Face Localization Using Illumination-dependent Face Model for Visual Speech Recognition

Robert E. Hursig, Jane X. Zhang

Abstract—A robust still image face localization algorithm capable of operating in an unconstrained visual environment is proposed. First, construction of a robust skin classifier within a shifted HSV color space is described. Then various filtering operations are performed to better isolate face candidates and mitigate the effect of substantial non-skin regions. Finally, 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. Experimental results show a 90% face detection success rate despite the demands of the visually noisy environment.

Keywords—Audio-visual speech recognition, Bhattacharyya coefficient, face detection,

I. INTRODUCTION

AUTOMATIC speech recognition (ASR) is a well-researched field of study aimed at augmenting the man-machine-interface through interpretation of the spoken words. Examples of ASR include automated telephone directories, voice-activated cell phone commands, and speech-based in-car music and control systems. Traditional ASR system utilizes audio-only information and its performance degrades when employed in noisy environments such as moving vehicles. In fact, even in controlled environment, state-of-the-art ASR systems still underperform human's ability by over an order of magnitude [1]. It has been determined that human speech perception is bimodal in nature – that is, both audio and visual information are analyzed for speech perception. The latter information is more heavily valued by hearing impaired or normal humans when in noisy environments. Recognizing the benefit of visual speech, audio-visual automatic speech recognition (AVASR) system aims at improving the performance of speech recognition by combining both audio and visual speech data.

While previous research demonstrated that the visual modality is a viable tool for identifying speech [1,2], the visual information has yet to become utilized in mainstream ASR. Despite years of research attention, there has been limited success in creating a robust visual front end in an unconstrained imagery. This paper addresses one essential first step – accurately and reliably locate the face in a moving car. Accurate face localization plays a critical role in successful lip localization and subsequent interpretation of the spoken words through extracted 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.

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, training and test datasets are drawn from the AVICAR database [3]. 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. Datasets were created by extracting still images from the video files, which utilizes a wavelet-based, lossy audio-video interlaced (AVI) encoding scheme. Video and image resolution for this database is 240-by-360 pixels, height-by-width.

The human face is one of the most variable and common objects that humans interact with on a daily basis. Viola/Jones's face detector proposed in 2001 [4] is very popular. However, the detector runs directly on an entire image without taking advantage of the inherent color information of the face that could drastically reduce the search area. Many other facial recognition methods exist under controlled conditions including optimal lighting and camera angle, ample resolution and processing power [5-7]. While the results are commendable, the extensive calculations demanded by these methods are significant. Moreover, a majority of the existing techniques assume ample resolution and controlled lighting conditions, which is not feasible for a real world lip reading system. Within the visual unconstrained environment, AVASR systems must compete with constantly changing

R. Hursig was with California Polytechnic State University, San Luis Obispo. He is currently with Sandia National Laboratories, Albuquerque, NM 87123, USA (phone: 505-284-4890, e-mail: rehursig@sandia.gov).
J. Zhang is with California Polytechnic State University, San Luis Obispo, CA 93407, USA. (phone: 805-7567528, e-mail: jzhang@calpoly.edu).

lighting conditions and background clutter as well as subject movement in three dimensions.

Thus, the goal of this work is to develop a robust still image face localization algorithm within the unconstrained car environment that precedes lip localization. This algorithm is designed as a visual front end to the larger AVASR system as a whole. This paper is organized as follows: In Section 2, we first detail the construction of a robust skin classifier within a shifted HSV color space. Section 3 describes various filtering operations performed upon the classification to better isolate face candidates and mitigate the effect of substantial non-skin regions. Section 4 proposes a unique, illumination-dependent face model probability density function approximation derived through an extensive training set that will serve as the basis for face detection. Section 5 describes the Bhattacharyya-based face detection algorithm itself and results of the larger face detection algorithm as a whole as applied to our database. Finally, Section 6 offers conclusions and recommendations for future improvement.

II. Skin classification via sHSV color space

In order to efficiently detect skin and faces within an image, the respective classifier must be developed within an appropriate color space. Proper color space selection has the effect of simplifying the classification complexity and dimensionality while improving inter-class separation. Extensive research has attempted to determine the optimal color space for skin detection with mixed findings [8-12]. In such studies, Shin *et al.* determined that most color space conversions fail to deliver ample skin detection improvements [8], while Jones *et al.* determined NRGB was the optimal color space [9]. Ming-Hsuan *et al.* and Zhang *et al.* selected perceptual color spaces such as HSV [10,11], and Abdel-Mottaleb *et al.* selected TV color spaces such as YIQ [12]. Based on previous work done by Zhang *et al.* [11,13] and further experimentation, in this work the HSV color space was adopted as the optimal skin and face detection color space under the discussed operating conditions.

In [11], it was shown that the HSV color space provides an illumination-independent color component as well as a separate value component, making it ideal for skin classification via simple thresholding. Because of the hue's color wheel effect, the standard hue is shifted to the right by a value of 0.2 (72°) to simplify the thresholding operation, resulting in a shifted HSV, or sHSV, color space where the region of interest (skin color) incurs no discontinuity.

In the first stage of skin classification, each pixel in the image is examined and classified as one of the two classes: Skin or NonSkin. The optimal decision boundary can be determined by deploying Bayes' rule via the following equation:

$$P(h \mid Skin) \cdot P(Skin) \underset{\substack{< \\ h \in NonSkin}}{\overset{\substack{h \in Skin \\ >}}{}} P(h \mid NonSkin)P(NonSkin) \quad (1)$$

where h is the shifted hue component of the sHSV triplet for a given pixel. Here, the a priori probabilities for skin and non-skin classes within the AVICAR training set were

estimated (based on manual segmentation of training data) as P(Skin)=0.7853 and P(NonSkin) = 0.2147. The class conditional densities, $P(h|\text{class})$, were determined by approximating hue histograms of manually segmented face/non-face images from the same training set [11]. Hence, Eq. (1) reduces the implementation of skin classification to a simple thresholding operation.

Fig. 1 illustrates the un-normalized posterior distribution respective to the class in question, where shifted hue for skin class is approximated by $\mathcal{N}(0.34, 0.11^2)$ shown in red and the non-skin class by $\mathcal{N}(0.55, 0.17^2)$ shown in blue. $\mathcal{N}(\mu,\sigma^2)$ denotes the Gaussian distribution with mean μ and variance σ^2. The green lines represent the 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.

Letting t_{lo} and t_{hi} be the lower and upper boundaries of the classifier, respectively, the theoretical skin classifier is defined as

$$C_o(h) = \begin{cases} Skin & \text{if} \quad t_{lo} \le h \le t_{hi} \\ NonSkin & \text{otherwise} \end{cases} \quad (2)$$
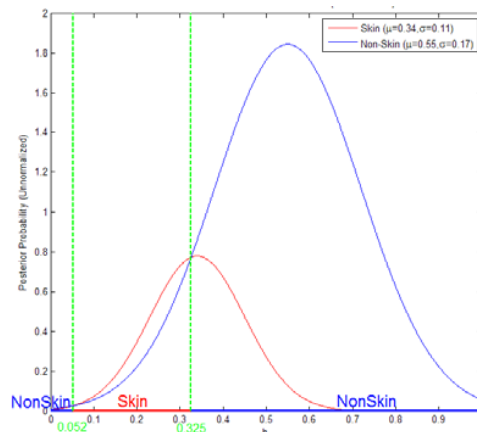
where $t_{lo} = 0.052$ and $t_{hi} = 0.325$

Fig. 1. Un-Normalized Posterior Distributions for Skin and Non-Skin Classes

When applied to the AVICAR database, the theoretical Bayes classifier yielded elevated partial facial skin detection. Incomplete facial skin detection is especially detrimental as the face detection methodology employed in this work benefits from a cohesive (continuous) skin classification mask that minimizes background pixel contamination. To promote skin region continuity, a hysteresis threshold that uses both spatial and hue information is then employed. Hysteresis thresholding results in skin classification if it satisfies the hard thresholding of Eq. (2) or if the soft thresholding is satisfied given at least one of the eight neighboring pixels satisfies Eq. (2). Additionally, to increase the skin detection robustness in low-light conditions, a minimum value component of 0.2 is set for all skin pixels. This is based on the study that more than 90% of skin pixels exist above illumination value of 0.15 [13].

The classification is applied to a manually classified subset of the AVICAR database. The set was constructed by selecting 20 male and 20 female candidates from the AVICAR database and extracting four separate images from the subject's video, comprising a 160-image training set in total. Subjects and images were selected such that the images provided a representative training set in regards to skin color (ethnicity) and provided a representative sampling of lighting conditions throughout the image. An accuracy of 93.8% is reported when the skin classification algorithm is applied to the above data set. Classification accuracy here is defined as when the number of correctly classified pixels based off of manually determined ground truth divided by total image pixels count is greater than 75%.

Fig. 1 contains sample complete and incomplete skin classifications, respectively. Part (b) in each case is the original RGB image converted to the sHSV color space, where the shifted hue, saturation, and value color components are displayed as the red, green, and blue RGB components, respectively. As seen especially within the right half of the second subject's face, low-light conditions tend to distort hue information. Note the substantial increase in the shifted hue value (displayed as red) in (b) over the right half of the subject's face. Similarly, overly bright conditions were also seen to distort the hue information and disrupt skin classification.

Nonetheless, over- and underexposure occurred in less than 5% of all images tested and the resulting 93.8% accuracy of the skin classifier remains robust within the visually noisy unconstrained environment. Despite this robustness, the complex, cluttered, and ever-changing background environment still manages to yield significant false positives within each frame. The following section discusses how each classified image is filtered to reduce these false positives and better isolate face candidates for subsequent detection.
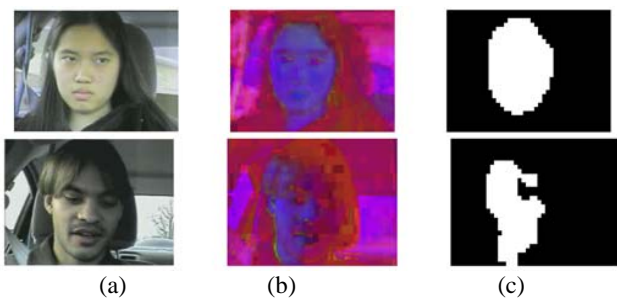

Fig. 2: Sample Successful Skin Classification (a) Original RGB Image (b) sHSV Image Displayed as RGB (c) Skin Classified Binary Image

III. FILTERING AN D BINARY CLUSTERING

The unprocessed skin-classified binary images in general suffer from two main undesirable effects. Impulse noise exists throughout the binary image and larger, false-positive regions tend to dominate background (non-skin) regions. As the skin-classified binary image will be used to locate the skin candidate for face detection, it is critical that these types of noise are reduced as much as possible.

The discussed single-element impulse noise manifests itself as false positives within background regions as well as false negatives within skin regions, namely within the face. Outlined in part within the green bounding boxes, Fig. 3(b) illustrates the appearance of impulse noise within skin classified region resulting from the original image in (a). While median filtering is generally used to combat salt-and-pepper noise, this method assumes equal undesirability of each false classification. 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.
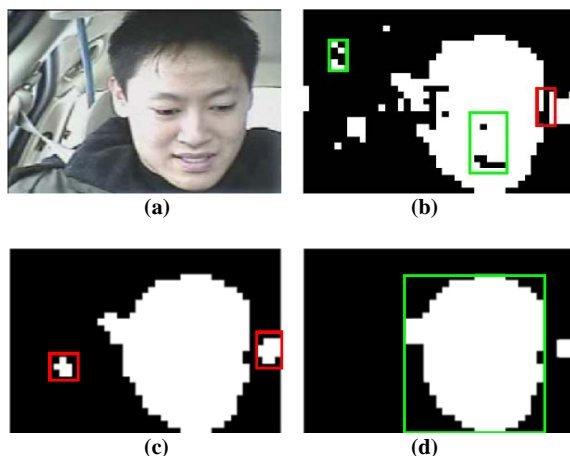

Fig. 3. Sample Post-Processing Imagery by Step
(a) Original Image (b) Skin Classified Binary Image
(c) 33rd Percentile Filtered (d) Application of Opening Operation

Larger regions of false classification can also be problematic when attempting to locate a face within a frame. Fig.3(c) outlines such falsely classified skin clusters within the red bounding boxes. To minimize the effect of these larger elements in the background, the binary morphological operations opening is utilized. Notice the elimination of the leftmost background cluster in (c) and the reduction in size of the rightmost cluster which was at least the size of the structuring element. 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 face detection algorithm. Note this assumption could be easily extended to multiple candidates per image if desired for future work.

IV. 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 algorithms. Building upon previous work [13], a joint shifted hue and saturation feature space was selected as the basis for face detection since it captures skin color information as well as the variation in saturation incurred around facial features such as eyes, nose, and mouth. With the feature space selected, another design decision was to approximate the joint probability density function as a histogram which quantizes the discussed two-dimensional feature space into a finite number of bins. A histogram is a nonparametric density estimation method which yields memory efficient and intuitive results. Moreover, quantizing the model and candidate's density functions achieves two important goals. First, the histogram approach reduces the computational complexity of PDF estimation and subsequent comparison. Furthermore, histogram approximation maintains a scale- and rotation-invariant comparison environment. In addition, the Epanechnikov kernel weights a pixels contribution to the estimated PDF per its spatial location.
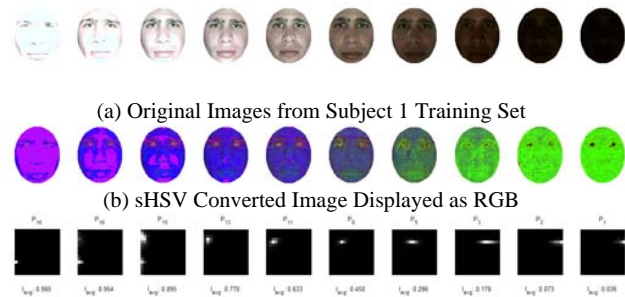
A. Forming the Face Model Joint Density Estimator

While previous work based on the same AVICAR database employed three face models extracted directly from arbitrarily chosen images containing light-, medium-, and dark-skinned individuals [13], it is the goal of this work to consolidate the face model into a single, cohesive, and more representative model. 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, which will be shown below. 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 average 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, a training set containing 150 images from five individuals of varying skin tone taken under a range of ambient lighting conditions was established. These subjects were centered in front of a video camera utilizing the same AVI compression and comparable resolution employed by the AVICAR database. Subjects were instructed to maintain a neutral, expressionless face while a series of images were taken under lighting conditions ranging from bright to dark. 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 one of the five subjects are detailed in Fig. 4. It can be seen that changes in average 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.

Hence 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. The resulting face model PDF histogram approximation across each illumination level is displayed in Fig. 5. Here the value of I_{bin} refers to the illumination component value which corresponds to the center (midpoint) of the discrete illumination bin, i. This face model histogram set $\mathbf{Q}_i$ will be stored in memory to be accessed by the face detection algorithm discussed in Section 5 to follow.

(a) Original Images from Subject 1 Training Set

(b) sHSV Converted Image Displayed as RGB

(c) Corresponding Joint sHue and Saturation Histogram
Fig. 4. Face Model Illumination Dependence Training Set.

B. 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 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.
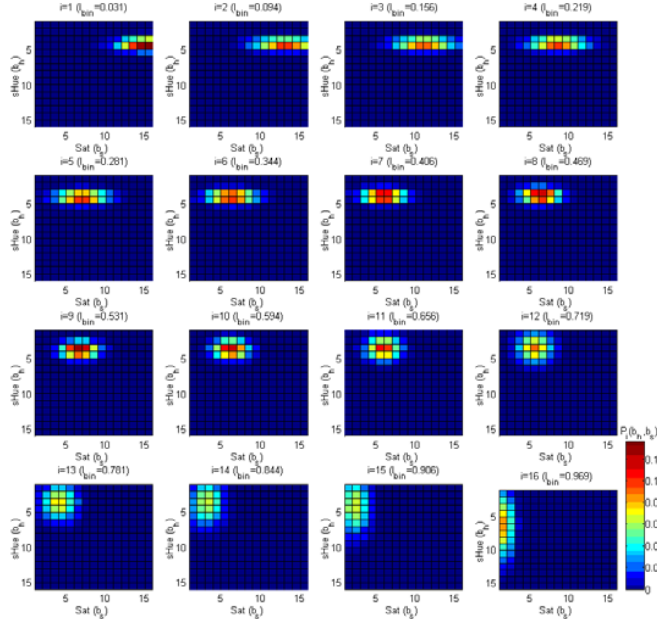

Fig. 5. Joint sHue and Saturation Histogram-Estimated PDF's over Average Illumination Bin Number

V. 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 to measure the similarity between the generated face model joint histogram and that of a candidate ROI. An important advantage of the Bhattacharyya coefficient calculation is that it does not require statistical measures from each distribution, significantly reducing computation time and complexity.

Remapping the definition of the Bhattacharyya to two dimensions, the Bhattacharyya coefficient can be defined as

$$\rho(\mathbf{P},\mathbf{Q}) = \sum_{h=1}^{m}\sum_{s=1}^{n}\sqrt{P(h,s)\cdot Q(h,s)} \qquad (3)$$

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 $[h, s]$. When both distributions are equal, the Bhattacharyya coefficient equals unity, meaning a perfect match. Conversely, lower valued coefficients indicate a poor match between the two distributions. Hence, a Bhattacharyya coefficient face detection scheme will be implemented such that sufficiently high ρ values will result in face classification of the candidate ROI. Based on iterative analysis over a training set of 160 images, Bhattacharyya coefficient of 0.5 was selected as the threshold to minimize false negative and false positive error rates.

To test the performance of the face detection algorithm, another 160-image test set was created from the AVICAR database, not containing any image found in the face-model or skin classification training sets. 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. 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. Note that of the 160 images tested 147 incurred successful face localization.

As can be seen, the face detection algorithm achieved an overall accuracy of 90% across the 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.6 (a) and (b), respectively.
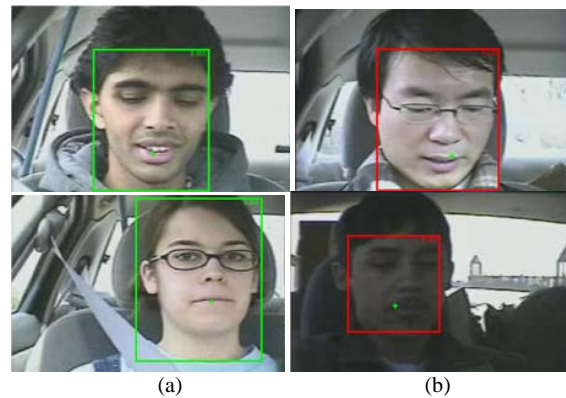

Fig.6. Sample (a) Positive Face and (b) Negative Face Detections

Significant sources of negative face detections involve changes in lighting conditions, specifically in dark environments. Ninety-percent of false negative classifications resulted from average candidate illumination values less than 0.5, or illumination level 8. Additionally, shifts in light chromaticity (color) away from "pure" white light significantly altered facial candidate's spectral content within the shifted hue and saturation feature space. Time of day and reflective surfaces around the car are two of many factors which have the ability to change the spectral content of ambient (visible) light. Illustration of this effect can be shown via the contrast in ambient lighting between Fig.7 (a) and (b). From the relatively white ambient lighting conditions in (a) to the less luminous and yellow-colored light in (b) a noticeable positive, 2-bin hue shift occurs in the candidate's peak histogram density consistent with this change in lighting conditions. Note that the green bounding box and text in Fig.7 (a) indicates a positive face detection with a Bhattacharyya coefficient of 0.878, while the red bounding box and text in (b) indicates a negative face detection with a coefficient of 0.422.

TABLE I
FACE DETECTION RESULTS
*Refer to Section III for definition

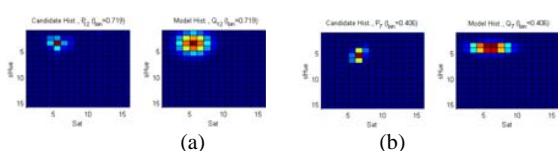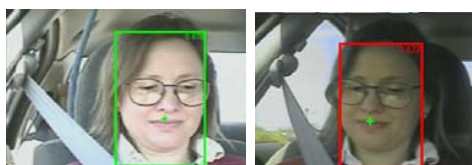| Face Detection	Successful Localization Set*		Complete Test Set	
	Instances	Percentage	Instances	Percentage
True Positive	139	94.6%	144	90.0%
False Negative	8	5.45%	16	10.0%
Total Images	147		160	

(a)   (b)

Fig.7. Effect of Ambient Light Chromaticity on Face Detection
Original RGB Image, Face Candidate ROI, and Model-Candidate
Histogram Pair for (a) Face Detection Success and (b) Face
Detection Failure with Same Subject

VI. CONCLUSION AND FUTURE WORK

Relative to previous work, positive face detection rates rose from 75% to 90% [13]. Among many techniques considered, the unique illumination-dependent face model and the adjusted skin classifier via filtering are considered successful and critical to the stated performance increase in face detection.

Despite the stated performance increases, common sources of error throughout the testing process highlight important system limitations of this face localization algorithm. Among these issues are limited image resolution, skin-colored car environments, and overly bright and dark operating conditions without sufficient image dynamic range. To mitigate the effects of dark lighting conditions or colored ambient lighting, techniques that improve color constancy should be considered. Color constancy is the ability to measure color of objects without the influence of the color of the light source. To solve the problem one estimates the color of the light source and remove it. Some techniques one might consider include Max-RGB and Grey-world [14,15]. In addition, the temporal element of a video sequence can also be incorporated into the face localization and/or the face model itself. Through tracking algorithms and periodic candidate and model updates, the face classification accuracy of 90% could potentially be increased further. Nonetheless, the performance of the skin classifier, filtering, face candidate localization, and face classifier algorithms yielded commendable results in the unconstrained car environment captured within the AVICAR database.

REFERENCES

[1] G. Potamianos, J. Luettin, and I. Matthews, "Audio-Visual Automatic Speech Recognition: An Oveview," *Issues in Visual and Audio-Visual Speech Processing*, MIT Press, rCh 10, 2004.
[2] D.G. Stork and M.E. Hennecke, "Speechreading by Humans and Machines" in NATO ASI Series F, vol 150, Springer Verlag, 1996.
[3] B. Lee, M. Hasegawa-Johnson, C. Goudeseune, S. Kamdar, S. Borys, M. Liu, T. Huang, "AVICAR: Audio-Visual Speech Corpus in a Car Environment," *INTERSPEECH2004-ICSLP*, 2004.
[4] Viola, Jones, "Robust Real-time Object Detection," IJCV 2001.
[5] P. Delmas, M. Lievin, "From Face Features Analysis to Automatic Lip Reading. *Seventh International Conference on Control, Automation, Robotics and Vision*, vol. 3, Dec. 2-5, 2002.
[6] K. Kumar, C. Tsuhan, R.M. Stern, "Profile View Lip Reading," *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 4, April 15-20, 2007.
[7] L.G. da Silveira, J. Facon, D.L. Borges, "Visual Speech Recognition: A Solution from Feature Extraction to Words Classification," sibgrapi, *XVI Brazilian Symposium on Computer Graphics and Image Processing*, 2003.
[8] M.C. Shin, K.I. Chang, L.V. Tsap, "Does Color Space Transformation Make Any Difference on Skin Detection?" *WACV: Proceedings of the Sixth IEEE Workshop on Applications of Computer Vision*, Washington DC, IEEE Computer Society, 2002.
[9] M.J. Jones, J.M. Rehg, "Statistical Color Models with Application to Skin Detection," *International Journal of Computer Vision*, vol. 46, no.1, 2006
[10] Y. Ming-Hsuan, A. Narendra, "Detecting Human Faces in Color Images," *Proceedings of the International Conference on Image Processing*, vol. 1, 1998.
[11] X. Zhang, H. A. Montoya, and B. Crow, "Finding Lips in Unconstrained Imagery for Improved Automatic Speech Recognition," *Proceedings of 9th International Conference on Visual Information Systems*, 2007.
[12] M. Abdel-Mottaleb, A. Ellgammal, "Face Detection in Complex Environments from Color Images," *Proceedings of the International Conference on Image Processing*, vol.3, 1999
[13] B. Crow, "Automated Location and Tracking of Facial Features in an Unconstrained Environment," Master's Thesis, California Polytechnic State University, 2008.
[14] J. van de Weijer, Th. Gevers, and A. Gijsenij, "Edge-based color constancy," in *Trans. On Image Processing*, 2007.
[15] G.D. Finlayson and E. Trezzi, , "Shades of gray and colour constancy," in *Proc. Of the 12th Color Imaging Conference,* 2004.