Finding Lips in Unconstrained Imagery for Improved Automatic Speech Recognition

Xiaozheng Jane Zhang, Higinio Ariel Montoya, and Brandon Crow

Abstract. Lip movement of a speaker conveys important visual speech information and can be exploited for Automatic Speech Recognition. While previous research demonstrated that visual modality is a viable tool for identifying speech, the visual information has yet to become utilized in mainstream ASR systems. One obstacle is the difficulty in building a robust visual front end that tracks lips accurately in a real-world condition. In this paper we present our current progress in addressing the issue. We examine the use of color information in detecting the lip region and report our results on the statistical analysis and modeling of lip hue images by examining hundreds of manually extracted lip images obtained from several databases. In addition to hue color, we also explore spatial and edge information derived from intensity and saturation images to improve the robustness of the lip detection. Successful application of this algorithm is demonstrated over imagery collected in visually challenging environments.

1 Introduction

Speech-based user interface allows a user to communicate with computers via voice instead of a mouse and keyboard. The use of speech interface in emerging multimedia applications is growing in popularity because it is more natural, easier, and safer to use. The key technology that permits the realization of a pervasive speech interface is automatic speech recognition (ASR). While ASR has witnessed significant progress in many well-defined applications, the performance of such systems degrades considerably in acoustically hostile environments such as in an automobile with background noise, or in a typical office environment with ringing telephones and noise from fans and human conversations. One way to overcome this limitation is to supplement the acoustic speech with visual signal that remains unaffected in noisy environment.

While previous research demonstrated that visual modality is a viable tool for identifying speech [1,2], the visual information has yet to become utilized in

mainstream ASR systems. One obstacle is the difficulty in building a robust visual front end that tracks lips accurately in a real-world condition. To date majority of the work in automatic speechreading has focused on databases collected in studio-like environments with uniform lighting and constant background, such as CMU database [3], XM2VTS database [4], Tulips1 [5], DAVID [6], CUAVE [7], and AVOZES [8]. Hence there is a high demand for creating a robust visual front end in realistic environments. Accurately detecting and tracking lips under varying environmental conditions and for a large group of population is a very difficult task due to large variations in illumination conditions, background, camera settings, facial structural components (beards, moustaches), and inherent differences due to age, gender, and race. In particular, when we consider a real-world environment, strong illumination, uneven light distribution and shadowing, cluttered/moving background can complicate the lip identification process considerably.

In this paper we present our current progress in obtaining a robust visual front end in real-world conditions. In Section 2, we examine the use of color information and report our results on the statistical analysis and modeling of lip pixels by examining hundreds of manually extracted lip images obtained from several databases. In Section 3 we explore the effectiveness of spatial and edge information in extracting the lips. Finally Section 4 offers our conclusions and future work.

2 Color-Based Image Segmentation

2.1 Statistical Modeling and Color Space Selection

Color is an important identifying feature of an object. Prominent colors can be used as a far more efficient search criterion for detecting and extracting the object, such as the red color for identifying the lips. While color has been studied extensively in the past decade, especially in the field of face recognition, limited work was done in finding lips reliably in real-world conditions. The first critical decision for using color to find lips is to determine the best color space. RGB is the most commonly used color space for color images. However its inability to separate the luminance and chromatic components of a color hinders the effectiveness of color in object recognition. Previous studies [9] have shown that even though different people have different skin colors, the major difference lies in the intensity rather than the color itself. To separate the chromatic and luminance components, various transformed color spaces can be employed. Two such color spaces, HSV and YCbCr, were frequently used in various previous work in face and lip identification [10,11]. To determine which color space works the best for identifying the lips in unconstrained imagery, we perform statistical comparisons between the two on 421 images of human faces collected from the following three databases: 1.) Images collected using a Logitech web cam in a car or office environment with natural lightings (these images were collected by us and no compression was used); 2.) Images collected by browsing the Internet (these images were primarily face shots of celebrities so the images were most likely

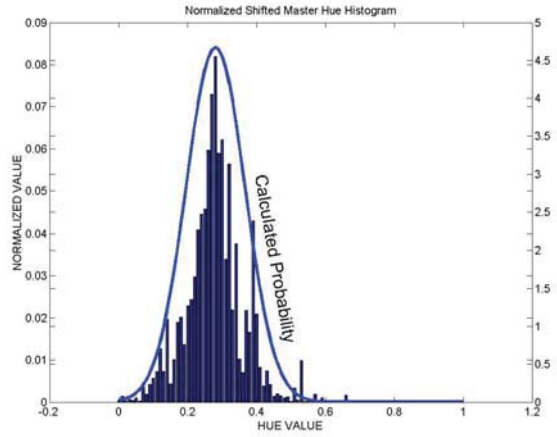
enhanced); 3.) Images collected from the AVICAR [12] Database, where 100 speakers were recorded in a natural automobile environment and under various driving conditions.

(a) (b) (c)

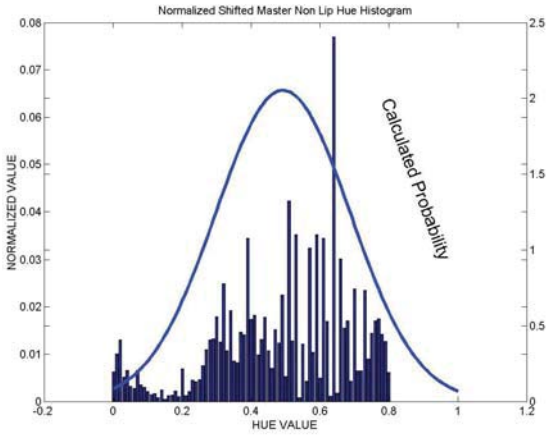
Fig. 1. Examples of Images from Three Databases

In our previous work [13], we reported the test results that were performed on all 421 images from the three databases. It was found that the Cb and Cr components offer very little distinction between the lip region pixel values and the background pixel values, while the hue component offers the best separation between the two based on the histograms shown in Figure 2. Hence the hue is better suited for distinguishing pixels between lip and non-lip regions and therefore was chosen as the primary feature for our visual front end. In Figure 2 the histogram was built by manually segmenting lip region for each of the 421 images and then normalized for equal comparison. In addition we shift the hue to the right by 0.2 because the red hue falls into two separate subsets at the low and high ends of the whole color range, as a result of the wrap-around nature of hue (hue is defined on a ring). After a right shift the red color falls in a connected region that lies at the low end close to 0. This can then be approximated by a Gaussian distribution that is shown together with the histogram in Figure 2(a). It can be shown that the hue values of the lip region are concentrated between zero and 0.6 with a mean of 0.28 and a variance of 0.0073.

Since the segmentation of the lip region will now occur in the shifted hue color space, then the histogram of the background must also be shifted to compare any overlapping sections. This shifted hue histogram together with the corresponding Gaussian model for background is shown in Figure 2(b). It is observed that the background has a much wider distribution since the background is un-controlled and varies greatly in an unconstrained environment. Note that both histograms for the lip and non-lip region are obtained by counting color pixels in manually segmented images from hundreds of images in our three databases. These histograms can well approximate the probability density functions (pdfs) for the lip and non-lips hue colors that are essential in the following lip segmentation procedure.



(a)



(b)

Fig. 2. Normalized Hue Histogram and Calculated PDFs for (a) Lips, (b) Background

2.2 Classification Design

To segment the lips a Bayesian classifier is employed. In Bayesian classifier a pixel x is classified as a lip pixel if the posteriori probability for the lips, $P(L \mid x)$, is larger than the posteriori probability for the background, $P(B \mid x)$, where L represents the class of lips, and B be the class for the background. By using Bayes Rule to calculate the posteriori probabilities, it can be shown that a pixel is classified as a lip pixel if

$$P[x \mid L]P[L] \geq P[x \mid B]P[B] \tag{1}$$

In equation (1), $P[x|L]$ and $P[x|B]$ are the class conditional densities according to the Gaussian models in Figures 2(a) and (b). $P[L]$ is the a priori probability for the lip region and can be estimated by computing the percent area of the lip region averaged

throughout all of our sample images. $P[B]$ is the a priori probability for the background and can be estimated by computing the percent area of the background region averaged throughout all of our sample images. The implementation of the Bayesian classifier results in the following segmentation results shown in Figure 3.

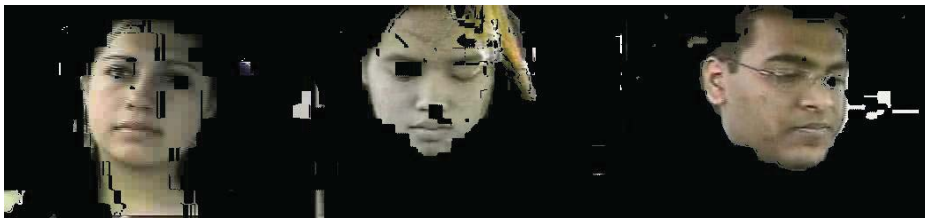


Fig. 3. Image Segmentation Results by Using Bayesian Classifier

The results of applying Bayesian classifier to our test images were that we were able to segment not only the lip region from the background, but also the entire face region. This result was neither expected nor desirable. To understand this we also calculated the hue pdf for skin around the lip region. To do so we scale the lip masks by 2 then subtract the original lip mask from the result. In Figure 4 all three hue pdfs for the lips, skin, and non-lip regions are combined in one graph for easy comparison. This figure clearly shows that there is very little hue contrast between the lips and facial skin, thus segmentation of lips from the surrounding face is difficult by using the hue color model only. However large variation of face color and the background scene allows an easy separation of both. We are able to use Bayesian classifier to successfully find faces in all images we tested. This is a significant result considering the fact that the images were collected in realistic environments that represent challenges due to large variations in illumination conditions.

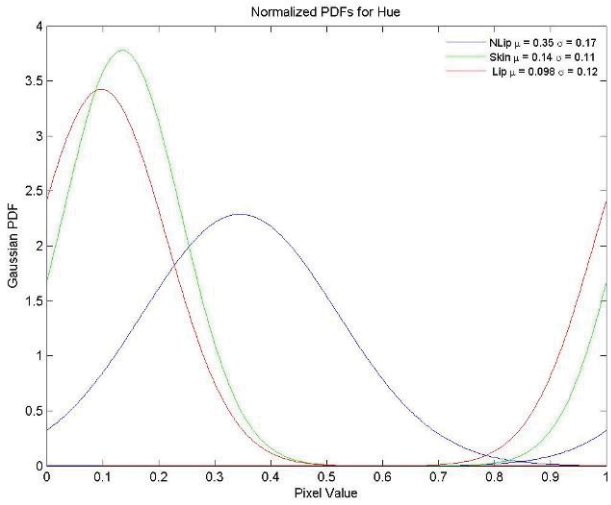


Fig. 4. Normalized Hue PDFs for Lip, skin, and non-lip regions (with no hue shift)

3 Finding Lips

To find the lip region from the face, we first reduce the search space by cropping the face image obtained from the previous module. Specifically, we remove the upper half and 1/8th from each side. The resulting image ("half_face") is then passed to the next module.

It was observed that all inner lips are bright, near horizontal lines in the saturation color space, see Figure 5(a). At the same time the intensity values of the lips are darker than the surrounding regions. We therefore use both the saturation and value components in the HSV space to determine the lip lines. We create a mask image where a pixel is assigned to a value of 1 if its value component is less than a pre-defined threshold T_{Value} and its saturation component is greater than $T_{Saturation}$. Experimentally T_{Value} is set to be 0.3, and $T_{Saturation}$ set to be 0.45. The resulting image in Figure 5(b) is then analyzed for the component that has the longest major axis with an angle from horizontal of less than 15°, this is indicated as * in Figure 5(c). Therefore, as long as the mouth is less than 15° off axis, it should be recognized. Once this component is found, all other components are removed; the resulting image

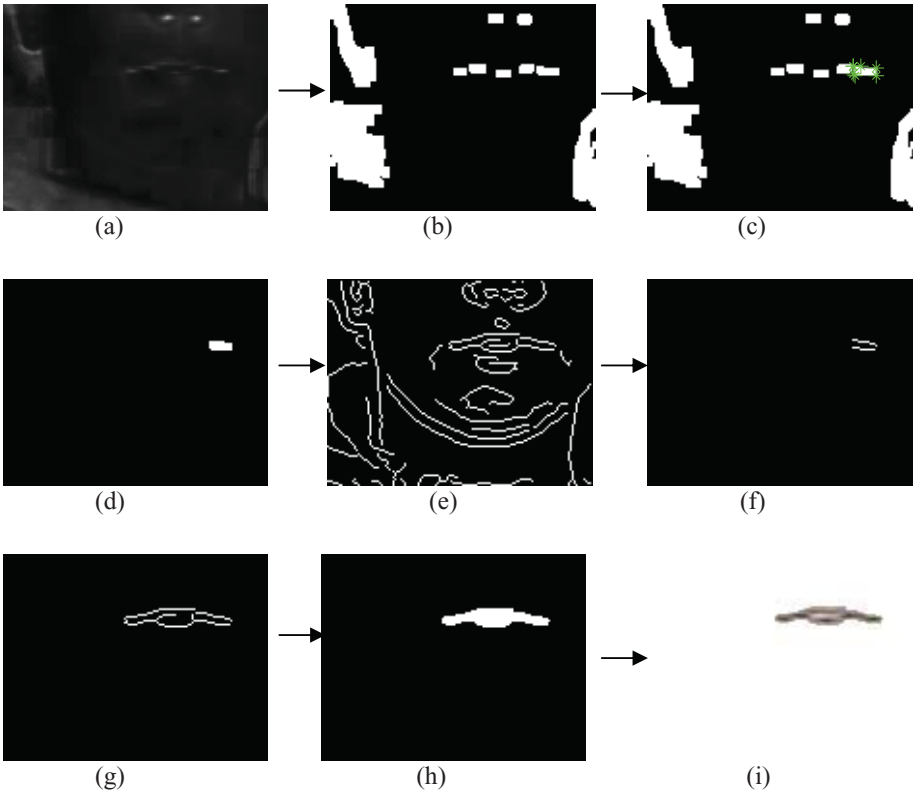


Fig. 5. Steps in Lip Finding Procedure

in Figure 5(d) is used as a mask in the subsequent step. Next a Canny edge detector is performed on the intensity values of the image and the resulting image in Figure 5(e) is multiplied by the mask image. The result is an image comprised of segments of Canny edge lines, see Figure 5(f). These segments are then analyzed for spatial location, which is then used to create a final mask image, comprised of the complete Canny components lying under the original mask. This image in Figure 5(g) is then dilated to force nearby components to connect, thus allowing filling of the region. Finally, this mask image in Figure 5(h) is inverted and then added to the original RGB image – resulting in a white background with only the lip region from the original image making it through as seen in Figure 5(i). When this procedure is applied to images in our databases, errors found are mostly due to weak edges from the lip lines extracted from Canny edge detector. Additionally, although the image was dilated prior to filling the final mask, not all regions are connected, preventing the complete mouth region from being displayed.

4 Conclusion and Future Work

Detection of lips in video sequences serves as an essential initial step towards building a robust audio-visual speech recognition system. In contrast to earlier works where the visual front end was designed based on databases collected in studio-like environment, this work seeks to find lips in imagery collected in visually challenging environment. A Bayesian classifier was developed that segments the face region in most unconstrained imagery. This was only possible after careful statistical analysis and modeling of hundreds of manually segmented lip images in various lighting and background conditions. These results demonstrate a very efficient way to narrow the search for the lip region substantially. From the detected face region, we subsequently extract lip region by incorporating spatial and edge information of the lips using both the saturation and value components in the HSV color space. From the detected lip region we can then extract physical dimensions of the lips that will be input to a recognition engine where spoken words can be classified.

In our future work we will explore additional features to further improve the robustness of our visual front end. One such feature is the motion feature. Here we notice that when a person is speaking, the motion fields around the lips are very different from those around the face. We expect this will result in improved accuracy in lip detection.

Acknowledgments. This work was sponsored by the *Department of the Navy, Office of Naval Research*, under Award # N00014-04-1-0436.

References

1. Stork, D.G., Hennecke, M.E. (eds.): Speechreading by Humans and Machines. NATO ASI Series F, vol. 150. Springer, Heidelberg (1996)
2. Potamianos, G., et al.: Audio-Visual Automatic Speech Recognition: An Overview. In: Bailly, G., Vatikiotis, E., Perrier, P. (eds.) Issues in Visual and Audio-Visual Speech Processing, MIT Press, Cambridge (2004)

3. Advanced Multimedia Processing Lab, Carnegie Mellon, Project-Audio-Visual-Processing (last accessed June 1, 2007), http://amp.ece.cmu.edu/projects/audiovisualspeechprocessing
4. The Extended M2VTS Database (last accessed June 1, 2007), http://www.ee.surrey.ac.uk/research/vssp/xm2vtsdb
5. Movellan, J.R.: Visual speech recognition with stochastic networks. In: Tesauro, G., Touretzky, D.S., Leen, T. (eds.) Advances in Neural Information Processing Systems, vol. 7, MIT Press, Cambridge, MA (1995)
6. Chibelushi, C.C., Gandon, S., Mason, J.S.D., Deravi, F., Johnston, R.D.: Design issues for a digital audio-visual integrated database. In: IEE Colloquium on Integrated Audio-Visual Processing for Recognition, Synthesis and Communication, Savoy Place, London (1996)
7. Patterson, E.K., Gurbuz, S., Tufekci, Z., Gowdy, J.N.: CUAVE: A new audio-visual database for multimodal human-computer interface research. In: Proc. ICASSP (2002)
8. Goecke, R., Millar, B.: The Audio-Video Australian English Speech Data Corpus AVOZES. In: Proc. ICSLP (2004)
9. Yang, J., Stiefelhagen, R., Meier, U., Waibel, A.: Real-time face and facial feature tracking and applications. In: Proc. of the 3rd IEEE Workshop on Applications of Computer Vision, pp. 142–147. IEEE Computer Society Press, Los Alamitos (1996)
10. Hsu, R., Abdel-Mottaleb, M., Jain, A.K.: Face detection in color images. J. IEEE Trans. Pattern Anal. Mach. Intell. 24(5), 696–706 (2002)
11. Zhang, X., Broun, C.C., Mersereau, R.M., Clements, M.A.: Automatic Speechreading with applications to human-computer interfaces. EURASIP Journal Applied Signal Processing, Special Issue on Audio-Visual Speech Processing 1, 1228–1247 (2002)
12. Lee, B., Hasegawa-Johnson, M., Goudeseune, C., Kamdar, S., Borys, S., Liu, M., Huang, T.: AVICAR: Audio-Visual Speech Corpus in a Car Environment. In: INTERSPEECH2004-ICSLP (2004)
13. Zhang, X., Montoya, H.A.: Statistical Modeling of Lip Color Features in Unconstrained Imagery. In: Proc. 11th World Multiconference on Systemics, Cybernetics and Informatic, Orlando, Florida (2007)