Lip Localization Algorithm Using Gabor Filters

Robert E. Hursig[1]**, Jane Xiaozheng Zhang**[2]**, and Chiweng Kam**[2]
[1]Sandia National Laboratories, USA
[2]Electrical Engineering Department, California Polytechnic State University, San Luis Obispo, CA, USA

Abstract – *This paper describes a lip localization algorithm within a still image frame for subsequent tracking and audio-visual speech recognition processing. A Gabor filter-based feature space is promoted as a means to localize lips within an image based off of shape. This filtered space is shown to effectively differentiate facial features, including lips, from their backgrounds and to bound the full extent of the lips within a face-classified region of interest. Extensive training and test sets are used to justify design decisions and performance.*

Keywords: Audio-Visual Automatic Speech Recognition (AVASR), lip localization, Gabor filter

1 Introduction

Audio-Visual Automatic Speech Recognition (AVASR) aims at improving Automatic Speech Recognition (ASR) by exploiting additional information contained in visual channel. Despite years of research attention on AVASR [1-3], visual speech information has yet to become incorporated into mainstream ASR. One major obstacle in the development of a practical AVASR system is the difficulty in creating a system that reliably detects and extracts lips in an unconstrained imagery. Within a real-world environment, AVASR systems must compete with constantly changing lighting conditions and background clutter as well as subject movement in three dimensions. In this paper, we focus on the problem of lip detection and extraction within a still image in an unconstrained visual environment.

Since the first AVASR system developed by Petajan in 1984 [4], numerous techniques have been reported in the literature that enable detection and extraction of the lips in an image. These techniques can be broadly classified into two categories: region-based and contour-based approaches.

In a region-based approach, pixels classified as lip pixels are grouped together and identified as region of interest (ROI). One simple method is to perform a thresholding operation on luminance or a particular chromatic component. Various thresholding methods were proposed including iterative threshoding [5], Fuzzy-based method [6], and adaptive threshoding [7,8]. Automatic computation of robust thresholds in various lighting conditions is the main challenge and limitation. However, simple thresholding methods are often used as an efficient preprocessing step to reduce the search range before other sophisticated methods are employed. Besides thresholding, others use various statistical techniques such as Bayesian decision theory [9,10,11], Markov random fields [12], neural networks [13], support vector machines [14], fuzzy C-means clustering [15], and deep belief networks fused with particle filtering [16]. A more recent approach applied Adaboost classifier using Haar-like features for lip segmentation [17-19] following the successful Adaboost algorithm for face detection by Viola and Jones [20].

The algorithms in a contour-based approach are based on deformable models and are divided into two categories: active contours and parametric models. Both employ an appropriate mathematical model of the contour (a spline or a parametric curve), and the model's energy terms, internal term for curve geometric properties, and external term from image data are then minimized to lock the initial contour to the object of interest. In Active Contours, also known as "snake", model points that define the active contour are modified one by one to the edges [21]. An application of snake and a gradient criterion to detect lips is found in [22]. Several other snake algorithms were proposed and used for designing robust lip segmentation, for example, the Gradient Vector Flow (GVF) snake is used by [23]; the genetic snakes is implemented by [24], and B-splines snake is utilized by [25]. In parametric model, also known as deformable template, an explicit assumption about the object shape is utilized. In [26,27], the parametric curves are several independent cubic polynomial curves and they are fitted by using gradient information based on hue and luminance. In [28], a parametric model of two cubic curves and a broken line is used for closed mouth whereas four cubic curves are used for open mouth. Recently, instead of parametric models with varied parameters, another approach is proposed using improved level set method for lip contour detection [29].

Unfortunately it is difficult to determine the best lip segmentation algorithm because there are no unified database or benchmark test images. Furthermore, there is no unified performance evaluation criterion or protocol among researchers to evaluate the result quantitatively.

While most of the developed techniques provided satisfying results, the extensive calculations demanded are significant. More importantly, a majority of existing lip localization techniques focus on lip extraction within controlled environments with ample image resolution and fail in real-world environments.

In this work we directly address unconstrained imagery in development of the visual front end. Generally, the in-car audio-visual environment can be considered as a worst-case scenario for AVASR. Background noise and mechanical vibrations from moving 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 [30] 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 of varying speed. Video and image resolution for this database is 240-by-360 pixels, height-by-width.

Thus, the goal of this work is to develop a robust still image lip localization algorithm designed as a visual front end of a practical AVASR system that is subject to the challenging real-world environment. In the next section, we present our Gabor filter-based facial feature extraction for lip localization. Extensive training and test sets will be used to justify design decisions and performance.

2 Preliminaries on Gabor Filters

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

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

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

$$\text{with } \alpha= {}^{F_o}\!/\!_{\gamma},\ \beta= {}^{F_o}\!/\!_{\eta},\ x_o \quad {}^{N_x}\!/\!_{2},\ y_o \quad {}^{N_y}\!/\!_{2} \qquad (1)$$

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 [31].

Utilizing a 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,

$$\begin{aligned}
&\mathbf{G} \quad \left\{G_{n,t,f} \quad G(x,y\,|\,\theta \quad \theta_t,F_o \quad F_f,N_x \quad N_n,N_y \quad N_n,\gamma,\eta,\phi)\right\}\\
&N_n \in \left\{\text{floor}\left(\frac{M_c}{8}\right), \text{floor}\left(\frac{M_c}{4}\right)\right\} \quad n \quad 1,2\\
&\theta_t \in \left\{\frac{3\pi}{8},\ \frac{\pi}{2},\ \frac{5\pi}{8}\right\} \quad t \quad 1,2,3\\
&F_f \in \left\{\frac{4}{N_n},\frac{8}{N_n}\right\} \quad f \quad 1,2
\end{aligned} \qquad (2)$$

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

where G is defined in Eq. (1) 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 Gabor Filtering Algorithm

With the establishment of the lip-specific Gabor filter set, processing of the face-classified region of interest can proceed. Here we assume the face region has already been obtained by using face detection algorithms such as in [20,32]. In the following, the sHSV triplet's value component (in other word, the intensity component) is selected as the feature space of choice for Gabor filtering since it best separates similarly colored lip and surrounding skin.

Figure 1 contains a block diagram of the entire Gabor filtering processing. 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, $G_{n,t,f}$. Next, all 12 Gabor responses are summarized element by element such that the pixel value at any location within the candidate's ROI is the sum of each Gabor responses, also called Gabor jets, at the same location. We denote the M_c-by-N_c total Gabor response as G_{tot}, where M_c and N_c are the row and column sizes of the face candidate, respectively.

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 normalization and remapping procedure is defined below as

$$G_f(r,c) \quad 2\left|G_{norm}(r,c)-0.5\right| \qquad \begin{matrix} r & 1,2,\ldots,M_c \\ c & 1,2,\ldots,N_c \end{matrix} \tag{3}$$

$$\text{where } G_{norm}(r,c) \quad \frac{G_f(r,c)-\min_{r,c}(G_f)}{\max_{r,c}(G_f)}$$

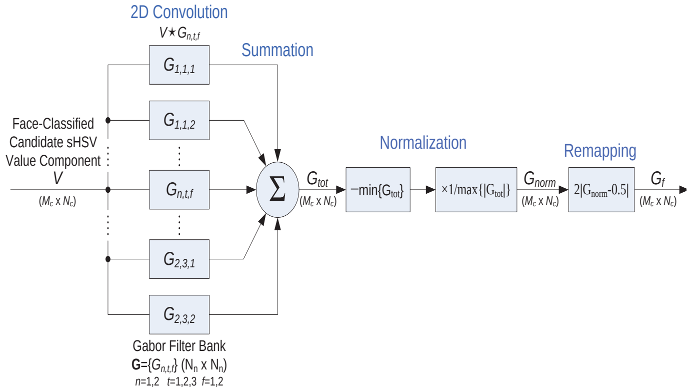


Fig. 1. **Gabor Filtering Process Block Diagram**

Let the final, normalized, and remapped Gabor filter response be defined as G_f, which has size M_c-by-N_c. In the case of zero phase shift, ϕ, normalization preceded remapping as the negative modes of the Gabor filter are attenuated more heavily by the Gaussian envelope than the central, positive mode. Supporting the need for the remapping process, an illumination-invariant design demands detection of absolute changes in achromatic intensity—both from high to low and low to high illumination. Figure 2 shows sample total Gabor filter responses of several images. Referring to Fig.2 (a) and (b), the cross section of the lip from chin to the region above the lip involves many such oscillatory changes in illumination value. Note the contrast facial features have against the face's background. Smooth skin surfaces, such as the cheeks, provide minimal response while the mouth opening, lips, nostrils, eyes, and eyebrows provide much elevated responses. This phenomenon can be attributed to the spatial transitions in illumination (both positive and negative) around these features. Interestingly, facial hair increases contrast between the hair and facial features, as is seen in the bottommost subject in Fig. 2. Also note that the near-vertical edges of the face provide low responses while the near-horizontal edges, such as the chin region, provide more noticeable responses. With these positive feature qualities, the final Gabor filter response will now be used as the preferred feature space for lip localization in the following sections.

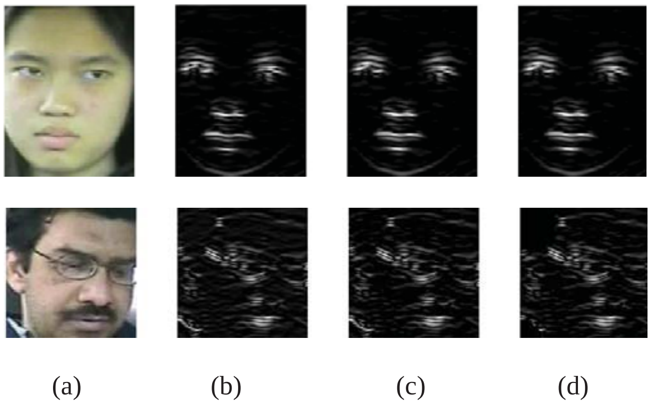


Fig. 2. **Sample Total Gabor Filter Responses** (a) original RGB image, (b) total Gabor response, G_f, (c) mean-removed total response, (d) mean-removed and masked total response, G_{mr}.

4 Lip Center Coordinate Estimation

To further remove false responses, 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. Furthermore, to remove false pisitives within the background surrounding the face, the skin-classified binary mask is applied over the mean removed response. Given the mean-removed and masked Gabor response, G_{mr}, a number of possible lip locations, called seeds, will be generated. Following seed generation, key parameters which are indicative of the presence of lips will be calculated. Utilizing these parameters, a figure of merit will be then calculated for each seed point.

Here, a column concentration signal, D_c, is first calculated from the G_{mr}.

$$D_c(r) \quad \sum_{c\ 1}^{N_c} G_{mr}(r,c) \qquad \begin{matrix} r & 1,2,\ldots,M_c \\ c & 1,2,\ldots,N_c \end{matrix} \tag{4}$$

Even though the subject face within the ROI may not be perfectly vertical, the column concentration still conveys general information about how the filter response is distributed throughout the image's vertical axis.

Next, the row mean is defined as

$$\mu_r \quad \sum_{r\ 1}^{M_c} r \frac{D_c(r)}{N_c} \tag{5}$$

Also let ROI_{valid} denote the region of the candidate's ROI defined by $\mu_r \le r \le \text{floor}\left(p_{bor} \cdot M_c\right) =$ and $\text{floor}\left((1-p_{bor}) \cdot N_c\right) \le c \le \text{floor}\left(p_{bor} \cdot N_c\right)$, where p_{bor} is experimentally set to 0.95 to eliminate false response returns at the candidate ROI's border.

Now, the seed point locations can be generated by finding the peaks of the column concentration signal over the region ROI_{valid}. Peak, or local maximum, detection is achieved via locating the concentration signal indices over the stated range, $\mu_r \le r \le \text{floor}\left(p \cdot M_c\right)$, such that the concentration immediately above and below that peak is less. Constant sequential concentration values are handled such that the peak occurs at the midpoint of the "plateau." Moreover, all peaks of height below a minimum peak height, denoted *mph*, are not considered. Let P_j and H_j be the row index and peak height, respectively, of the j^{th} peak of the column concentration signal, D_c, within the ROI_{valid} space. Now, let the seed point locations for lip coordinate estimation be

$$\begin{aligned}\mathbf{r_{pk}} &\quad \{r_{pk,i}\} \quad \{P_j \mid H_j \geq mph\} \quad \forall j \qquad i \quad 1,2,...,n_{cand}\\ \mathbf{h_{pk}} &\quad \{h_{pk,i}\} \quad \{H_j \mid H_j \geq mph\} \quad \forall j\end{aligned} \tag{6}$$

(4.1)

with $mph \quad \dfrac{1}{M_c}\sum_{r \;\; 1}^{M_c} \cancel{D}_c(r)$

where n_{cand} is the number of returned seeds, $r_{pk,i}$ is i^{th} row index for the peak of height $h_{pk,i}$ within the ROI_{valid} space which is at least as much as *mph*. Here, the minimum peak height was empirically selected as the sample mean of the column concentration signal. Completing the seed points' Euclidean coordinates, let $c_{pk,i,}$ be the i^{th} element of the set $\mathbf{c_{pk}}$, be defined as the midpoint coordinate of the longest consecutive non-zero chain in row $r_{pk,i}$ of the image's total Gabor response. Hence, the i^{th} seed point now has the location vector $[r_{pk,i}, c_{pk,i}]$.

Combined, the $\mathbf{r_{pk}}$ and $\mathbf{c_{pk}}$ sets convey spatial location within the ROI. Due to the striation of the lip's structure from chin to nose, the concentration of peak Gabor responses along the row-axis is also pertinent to lip localization. Hence, let the peak concentration set be defined as

$$\mathbf{D_{pk}} \quad \{D_{pk,i}\} \quad \left\{ \sum_{= r_{pk,i}-w}^{r_{pk,i}+w} = r \in r_{pk,i} \right\}^{=} \forall i, \;\; w \quad \text{floor}(M_c/5) \tag{7}$$

where $\mathbf{D_{pk}}$ is the set of n_{cand} concentration values, $D_{pk,i}$, which are the sum of all peaks contained within the $2w$+1-row window centered about $r_{pk,i}$. The windowing value of w was experimentally determined to not exceed the ratio of average lip height to ROI height and to provide optimal results within the lip region. Lastly, let the local two-dimensional mean-removed Gabor response concentration set be defined as

$$\mathbf{D_{loc}} \quad \{D_{loc,i}\} \quad \left\{ \sum_{= r_{pk,i}-w_r}^{r_{pk,i}+w_r} = \sum_{c \;\; c_{pk,i}-w}^{c_{pk,i}+w} = G_{mr}(r,c) \right\}^{=}_{=} \quad \forall i \tag{8}$$

where $w_r \quad \text{floor}\left({}^{w}\!/_{2} \right) =$

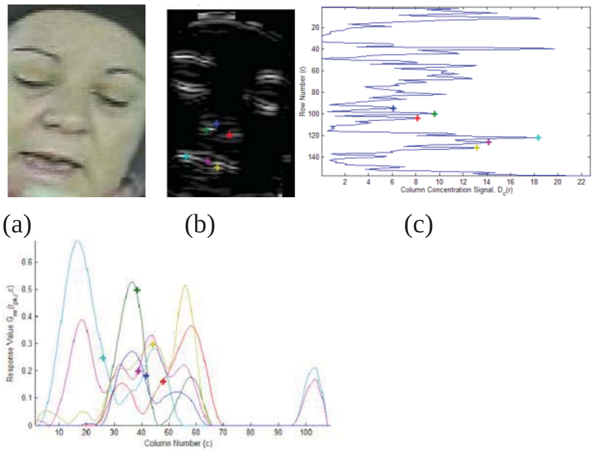
where $\mathrm{D_{loc}}$ is the set of n_{cand} local two-dimensional response concentration values, $D_{loc,i}$, at seed index i and w is defined in Equation (6).

Fig.3 contains a graphical overview of the lip coordinate estimation algorithm. Fig.3(b) shows the mean-removed and masked Gabor response, G_{mr}, overlaid with the seed locations indicated by the colored crosses. Fig.3(c) contains the concentration signal, $D_c(r)$, while part (d) contains the total Gabor response row signals at the seed row indices, $G_{mr}(r_{pk,i},c)$, respectively, with the row and column seed components, respectively, overlaid on the signal.

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

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

where **FOM** is the set of all figure of merit values, FOM_i, at seed index i, D_{pk} is the peak density, D_{loc} is the local concentration, 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.



(a) (b) (c)

(d)

Fig.3. **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 D_c Plot, (d) Seed Row's Column Signals.

Hence it has been argued that the lip's central coordinates are the coordinates for which the established figure of merit is maximal. In other words, the estimated lip central coordinate are now given by

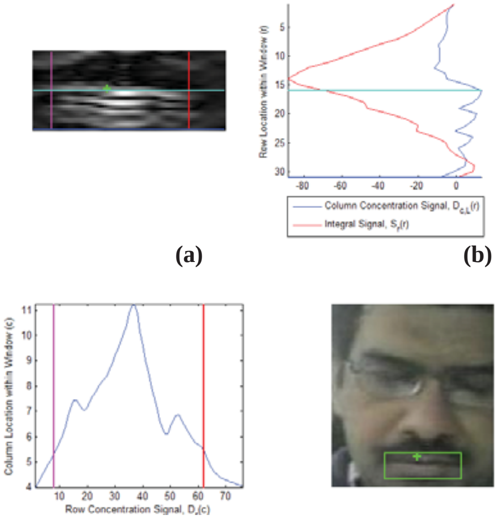
$$\mathbf{x_{est}} \quad [r_{est}, c_{est}] \quad [r_{pk,i}, c_{pk,i}] \Big| i \quad \arg_i \max\left(FOM_i\right) = \tag{10}$$

where $\mathbf{x_{est}}$ is the lip's estimated coordinates relative to the candidate image's coordinates FOM_i as defined in Equation (9), and r_{est} and c_{est} are the row and column estimated lip locations, respectively.

The lip center coordinate estimator was applied to160 test images from AVICAR database. 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 [32]. 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.

5 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. 4 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.



(a) **(b)**

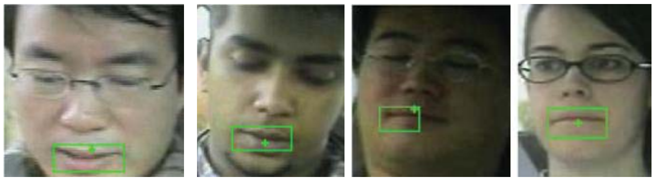
(c) **(d)**

Fig.4. 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, 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 (lines) 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.5(a) and (b), respectively. When applied to the 160-image test set, factoring in face detection, which has an accuracy of 90% from our previous work [32], the overall accuracy for lip detection is 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. 5 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.

6 Conclusions and Future Work

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.



(a) **(b)**

Fig.5. Sample Lip Localization (a) Success and (b) Failures

Common sources of error include 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/bright conditions and colored ambient lighting, techniques that improve color constancy should be considered in future work, such as Grey-world and Max-RGB algorithms. In addition, inclusion of temporal information and 3D information should further improve the system performance.

7 References

[1] D. G. Stork and M.E. Hennecke, "Speechreading by Humans and Machines" in NATO ASI Series F, vol. 150, Springer Verlag, 1996.

[2] G. Potamianos, J. Luettin, and I. Matthrews, "Audio-Visual Automatic Speech Recognition: An Overview" Issues in Visual and Audio-Visual Speech Processing, G. Bailly, E. Vatikiotis, and Perrier, Eds, MIT Press, Ch 10, 2004.

[3] A. Liew, S. Wang, " Visual Speech Recognition: Lip Segmentation and Mapping", Medical Information Science Reference, 2009.

[4] E.D. Petajan, "Automatic Lipreading to Enhance Speech Recognition", Ph.D thesis, University of Illinois at Urbana-Champaign, 1984

[5] R. Gocke, "Audio-video automatic speech recognition: an example of improved performance through multimodal sensor input", NICTAHCSNet Multimodal User Interaction Workshop, 2005

[6] T.T. Pham, M.G. Song, J.Y. Kim, S. Y. Nam and S.T. Hwang, "A Robust Lip Center Detection in Cell Phone Environment", Int. Symposium on Signal Processing and Information Technology, 2008.

[7] J.M. Zhang, L.M. Wang, D.J. Niu, and Y.Z. Zhan, "Research and implementation of a real time approach to lip

detection in video sequences," Int. Conf. on Machine Learning and Cybernetics, 2003.

[8] L. Wang, J. Xu, Y. Zhao, "Research of Visual Features Detection and Tracking Methods about Audio-Visual Bimodal Speech Recognition," International Forum on Information Technology and Applications, 2010.

[9] M. Sadeghi, J. Kittler, and K. Messer, "Modelling and Segmentation of lip area in face images", in IEEE Proceedings on Vision, Image and Signal Processing, 2002.

[10] P. Lucey and G. Potamianos, "Lipreading Using Profile Versus Frontal Views," IEEE 8th Workshop on Multimedia Signal Processing, 2006.

[11] B. Crow, H.A. Montoya, and X. Zhang, "Finding Lips in Unconstrained Imagery for Improved Automatic Speech Recognition," 9th Int. Conference on Visual Information Systems, 2007.

[12] X. Zhang and R.M. Mersereau, "Lip feature extraction towards an automatic speechreading system," International Conference on Image Processing, 2000.

[13] P. Daubias and P. Deleglise, "Statistical Lip-Appearance Models Trained Automatically Using Audio Information," AURASIP Journal on Applied Signal Processing, 2002.

[14] B. Castaneda and J.C. Cockburn, "Reduced support vector machine applied to real-time face tracking," in IEEE ICASSP, 2005.

[15] S. Wang, A.W. Liew, W.H. Lau, and S.H. Leung, "Lip Region Segmentation with Complex Background," in [3].

[16] G. Carneiro and J.C. Nascimento, "The Fusion of Deep Learning Architectures and Particle Filtering Applied to Lip Tracking," ICPR 2010.

[17] Y.H. Huang, B.C. Pan, J. Liang, and X.Y. Fan, "A new lip automatic detection and location algorithm in lip-reading system," IEEE International Conference on Systems Man and Cybernetics, 2010.

[18] R. Navarathna, P. Lucey, D. Dean, C. Fookes, and S. Sridharan, "Lip detection for audio-visual speech recognition in car environment," 10th International Conference on Information Sciences Signal Processing and their Applications, 2010.

[19] L. Wang, X. Wang, and J. Xu, "Lip Detection and Tracking Using Variance Based Haar-Like Features and Kalman Filter," 5th International Conference on Frontier of Computer Science and Technology, 2010.

[20] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2001.

[21] M. Kass, A. Witkin, and D. Terzopoulos, "Snake: Active contour models," International Journal of Computer Vision, 1987.

[22] P. Delmas, N. Eveno, and M. Lievin, "Towards Robust Lip Tracking," International Conference on Pattern Recognition, 2002.

[23] Z. Wu, A.Z. Petar, and A.K. Katsaggelos, "Lip Tracking for MPEG-4 Facial Animation," International Conference on Multimodal Interfaces, 2002.

[24] R. Seguier and N. Cladel, "Genetic Snakes: Application on Lipreading," International Conference on Artificial Neural Networks and Genetic Algorithms, 2003.

[25] T. Wakasugi, M. Nishiura, and K. Fukui, "Robust lip contour extraction using separability of multi-dimensional distributions," in Proceedings of 6th IEEE International Conference on Automatic Face and Gesture Recognition, 2004.

[26] N. Eveno, A. Caplier, and P.Y. Coulon, "Accurate and quasi-automatic lip tracking," IEEE Transactions on Circuits and Systems for Video Technology, 2004.

[27] N. Eveno, A. Caplier, and P.Y. Coulon, "A parametric model for realistic lip segmentation," 7th Int. Conf. on Control, Automation, Robotics and Vision, 2002.

[28] S. Stillittano, V. Girondel, and A. Caplier, "Inner and outer lip contour tracking using cubic curve parametric models," 16th IEEE International Conference on Image Processing, 2009.

[29] K. Li, M. Wang, M. Liu, and A. Zhao, "Improved level set method for lip contour detection," ICIP, 2010.

[30] 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.

[31] J. Kamarainen, V. Kyrki, "Invariance Properties of Gabor Filter-Based Features – Overview and Applications," in IEEE Transactions on Image Processing, vol.15, no. 5, 2006.

[32] R.E. Hursig, and X. Zhang, "Face Localization Using Illumination-dependent Face Model for Visual Speech Recognition," in Proc. WASET 10th International Conference on Signal and Image Processing, 2010.