and decrease computational requirements for subsequent face detection. To promote skin region continuity, a hysteresis threshold that uses both spatial and hue information was then employed. Additionally, to increase the skin detection robustness in low-light conditions, a minimum value component of 0.2 is set for all skin pixels, due to the study showing that more than 90% of skin exists above luminosities value of 0.15 when approximated by a Gaussian distribution [12]. To decrease computational demands, the original input image is downsampled to reduce computational complexity when these operations are performed.

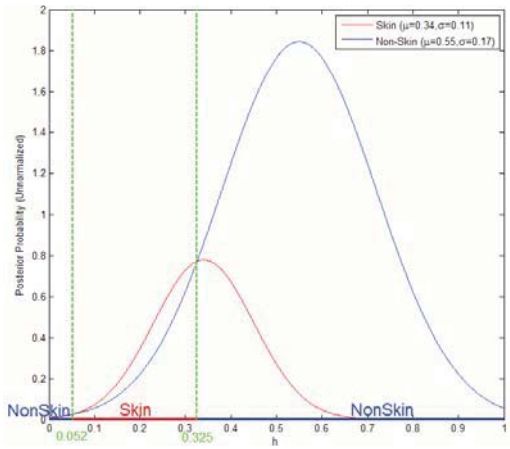


Fig. 2. Un-Normalized Posterior Distributions for Skin and Non-Skin Classes

2.2 Filtering and binary clustering

The unprocessed skin-classified binary images suffer from two main undesirable effects. One type of the error includes single-element impulse noise existing throughout the binary image as false positives within background regions as well as false negatives within skin regions, shown in green boxes in Fig. 3(b). Since false positives were deemed more detrimental to locating the dominant facial skin region, a 33rd percentile order-statistic filter of size 3x3 was selected as a more appropriate filter than the 50th percentile standard median filter. An extra benefit of this filter is that it better separates facial skin regions with skin colored car backgrounds. The red bounding box in Fig. 3(b) illustrates such a boundary, which is preserved via the 33rd percentile filter from (b) to (c). Had a median filter been applied to this image, the segregation would have disappeared and complicated face candidate localization and subsequent face detection. This is an important performance increase as the cluttered and similarly colored car backgrounds often result in false skin detection.

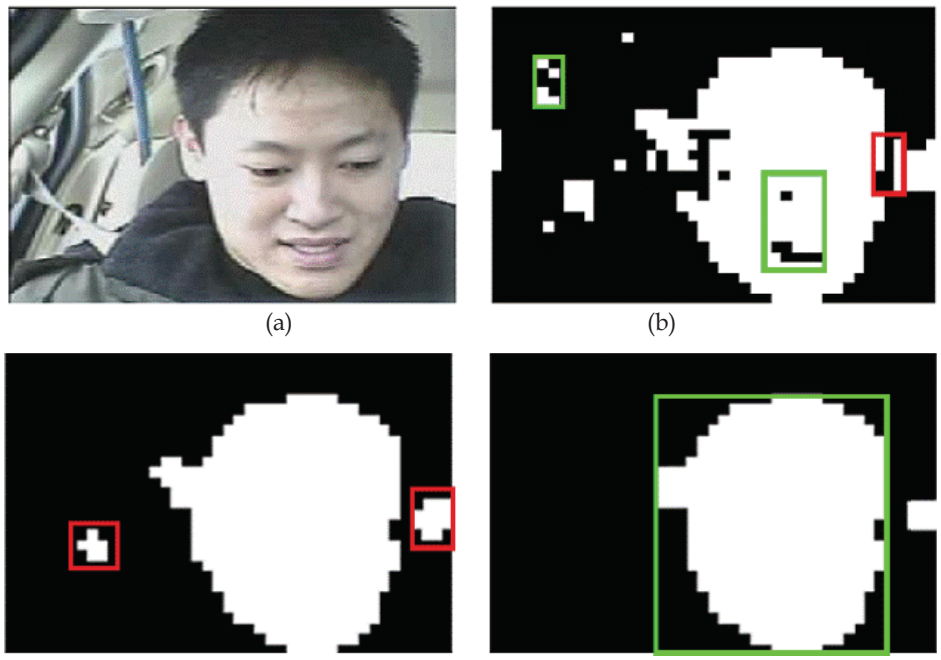


Fig. 3. Sample Post-Processing Imagery by Step (a) Original Image (b) Skin Classified Binary Image (c) 33rd Percentile Filtered (d) Application of Opening Operation.

The second type of error includes larger, false-positive regions that tend to dominate background (non-skin) regions. The binary morphological operation opening is utilized to minimize this effect. Notice the elimination of the leftmost background cluster in (c) and the reduction in size of the rightmost cluster. Since one face is assumed in each image, the largest skin cluster is selected as the region of interest, shown as the green bounding box in

(d), via the connected component labeling. This cluster will now be the input to the subsequent face localization algorithm.

2.3 Face candidate localization algorithm

Despite the filtering and classification methods employed, large regions of falsely classified background pixels still comprise part of the largest cluster returned by the pre-processing algorithm outlined in Section 2.2. Resulting from the unconstrained environment, these problem regions include skin-colored car interior regions, such as a car's roof, and window areas. Fig. 4 illustrates one such distinct, false positive protrusion resulting from a skin colored brick wall behind the car's back windshield. The goal of the face candidate localization algorithm is to simply determine such falsely classified regions attached to the largest cluster and remove them from the region of interest's (ROI's) bounding box. Fig. 5 provides an face candidate localization algorithm flow diagram to be developed within this section.

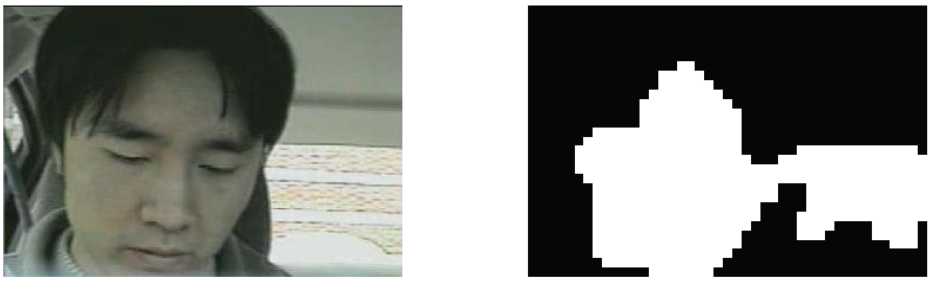


Fig. 4. Example Face Candidate Protrusion

Per Fig.5, the $M_c x N_c$ binary image face candidate, BW_c, is first input to the algorithm. To more effectively separate face candidates without significant background inclusions, an initial candidate screening takes place at the beginning of the algorithm. Sources cite that the average height-to-width ratio of the human face is approximated by the well-known golden ratio of 1.618:1. Accounting for facial tilt and out-of-frame rotation, typical face candidate ROI height-width ratio were found to exist between values of 1.2:1 and 1.7:1 through database measurements over the test subset. Hence, all face candidate ROI's whose height-to-width ratio, M_c/N_c, does not fall within the range [1.2, 1.7] will be subject to the remainder of the ROI pruning process.

For images which fall outside of the acceptable height-width ratio, further filtering takes place. To eliminate clear protrusions which are comparable in size to the face region itself a two-pass spatial filtering technique was employed. This technique locates sudden deviations in cluster configuration between the top and bottom of the face candidate cluster. While other more accurate methods, such as flood-fill techniques, exist to segment binary clusters, these methods are more computationally intensive, requiring several iterations of initial condition- and parameter-dependent morphological operations. Hence, the following computationally inexpensive method was employed to roughly locate distinct binary cluster protrusions similar to that in Fig.4, while preserving the roughly vertically oriented elliptical face region.

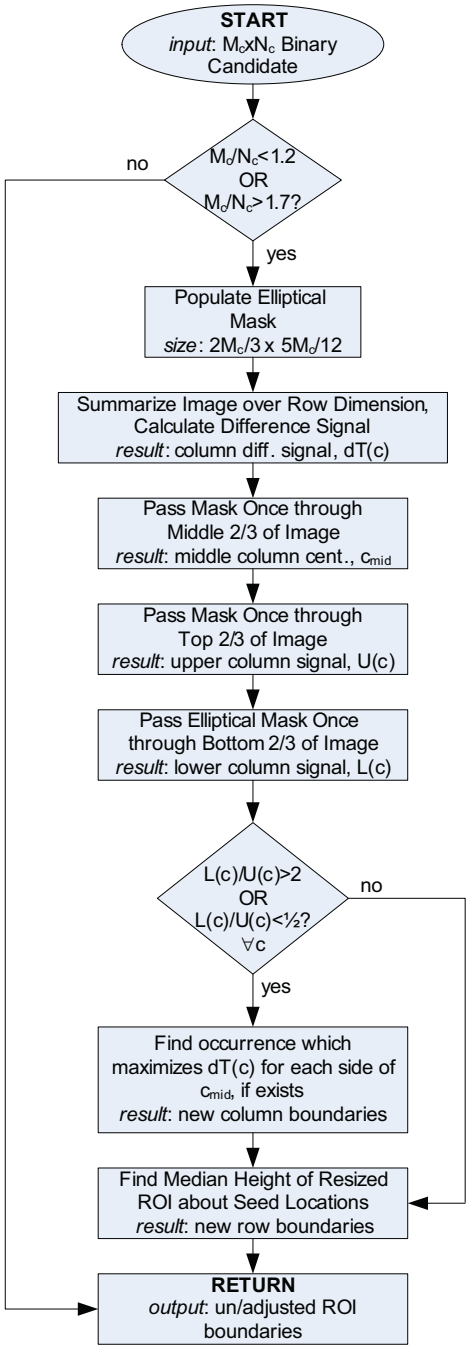


Fig. 5. Face Candidate Localization Algorithm Flow Diagram

The spatial filtering discussed is the result from passing an elliptical binary mask once through the top two-thirds and bottom two-thirds of the face candidate binary image, BW_c. The height of the elliptical binary mask, called H, was chosen to be two-thirds the input candidate ROI's height, M_c. The width of the ellipse was chosen to mimic the average dimensions of the human face, which is 1.6 times less than its height. Hence, the final size of the elliptical mask is M_h x N_h, where $M_h = floor(2M_c/3)$ and $N_h = floor(5M_c/12)$. The composition of the mask, H, is defined per the following equation

$$H(\mathbf{z}) = \begin{cases} 1 & \text{if} \quad \mathbf{z} \cdot \mathbf{z}^{\mathbf{T}} < 1 \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

where $\mathbf{z} = \begin{bmatrix} \dfrac{r - r_{H,cen}}{r_{H,cen}} & \dfrac{c - c_{H,cen}}{c_{H,cen}} \end{bmatrix}$, $c_{H,cen} = N_h / 2$, and $r_{H,cen} = M_h / 2$

where r and c are the row and column location of the elliptical mask. Thusly defined, the elliptical mask is not convolved with the face candidate binary image in the strictest sense. Rather, the elliptical mask, H, is passed once through the top two-thirds and once through the bottom two-thirds of the candidate ROI, centered about one-thirds and two-thirds of the candidate ROI's height, respectively. At each column location, the mask and image are multiplied and then summed by element, returning a value equivalent to the total number of skin-classified pixels enclosed within the mask H at that location. Let $U(c)$ and $L(c)$ be the column signals resulting from the upper and lower passes through the candidate ROI, BW_c, respectively. To preserve the accuracy of the spatial filtering, it should be noted that the input binary image, BW_c, was padded column-wise with $N_h/2$ zeros on each side of the largest cluster. Then the ratio of the upper signal to the lower signal is given by:

$$R(c) = \frac{U(c) + \varepsilon}{L(c) + \varepsilon} \qquad c = 1, 2, \ldots, N_c \tag{2}$$

where ε is a small positive integer introduced to safeguard against $L(c)=0$. This ratio signal effectively shows the relative distribution of the face candidate cluster with $R(c)>1$ indicating a greater concentration at the cluster's top and with $R(c)<1$ indicating a greater concentration at the cluster's bottom. Fig. 6 (a) contains an annotated example of the relative size and shape of the elliptical mask, the resulting upper and lower column signals, U(c) and L(c), as well as the ratio signal, R(c). Note that for clarity this example normalizes each column signal to the area of the elliptical mask.

After the ratio signal has been calculated over the width of the binary image, the binary image is summed across the row dimension yielding a total column vector, T(c). Equivalently, this total signal can be expressed as

$$T(c) = \sum_{r=1}^{M_c} BW_c(r, c) \qquad c = 1, 2, \ldots, N_c \tag{3}$$

Where r and c are the row and column indices, respectively, from the face candidate binary image. Next, an absolute difference signal, $dT(c)$, is derived from $T(c)$ per the following equation:

$$dT(c) = abs\big(T(c+1) - T(c)\big) \qquad c = 1,2,\ldots,N_c - 1 \tag{4}$$

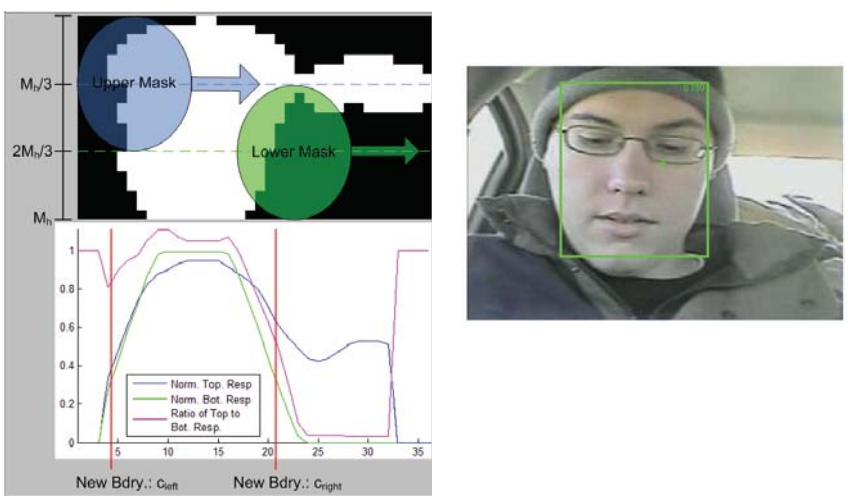
Next, a value of two is chosen to select the factor by which the upper and lower signals can deviate and still be considered part of the facial region. Then, letting $\boldsymbol{C}$ be the set of all column locations for which $R(c)>2$ or $R(c)<0.5$, the new horizontal boundaries, $c_{c,left}$ and $c_{c,right}$, of the candidate ROI is then selected by the following equation.

$$c_{c,left} = \begin{cases} \arg_c \max\{dT(c) \mid c \in \mathbf{C}\} & 1 \le c \le c_{mid} \quad \text{if} \quad \mathbf{C} < c_{mid} \ne \varnothing \\ 1 & \text{otherwise} \end{cases}$$

$$c_{c,right} = \begin{cases} \arg_c \max\{dT(c) \mid c \in \mathbf{C}\} & c_{mid} < c \le N_c \quad \text{if} \quad \mathbf{C} > c_{mid} \ne \varnothing \\ 1 & \text{otherwise} \end{cases} \tag{5}$$

where $$c_{mid} = \text{median}\big\{c \mid T(c) = \max\big(T(c)\big)\big\} \qquad c = 1,2,\ldots,N_c$$

where $c_{c,left}$ and $c_{c,right}$ are the new left and right ROI boundaries, respectively, and c_{mid} is the median value of c for which $T(c)$ is maximum over the candidate's entire width. In words, the new boundaries are selected by maximizing the difference signal for all locations where the upper and lower mask differ by a factor of two. This method effectively selects new boundaries located where an abrupt change in top-bottom concentration occurs.



(a) (b)

Fig. 6. Sample Face Candidate Localization Process (a) Original Face Candidate Cluster and Spatial Filter and Ratio Responses; (b) Successfully Modified Bounding Box

Lastly, new top and bottom boundaries, $r_{c,top}$ and $r_{c,bot}$, are created by median filtering the top and bottom cluster edges within $N_c^{'}/20$ pixels of the new ROI's horizontal center. Hence, the new face candidate ROI is now bounded horizontally over [c_{left}, c_{right}] and

vertically over [r_{top}, r_{bot}], noting that these ranges are referenced to the origin of the original candidate binary image, BW_c. Fig. 6 (b) illustrates a successfully modified ROI bounding box resultant from this algorithm. Note the correspondence between where the ratio signal drops below one-half and where the new boundaries are located. Also note that these new coordinates are referenced to the downsampled (M_dxN_d) image space and will require conversion back to the original resolution space. Now that the face candidate is localized by its four boundaries, it is subject to the next step, where the face detection algorithm determines whether it indeed is a face.

2.4 Face model joint histogram estimation

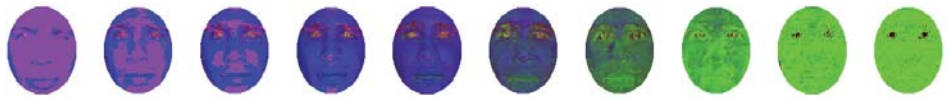
A critical component of face detection is modeling the variable human face such that a given algorithm provides accurate, repeatable, and reliable results. For this reason, selection of a proper feature set and development of an extensive, representative training set is critical for successful face detection algorithm. Building upon previous work [12], a joint shifted hue and saturation feature space was selected as the basis for face representation since it captures skin color information as well as the variation in saturation incurred around facial features such as eyes, nose, and mouth. Next, the joint probability density function was approximated as a histogram which quantizes the discussed two-dimensional feature space into a finite number of bins. To incorporate spatial information as well, the Epanechnikov kernel is employed in the histogram estimation. The Epanechnikov kernel weights a given ROI heavier towards the center and radially less towards the ROI's perimeter. Hence, it minimizes the effect of background pixels and skin edge pixels which are not always representative of the face itself. Crow utilized the Epanechnikov kernel noting similar advantages and associated performance increases [12]. Another benefit of the Epanechnikov kernel is that it is elliptically symmetric about the ROI's central coordinate, mirroring the natural shape of the human face within the ROI.

2.4.1 Forming the face model joint density estimators

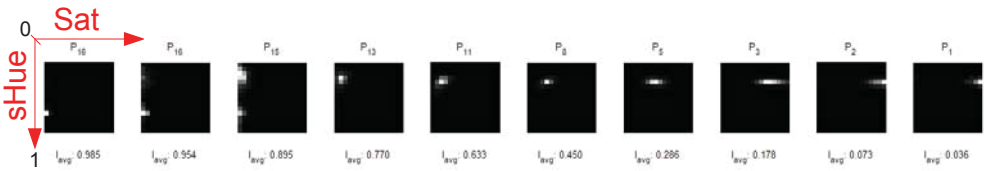
It is observed that while illumination content remains relatively constant within any given image, the average illumination within a given ROI directly impacts the distribution of the face within the joint shifted hue and saturation feature space. Hence, average intensity was chosen as an easily calculable metric which represents the face's ambient lighting conditions. For the sake of consistency, the illumination space was also quantized into a discrete number of bins and the Epanechnikov kernel will weight a pixel's contribution to the average illumination. Borrowing from previous work, the histogram bin count for each feature component, h and s, and the intensity information, I_{avg}, will be segmented into 16 discrete bins uniformly spread about the respective spaces. This value minimizes storage requirements while mitigating the risk of overfitting the actual distribution.

To construct the face model joint density estimators, training set containing 150 images from five individuals of varying skin tone taken under a range of ambient lighting conditions were collected. Care was taken to ensure that across each subject average illumination levels remained within 1/30 of each of the 30 values uniformly spread over the range [0,1]. For each image within the training set, the kernel-weighted intensity and the joint PDF histogram were calculated for each image after conversion to the sHSV color space. Selected results obtained by three of the five subjects are detailed in Fig. 7-9 representing light-, medium-, and dark-skinned individuals, respectively. It can be seen that changes in average

(a) Original Images from Subject 1 Training Set



(b) sHSV Converted Image Displayed as RGB

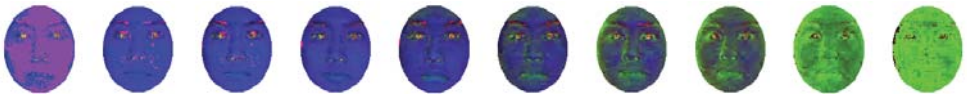


(c) Corresponding Joint sHue and Saturation Histogram*

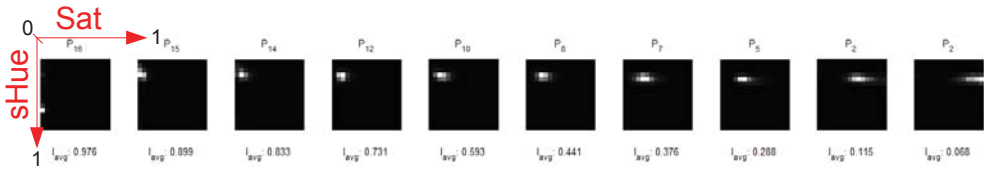
**for clarity each histogram has been normalized to its own maximum value*

Fig. 7. Face Model Illumination Dependence Training Set, Subject 1

(a) Original Images from Subject 2 Test Set



(b) sHSV Converted Image Displayed as RGB (sH→R,S→B,V→B)

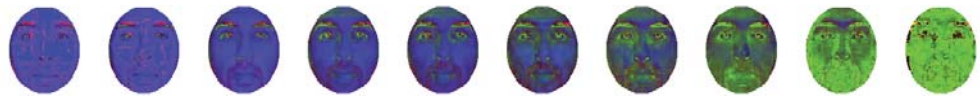


(c) Corresponding Joint Shifted Hue and Saturation Histogram*

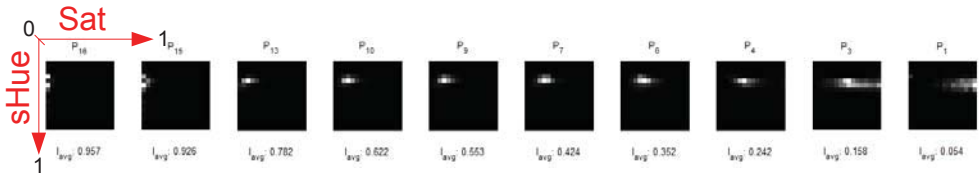
**for clarity each histogram has been normalized to its own maximum value*

Fig. 8. Face Model Illumination Dependence Training Set, Subject 2

(a) Original Images from Subject 3 Test Set



(b) sHSV Converted Image Displayed as RGB (sH→R,S→B,V→B)



(c) Corresponding Joint Shifted Hue and Saturation Histogram*

**for clarity each histogram has been normalized to its own maximum value*

Fig. 9. Face Model Illumination Dependence Training Set, Subject 3

illumination directly impact the distribution of the largely unimodal (singly peaked) shifted hue and saturation joint PDF. Furthermore, it can be seen across all PDF histograms that a majority of the hue content is contained within three or four histogram bins across all illumination values. However, saturation content varies from more tightly concentrated at low values under high illumination to roughly three times more spread about the saturation axis under low illumination. Differences in the PDF histograms between light and dark skin tones were slight, involving a positive one-bin shift of the general unimodal distribution along the hue axis. Moreover, at high illumination levels spreading about the hue axis occurred largely due to overexposure at the imaging device itself. Hence, the decision was made to replicate this dependence in the final face model.

The entire 150-image training database was utilized to construct a joint shifted hue and saturation histogram-estimated PDF for each discrete ROI average illumination bin. In words, the face model hisgram set is derived by summing each histogram over the training set whose parent image has the average illumination level and then normalizing each illumination level's PDF histogram independently to unity. The resulting face model PDF histogram approximation across each illumination level is displayed in Fig. 10. Here the value of I_{bin} refers to the average illumination component value which corresponds to the center (midpoint) of the discrete illumination bin, i. This face model histogram set will be stored in memory to be accessed by the face detection algorithm discussed in Section 2.5 to follow.

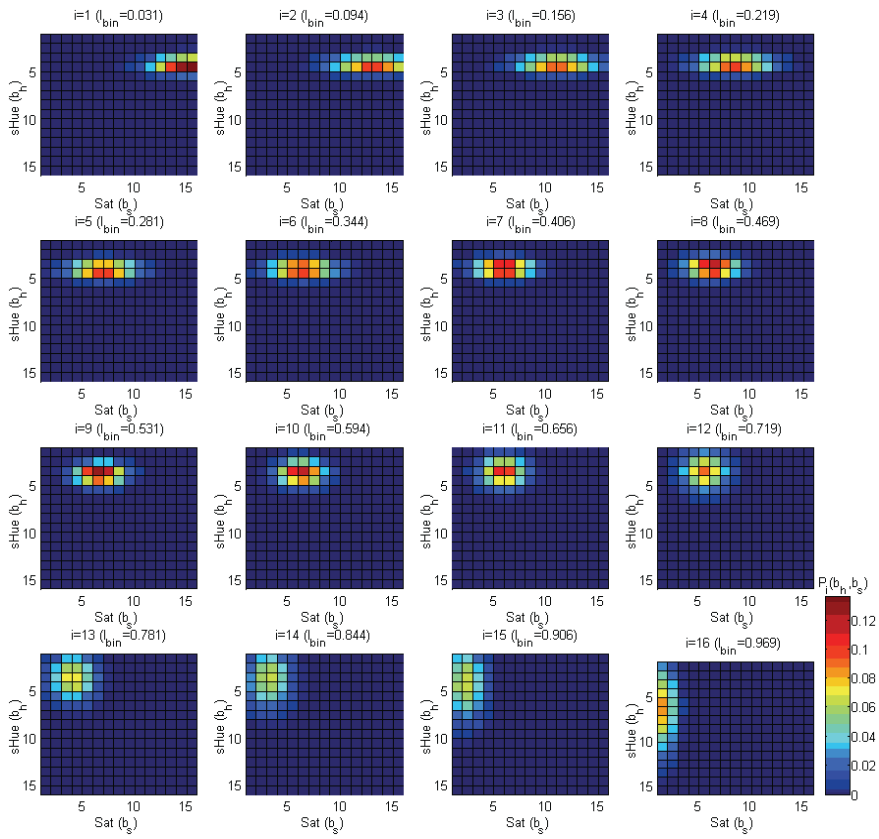


Fig. 10. Joint sHue and Saturation Histogram-Estimated PDF's over Average Illumination Bin Number

2.4.2 Forming the face candidate joint density estimators

With the face model density estimate in place, the face candidate density joint PDF must be constructed so that it can be compared with the model distribution. Derivation of the candidate's histogram approximated joint PDF is straightforward as it only entails the histogram associated with one ROI and its corresponding average illumination value. To complete this task, the face candidate which results from the face candidate localization algorithm (see Section 2.3) is converted to the original coordinate and resolution space. Next, the converted sHSV ROI will be kernel weighted and the histogram estimation process will take place. This face candidate joint density estimate, $\mathbf{P}_i$, will be compared with the face model histogram of the same illumination level, $\mathbf{Q}_i$, via the face detection algorithm outlined in the next section.

2.5 Face detection and test results

With a face model and candidate distributions in hand, candidate ROI's output from the skin detection and filtering algorithm can now be processed for the presence of a face. The

face detection algorithm implemented in this work utilizes the Bhattacharyya coefficient as a means by which the similarity between the generated face model joint histogram and that of a candidate ROI is measured.

The major advantage of the Bhattacharyya coefficient is that, unlike the Mahalanobis distance, it requires no statistical measure from each distribution, drastically reducing computational time and complexity. Remapping the definition of the Bhattacharyya to two dimensions, the Bhattacharyya coefficient can be defined as

$$\rho(\mathbf{P},\mathbf{Q}) = \sum_{h=1}^{m}\sum_{s=1}^{n}\sqrt{P(h,s)\cdot Q(h,s)} \tag{6}$$

where $\rho(\mathbf{P}, \mathbf{Q})$ is the Bhattacharyya coefficient between the m-by-n bin candidate histogram $\mathbf{P}$ and m-by-n bin model histogram $\mathbf{Q}$, and $P(h,s)$ and $Q(h,s)$ are the density of the candidate and model histograms, respectively, at bin location $[h, s]$. After the Bhattacharyya coefficient for a given set of candidate and model histograms has been calculated, a simple threshold is applied in order to claffiy the candidate ROI as either a *Face* or a *NonFace*. As expected, false positive error rates decreases as the threhold was increased as higher threholds effectively increased he similarity measure relative to the face model required for face detection. Conversely, false negative failure rates increased as the threshold was increased as an increased number of candidates failed to adequately compare in similarity to the model distribution. Via iterative analysis over the training set composed of 160 images, the Bhattacharyya coefficient threshold of 0.5 was then selected to minimize false negative and false positive error rates.

Face Detection Result	**Successful Localization Set***		**Complete Test Set**	
	Instances	**Percentage**	**Instances**	**Percentage**
Positive Face Detection ($\rho \geq 0.5$)	139	94.6%	144	90.0%
Negative Face Detection ($\rho < 0.5$)	8	5.45%	16	10.0%
Total Images	*147*		*160*	

**successful localization is defined as ROI contains 75% to 125% of the visible face.*

Table 1. Face Detection Algorithm Results

To test the performance of the face detection algorithm, another 160-image test set was created from the AVICAR database, not containing any images found in the face model or skin classification training sets. The test set was composed of 40 subjects at four different time instances throughout the video data. The performance of the face detector using this test set illustrates the success of the algorithm in response to variation in the subject's skin tone as well as any lighting or background changes over time. Recall that this test set also generated the face localization results from Section 2.3, where 147 of the 160 images incurred successful face localization. Table 1 details the true positive and false negative detection rates for both the complete test set and the subset for which the face candidate was successfully localized. As seen, the face detection algorithm achieved an overall accuracy of 90% across all test set images. The accuracy of the algorithm improves by 5% when the face itself is successfully bounded as a result of the face localization algorithm. Sample positive (*Face*) and negative (*NonFace*) classifications are contained within Fig.11 (a) and (b), respectively.

(a) (b)

Fig. 11. Sample (a) Positive Face and (b) Negative Face Detections

3. Lip feature extraction

The Gabor filter is a linear filter whose impulse response is defined as a sinusoidal function multiplied by a Gaussian function in the following form

$$G(x,y\,|\,\theta,F_o,N_x,N_y,\gamma,\eta,\phi)=\frac{\gamma\cdot\eta}{\pi}e^{-\left((\alpha x_r)^2+(\beta y_r)^2\right)}e^{j2\pi F_o\left(x_c\cos\theta+y_c\sin\theta+\phi\right)}$$

$$\forall x\in[1,N_x],\ y\in[1,N_y]$$

with
$$\alpha={F_o}/{\gamma},\ \beta={F_o}/{\eta},\ x_o={N_x}/{2},\ y_o={N_y}/{2} \tag{7}$$

where N_x and N_y are the width and height of the Gabor filter mask, respectively, ϕ is the phase of the sinusoid carrier, F_o is the digital frequency of the sinusoid, θ is the sinusoid rotation angle, γ is the Along-Wave Gaussian envelope normalized scale factor, and η is the Wave-Orthogonal Gaussian envelope normalized scale factor. These parameters define the size, shape, frequency, and orientation of the filter among other characteristics. G is the N_y-by-N_x Gabor filter and $[y,x]$ is the spatial location within the filter. The Gabor filter's invariance to illumination, rotation, scale, and translation, and its effective representation of

natural images, make the filter an ideal candidate for detecting the facial features in less than desirable circumstances [13].

Utilizing the 160-image training set from the AVICAR database, measurements of upper and lower lip thicknesses and orientations were recorded. It was found the upper lip thickness ratio h_{hi}/M_c and lower lip thickness ratio h_{low}/M_c, yield an average value of 0.136 and 0.065, respectively, where M_c measures height of the candidate's facial bounding box. Lip orientation, $\Delta\theta_{lip}$, was recorded as the absolute rotation of the mouth opening axis from horizontal and has an average measurement of 11.25^0. With this data, the Gabor filter set can now be created to more accurately represent the lip region. The final 12-component Gabor filter set, **G**, is thus defined as,

$$\mathbf{G} = \left\{ G_{n,t,f} = G(x, y \mid \theta = \theta_t, F_o = F_f, N_x = N_n, N_y = N_n, \gamma, \eta, \phi) \right\}$$

$$\begin{aligned} N_n &\in \left\{ \text{floor}\left(\frac{M_c}{8}\right), \text{floor}\left(\frac{M_c}{4}\right) \right\} \quad n = 1,2 \\ \theta_t &\in \left\{ \frac{3\pi}{8}, \frac{\pi}{2}, \frac{5\pi}{8} \right\} \quad t = 1,2,3 \\ F_f &\in \left\{ \frac{4}{N_n}, \frac{8}{N_n} \right\} \quad f = 1,2 \end{aligned} \tag{8}$$

with $\gamma = \eta = 1$ and $\phi = 0$

where G is defined in Eq. (8) and n, t, and f are the set indices of the (square) Gabor filter size, sinusoid angle, and digital frequency sets, respectively. In words, the Gabor filter set, **G**, is the set of Gabor filters for every combination of n, t, and f. The orientation values, $\theta_{t\in 1,2,3}$, were chosen such that the sinusoid orientation was vertically oriented (θ=90°) and $\pm 2\Delta\theta_{lip}$ away from vertical, where the factor of two was experimentally determined. In addition, the Gabor filter's size, N_n-by-$N_n|_{n\in 1,2}$, was selected such that over 80% of the total energy contained in the unbounded Gabor filter is contained within the N_n-by-N_n mask for any value of F_f (which depends upon N_n) and θ_t. The relative size and frequency of the Gabor filter to the candidate's height allows for a more scale-invariant design.

3.1 Gabor filtering algorithm

With the establishment of the lip-specific Gabor filter set, processing of the face-classified region of interest can proceed. Here, the sHSV triplet's value (illumination) component is selected as the feature space of choice for Gabor filtering since it best separates lip and surrounding face.

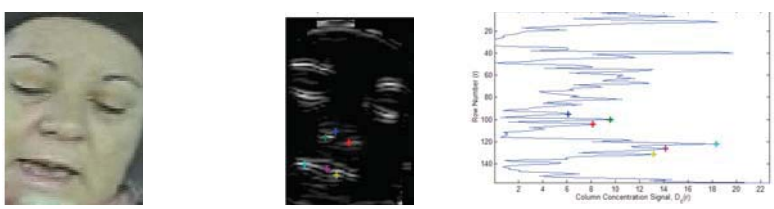
First, 12 Gabor filter responses are generated by performing two-dimensional convolution of the face-classified image's value component, V, independently with each Gabor filter configuration. Next, all 12 Gabor responses are summarized element by element. Due to the positive- and negative-valued modes of the Gabor filters, the total Gabor response is then normalized to the range [0,1] and further remapped to stress the maximal and minimal Gabor jet values. The final, normalized, and remapped Gabor filter response is denoted as

G_f, and has size M_c-by-N_c where M_c and N_c are the row and column sizes of the face candidate, respectively.

Subsequently, the Gabor filter response, G_f, undergoes mean-removal where all response pixel values are set to zero if they are less than the total response's sample mean and are left unchanged if the values are above the mean. Furthermore, to remove false positives within the background surrounding the face, the skin-classified binary mask is applied over the mean removed response. Fig. 12(b) shows a mean-removed and masked Gabor response, G_{mr}, of the original image in (a). As can be seen, smooth skin surfaces, such as the cheeks, provide minimal response while the mouth opening, lips, nostrils, eyes, and eyebrows provide much elevated responses. In addition, the cross section of the lip from chin to the region above the lip involves many oscillatory changes in intensity value. Mean removal effectively eliminates the contribution of background pixels to subsequent processing. The skin-classification masking also noticeably reduces the effect of several high-intensity non-face background regions.

3.2 Lip center coordinate estimation

Given the mean-removed and masked Gabor response, G_{mr}, a number of possible lip locations, called seeds, will be generated. Here, a column concentration signal, D_c, is first calculated from the G_{mr}. Then, seed row coordinates, $r_{pk,i}$ are chosen as local maxima of D_c, see colored crosses in Fig. 12(c). Peaks above image mean row value which do not exceed signal's mean are discarded. Finally, seed column coordinates $c_{pk,i}$, are chosen as midpoint of longest nonzero response chain in row. Hence, the ith seed point now has the location $[r_{pk,i}, c_{pk,i}]$. Fig. 12(b) shows the Gabor response, G_{mr}, overlaid with the seed locations indicated by the colored crosses.



(a) (b) (c)

Fig. 12. Sample Lip Coordinate Estimation Process. (a) Original RGB Face Candidate (b) Seed Locations within Mean-Removed, Masked Gabor Response (c) Seed Row Locations Overlaid on Dc Plot

Following seed generation, key parameters which are indicative of the presence of lips will be calculated. Utilizing these parameters, the *figure of merit* (FOM) will then be calculated as

$$\begin{aligned} &\mathbf{FOM} = \{FOM_i\} = \left\{D_{loc,i} \cdot D_{pk,i} \cdot r_{pk,i}\right\} \\ &D_{loc,i} \geq 1, \;\; D_{pk,i} \geq 1, \;\; r_{pk,i} \in [\mu_r, M_c] \end{aligned} \tag{9}$$

where **FOM** is the set of all figure of merit values, FOM_i, at seed index i, D_{loc} is the local two-dimensional concentration of G_{mr} about the seed, D_{pk} is the sum of all column concentration

signal peaks about the seed, and r_{pk} is the seed row location. Conceptually, the figure of merit in Eq. (9) combines the most visually apparent features of the lips into a single function. It has been argued that the lip's central coordinates are the coordinates for which the established figure of merit is maximal.

The lip center coordinate estimator was applied to the test set used in the previous sections. It was found that the figure of merit and Gabor filter system utilized in the lip coordinate estimate yields comparable results to those of the face detector algorithm of Section 2. Of the 139 images for which the face candidate ROI was successfully localized and classified as a face, the algorithm placed the lip coordinates on the lips for 89.2% of the time. When applied to the test set in its entirety, the lip coordinate estimation algorithm placed the estimated coordinate on the lips 83.8% of the time.

3.3 Lip localization and test results

Vertical lip localization within an image is inherently more complex than horizontal localization due to the striation (layers) of the Gabor response in the lip axis direction. Due to this, horizontal lip localization will be performed first to increase accuracy of the vertical localization. Fig. 13 illustrates lip localization procedure. To locate the lips in the horizontal axis, the row concentration signal $D_r(c)$ is computed over the lip region, shown in (c). Then, the left and right boundaries are determined where $D_r(c)$ is at 10% of that signal's maximum value above the mean.

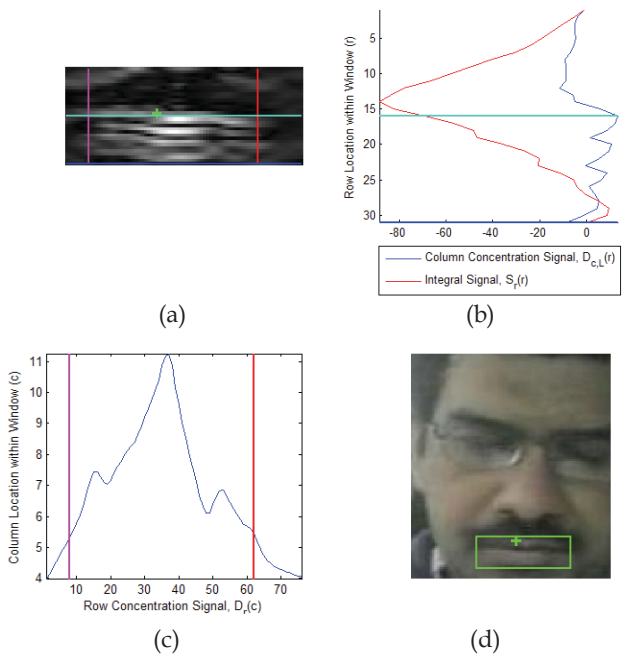


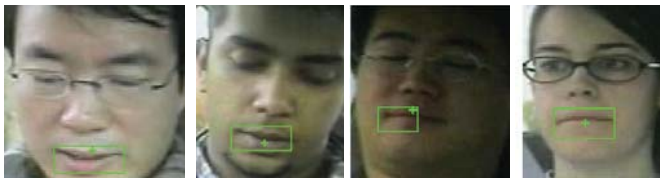
Fig. 13. Sample Horizontal and Vertical Lip Localization Procedure and Result. (a) Gabor Response within Lip Region (b) $D_{c,L}$ and S_r Signals over Lip Region Row (c) D_r Signal over Lip Region Column and (d) Lip Localization Result

After horizontal lip localization, vertical localization is undertaken, utilizing the returned left and right boundaries. To do so, the column concentration signal, $D_{c,L}(r)$, and column discrete integral signal, $S_r(r)$, are calculated between the left and right bounds only (see (b)). The integral signal is the summation of the mean-removed column concentration signal from the top of the lip region to row index r. Mean subtraction was performed on the column concentration signal such that lower intensity regions (rows) of pixels would count negatively toward the integral signal and higher intensity regions would positively count toward the signal. Finally, the lip localized upper and lower boundaries are found where the points are at 10% of S_{max} above the upper and lower minimum values, respectively. Sample lip localization success and failures are shown in Fig.14(a) and (b), respectively. When applied to the 160-image test set, factoring in face detection, the overall accuracy of 75.6%. Note that if the detected lip boundary is more than 5 pixels away from the lip corner or the closest lip point vertically, it is considered as a failure. The last image in Fig. 14 is considered a failure because the detected region contains more than 125% of the actual lips. While the overall accuracy is less than ideal, the challenges of the sub-optimal image quality and the unconstrained car environment make this a respectable value.

4. Conclusion and future work

Relative to previous work, positive face detection rates rose from 75% to 90% while effective lip localization rates rose from 65% to 75% when considering face detection as a front end to lip localization [12]. Among many techniques considered, the unique illumination-dependent face model and the adjusted skin classifier are considered successful and critical to the stated performance increase in face detection. The lip localization algorithm proposed a unique Gabor response feature space which relied upon a figure of merit rather than heuristic approximations, making it more versatile within the unconstrained environment.

Despite the stated performance increases, common sources of error include limited image resolution, skin-colored car environments, and overly bright and dark operating conditions without sufficient image dynamic range. The most notable improvement to the lip localization algorithm would be realized through the inclusion of time into the algorithm. Advanced difference imaging, the detection of movement between frames, would improve face localization and detection while reducing additional processing. Furthermore, face and lip spatial movement are generally orthogonal to each other, aiding the lip localization process even further.



(a) (b)

Fig. 14. Sample Lip Localization (a) Success and (b) Failures

5. References

[1] Stork, D.G. & Hennecke, M.E. (1996). *Speechreading by Humans and Machines*, in NATO ASI Series F, vol. 150, Sprinter Verlag.

[2] Zhang, X.; Broun, C.C.; Mersereau, R.M.; & Clements, M.A. (2002). Automatic Speechreading with applications to human-computer interfaces, *EURASIP Journal Applied Signal Processing, Special Issue on Audio-Visual Speech Processing*, vol. 1, pp.1228-1247.

[3] Potamianos, G. et al. (2004). Audio-Visual Automatic Speech Recognition: An Overview, in *Issues in Visual and Audio- Visual Speech Processing* by G. Bailly, E. Vatikiotis, and Perrier, Eds, MIT Press.

[4] Liew, A. & Wang, S.L. (2009) . *Visual Speech Recognition: Lip Segmentation and Mapping*, Medical Information Science Reference.

[5] Coulon, D.; Delmas, P.; Coulon, P.Y. & Fristot, V. (1999). Automatic Snakes for Robust Lip Boundaries Extraction, in *Proc. ICASSP*.

[6] Luettin, J.; Tracker, N.A. & Beet, S.W. (1995) . Active Shape Models for Visual Speech Feature Extraction, *Electronic System Group Report No95/44*, Univ. Of Sheffield, UK.

[7] Zhang, X. & Mersereau, R.M. (2000). Lip feature extraction towards an automatic speechreading system, *Proc. of IEEE ICIP*

[8] Wang, S.L.; Liew, A.; Lau W.H. & Leung, S.H. (2009). Lip Region Segmentation with Complex Background", in [4].

[9] Lee, B.; Hasegawa-Johnson, M.; Goudeseune, C; Kamdar, S.; Borys, S.; Liu, M.& Huang, T. (2004). AVICAR: Audio-Visual Speech Corpus in a Car Environment, in *INTERSPEECH2004-ICSLP*.

[10] Viola, P. & Jones, M. (2001). Robust Real-time Object Detection, in *Internationa Journal of Computer Vision.*

[11] Zhang, X.; Montoya, H.A. & Crow, B. (2007). Finding Lips in Unconstrained Imagery for Improved Automatic Speech Recognition, in *Proc. 9th International Conference on Visual Information Systems*, Shanghai, China.

[12] Crow, B. & Zhang, X. (2009). Face and Lip Tracking in Unconstrained Imagery for Improved Automatic Speech Recognition, in *Proc. 21th IS&T/SPIE Annual Symposium on Electronic Imaging*, San Jose, California.

[13] J. Kamarainen, V. Kyrki (2006). Invariance Peroperties of Gabor Filter-Based Features—Overview and Applications, in *IEEE Transactions on Image Processing*, vol. 15, no. 5, May 2006, pp. 1088-1099.