FACE AND LIP TRACKING IN UNCONSTRAINED IMAGERY FOR IMPROVED AUTOMATIC SPEECH RECOGNITION

Brandon Crow, Jane Xiaozheng Zhang
Department of Electrical Engineering, California Polytechnic State University
1. Grand Ave. San Luis Obispo, CA USA 93401

ABSTRACT

When combined with acoustical speech information, visual speech information (lip movement) significantly improves Automatic Speech Recognition (ASR) in acoustically noisy environments. Previous research has demonstrated that visual modality is a viable tool for identifying speech. However, the visual information has yet to become utilized in mainstream ASR systems due to the difficulty in accurately tracking lips in real-world conditions. This paper presents our current progress in tracking face and lips in visually challenging environments. Findings suggest the mean shift algorithm performs poorly for small regions, in this case the lips, but it achieves near 80% accuracy for facial tracking.

Keywords: Automatic visual speech recognition, face tracking, lip tracking, mean-shift, Bhattacharyya coefficient

1. INTRODUCTION

Automatic speech recognition (ASR) is a field of study concerning the interpretation of the spoken word into a machine instruction for use with computing. Common uses of ASR include automated telephone directories, cellular phone voice dialing, and in-car voice-activated systems such as Ford Motor Company's Sync. However, system performance degrades in noisy environments.

Summerfield has found that humans rely on visual cues in noisy environments to aid in speech comprehension [1]. Applying this concept to ASR, current research has been able to integrate both audio and video into an ASR system to provide robustness to noisy environments. This integration is known as audio-visual automatic speech recognition (AVASR). While AVASR systems outperform their audio-only counterpart, much of the research extracts the visual data (lip parameters) within a controlled environment of monotone backgrounds, one stationary face, and optimal lighting [2-7]. These tests do not represent a real-world environment in which the subject will frequently move in 3-dimensions within an environment with multiple subjects, cluttered backgrounds, and lighting changes. Thus the objective of this work is to accurately locate and track the face and lips within a real-world environment, providing a preprocessing component to currently developed lip parameter extraction and AVASR systems. This is further motivated in that facial tracking can be extended to many more applications, including human interface devices (HIDs), which enable computer control, to security in which a suspect could automatically be tracked via facial recognition.

Numerous facial tracking methods have been proposed. These include methods based on intensity gradients [9], graph matching (deformable templates) [10], and contour tracking (snakes) [11], all of which provide promising results but require numerous calculations or extensive statistical datasets. Elaborate systems implementing Kalman filtering and k-mean clustering [12] or adaptive hidden Markov model (HMM) classification [13], provide increased accuracy at the expense of still more additional computations. Background substitution methods (difference images) [14] offer a reduction in computation and complexity but require relatively static backgrounds.

The mean shift (MS) algorithm, a model-based tracking implementation, was first proposed in 1975 by Fukunaga and Hostetler [15] and has been used extensively in recent years for object tracking [16-18]. Its popularity stems from its simplicity and computational efficiency; requiring minimal storage and a non-exhaustive maximization method in which all calculations are region based. Furthermore, out-of-plane rotations are of minimal concern with the MS algorithm

Visual Communications and Image Processing 2009, edited by Majid Rabbani, Robert L. Stevenson, Proc. of SPIE-IS&T Electronic Imaging, SPIE Vol. 7257, 72571Y · © 2009 SPIE-IS&T · CCC code: 0277-786X/08/$18 · doi: 10.1117/12.817092

since the densities of a region are compared rather than its raw pixel data [19]. In this work, MS algorithm was adopted and extended to tracking face and lips in visually challenging environments.

The paper is organized as follows. Section 2.1 describes color space selection for face and lip detection. Section 2.2 examines the mean shift algorithm and details a specific implementation of the MS algorithm. To compensate for a tracked object's movement in 3-dimensions, a scaling algorithm is then developed in Section 2.3. Then Section 2.4 introduced a MS vector scaling to speed up the process. Second invocation of the mean shift algorithm in tracking the lips will be detailed in Section 2.5. Our tracking implementation introduces two benefits over that of previous implementations and these contributions are detailed in Section 2.6. Section 3 presents test results on over 300 videos. Finally, a brief summary is given in Section 4.

2. PROPOSED SYSTEM

2.1 Optimal Color Space for Face and Lip Detection

To determine the optimal color space for skin detection, a database of over 400 images, collected from [24] as well as the internet and personal photos, was collected. Manually drawn lip masks are then constructed over each subject's mouth and are used to develop the statistical models of *Lip*, *Non-Lip*, and *Skin* classes for the different color spaces. Histograms are generated for each class and color space and, when applicable, the Gaussian approximations are calculated. To aid in comparison between color spaces, the histograms for each color space component and the three classes are normalized, providing the expected frequency of occurrence. Several sample histograms for each of the five analyzed color spaces (RGB, nrgb, YCbCr, YIQ, HSV) are shown in Figure 1. It can be seen that the hue component of the HSI color space provided the maximum separation between regions and, therefore, is the strongest classifier. To simplify processing within the hue component a 0.2 shift was then applied, resulting in the modified color space, sHSI. More details on color space analysis can be found in our previous work [21].

2.2 Face Tracking with the Mean-Shift (MS) Algorithm

Since the MS algorithm provides a density estimation of the gradient, we first convert the face ROI to its corresponding density; this was accomplished using histograms, approximating the ROI's PDF. In addition, since sHSI is identified as the optimal color space for face and lip detection, hence in MS algorithm, face ROI is represented by 2D histogram of size 16x16 consisting of both the sH and S components.

To calculate the PDF approximations for the given model, each pixel within the given ROI is first converted to its equivalent bin representation. Assuming the pixel values are 8-bit integers, the conversion for each dimension of the ROI's color space (sH and S) to a bin representation, b_d, for the dimension, d, is given by

$$b_d(pv_{r,c,d}) = floor\left(\frac{pv_{r,c,d} \cdot Tb_d}{256}\right) + 1, \qquad d = 1,2 \tag{1}$$

where $pv_{r,c,d}$ is the 8-bit integer pixel value for row, r, column, c, and dimension, d, and Tb_d is the total number of bins for the dimension, d. Both Tb_1 and Tb_2 are chosen to be 16 for the proposed system. This process is repeated for all dimensions, rows, and columns within the ROI. Note the addition of 1 is due to Matlab starting matrix indexing at 1 and not 0. Then the pixel distance is normalized relative to the ROI's center to provide scale invariance. And the approximated PDF is given by:

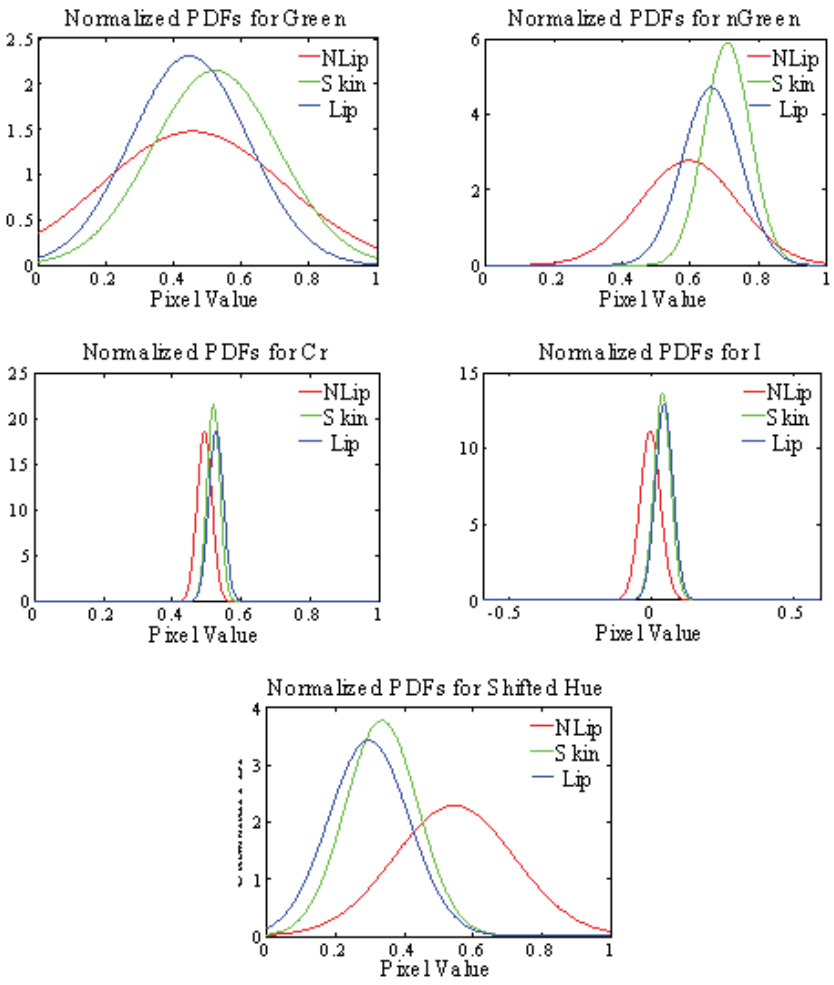


Fig. 1. Gaussian distribution for Lip, Skin, and Non-Lip regions for five color spaces – RGB, NRGB, YCbCr, YIQ, and sHSI.

$$\mathbf{H} \quad \mathbf{0},$$
$$\forall \mathbf{x} \in ROI \rightarrow H(b_1, b_2) = H(b_1, b_2) + \left(1 - \mathbf{x}_p^T \mathbf{x}_p\right) \tag{2}$$
$$\mathbf{H} \quad \frac{\mathbf{H}}{\sum_i \sum_j H(b_i, b_j)} =$$

where the b's are the bins for a given dimension (in this case the hue and saturation dimensions) and $\mathbf{x_p}$ is the normalized distance from the ROI center to the pixel at $[r,c]$, belonging to the bin (b_1,b_2). and **H** is the resulting 2-dimensional approximated PDF (normalized and weighted histogram).

Next, the Bhattacharyya coefficient, which relates the distributions of a candidate ROI to that of a model's, is given as (to simplify the mathematics we will restructure the 2-dimensional histogram as a 1-dimensional array):

$$\rho(\mathbf{y}) \quad \sum_{u} \sqrt{p_u(\mathbf{y}) q_u} \tag{3}$$

where,
$p_u(\mathbf{y})$ = density of candidate histogram bin,u, at frame location $\mathbf{y}$
q_u = density of model histogram bin,u

Through several steps of calculations [20] it can be shown that the probability of color u for the model and the candidate can be represented as

$$q_u \quad C \sum_{i=1}^{n} K_E\left(\left\|\frac{\mathbf{x}_i}{\mathbf{x}_{ctr}}\right\|^2\right) \delta[b_d(\mathbf{x}_i) - u] \tag{4}$$

$$p_u(\mathbf{y}) \quad C_p \sum_{i=1}^{n} K_E\left(\left\|\frac{\mathbf{y} - \mathbf{x}_i}{\mathbf{x}'_{ctr}}\right\|^2\right) \delta[b_d(\mathbf{x}_i) - u] \tag{5}$$

where C and C_p are the normalization constants and $\mathbf{x'_{ctr}}$ is the ROI center for the candidate distribution.
Now to maximize the Bhattacharyya coefficient, the following term needs to be maximized

$$\frac{1}{2} \sum_{u=1}^{m} p_u(\mathbf{y}) \sqrt{\frac{q_u}{p_u(\mathbf{y_0})}} \quad \frac{C_p}{2} \sum_{i=1}^{n} w_i K_E\left(\left\|\frac{\mathbf{y} - \mathbf{x}_i}{\mathbf{x}'_{ctr}}\right\|^2\right) \tag{6}$$

$$w_i \quad \sum_{u=1}^{m} \delta[b_d(\mathbf{x}_i) - u] \sqrt{\frac{q_u}{p_u(\mathbf{y_0})}}$$

where w_i is the weight function and is dependent on the model distribution and the current location of the candidate.

We can now introduce the mean shift vector given by:

$$\mathbf{y}_1 \quad m(\mathbf{y}_0) - \mathbf{y}_0$$

$$\mathbf{y}_1 \quad \frac{\sum_{i=1}^{n} \mathbf{x}_i w_i}{\sum_{i=1}^{n} w_i} \tag{7}$$

The recursive implementation of the mean shift vector of (7) is defined as the mean shift algorithm. Note that to determine the new location, only $p_u(\mathbf{y})$ must be updated between iterations since the model distribution, q_u, is not dependent on the candidate's location and, therefore, only needs to be calculated at initialization.

The mean shift tracking algorithm is based on the density gradient estimates, yet no gradients are ever calculated. Furthermore, while the math is extensive to show how it works, its implementation is simple and, hence, computationally inexpensive.

Utilizing the methods of detecting a face developed in [8], if a face is detected then the MS tracking algorithm, given below and based on [12], is initialized to track the face.

Initialize the system:

1. Determine a distance threshold, ε, to stop recursion.
2. Calculate the model's distribution, q_u, (4).
3. Initialize the candidate to the model's location.

Mean shift recursion

4. Calculate $p_u(\mathbf{y}_0)$, (5).
5. Calculate the weight function, w_i, (6).
6. Calculate the mean shift vector (7) to determine new location, $\mathbf{y}_1$.
7. If $\|\mathbf{y}_1 - \mathbf{y}_0\| < \varepsilon$ Stop. Otherwise $\mathbf{y}_0 \leftarrow \mathbf{y}_1$, and go to Step 4.

2.3 ROI Scaling

The MS algorithm provides tracking in the 2-dimensional space of an image. However, if a subject moves in a direction normal to the camera plane (closer or further to the camera) where the current candidate ROI will either no longer contain the full face or will contain additional background information, an *ROI scaling algorithm* was developed. It was found in previous research [21] that the expected value for the skin class in the shifted-hue (sH) color plane is lower than the background class. Therefore, if we determine the gradient of the candidate ROI for the sH color plane, then the gradient magnitudes are expected to be larger around the face perimeter. Additionally, assuming minimal scale change between frames of the candidate ROI, these increased values of gradient magnitude should lie near the perimeter of the elliptical ROI. To minimize the computational complexity, the gradient will be approximated using the pixel differences for both rows and columns.

To further reduce processing time, the scale algorithm is limited by two invocation restrictions. After every four frames the Bhattacharyya coefficient is calculated, if the coefficient is below 0.8, the scaling algorithm is called. Note: these values were determined using Monte Carlo simulation. Finally, to provide adaptive sizing to the ROI and increase the accuracy of the returned ROI size, the scale algorithm has been made recursive. A maximum of five iterations is allowed for the algorithm and, if any iteration results in an optimal ROI size corresponding to the unaltered current ROI size (middle size ring), the function terminates. The increased accuracy is the result of the current ROI being oversized by 10% of its diameter; therefore, each iteration will provide a percentage difference to that of the original.

An example of the scale algorithm is provided in Figure 2. Fig. 2(a) displays the frame from which the model ROI was selected while (b) provides an example of the increase in scale and (c) a decrease in scale. Note the offset in (b); this was caused by the subject's face just leaving the edge of frame, partially occluding the face.

(a) (b) (c)

Fig. 2. Example of the ROI scale algorithm.

2.4 MS Vector Scaling

As the distance between model and candidate ROI increases, so does the required number of MS iterations; this is the result of the candidate ROI size dictating the maximum value for the mean shift vector. Therefore, tracking frequency is inversely proportional to the MS iterations required. A method of *scaling the vector* was created to take advantage of this.

For sampling frequencies below 15 Hz (1 sample per 2 frames) it was found that the model and candidate ROIs would converge in fewer iterations if the result of the first iteration's mean shift vector was scaled. Specifically, comparing the MS vector for each iteration of a tracking implementation, one would find that the vector direction remained relatively constant. Additional testing provided a ratio relating the sampling frequency to that of the vector scalar. However, implementing a constant scalar provided near identical results while removing the calculation of the ratio from the system. The result is a modification of the mean shift tracking algorithm in which the last step is modified using the proposed vector scalar, *s*:

7. $\left\|\mathbf{y}_1 - \mathbf{y}_0\right\| \quad d$ If $d < \varepsilon$ Stop. Otherwise $\mathbf{y}_0 \leftarrow s \cdot \mathbf{y}_1$, and go to Step 4.

and,

$$s \quad \begin{cases} 3 & \text{if } I_M \quad 1, d >= 2 \cdot \varepsilon \\ \overline{\overline{1}} & \text{otherwise} \end{cases}$$

where I_M is the current mean shift iteration number.

The result is an MS tracking algorithm which will converge in fewer iterations than the MS target algorithm of for targets in motion. Finally, this addition has reduced the average iterations for our proposed system approximately 36% (from 1.88 to 1.21 iterations/frame).

2.5 Lip Tracking

Thus far only facial tracking has been presented. This is due to the similarity of the implementations; with the only major differences being that the lip tracking implementation does not implement an MS vector scalar, the MS iteration limit is increased to 10, and the histogram size is enlarged due to the small ROI size and mouths deformation.

Since the lip ROI is much smaller than that of the face ROI, an implementation of the MS vector scalar risks "overshooting" the lips and either increasing the MS tracking algorithm's required iterations for convergence or losing the target altogether (the system could determine the nose or chin is the mouth). Since the proposed system's objective is to provide a lip region in which to extract its parameters, these errors are not acceptable and, therefore, the MS vector scalar is excluded from the lips MS tracking algorithm.

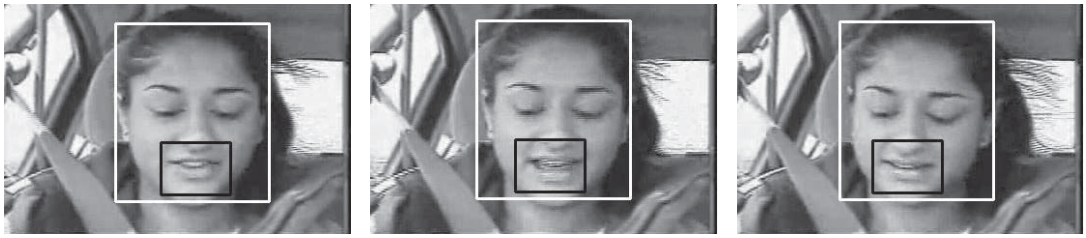
After experimentation, it was determined that the lip ROI histogram must be enlarged to compensate for the reduction in ROI detail. Specifically, since the perimeter of the lip ROI is skin, minimal discriminatory data is obtained which, in turn, requires an increase in the histogram size to provide additional detail within the mouth region and its surrounding skin. Furthermore, since the skin perimeter provides near radial symmetry, the mean shift vector's effectiveness is reduced; as a result, the lip ROI tends to lag behind the target. This is further complicated by the relative size of the ROI in comparison to the subject's potential movements, in which even typical movements of the head, including rotation, could cause a loss of the target. Therefore, after experimentation, the lip ROI histogram has been increased to 32x32.

To further reduce errors, the lip tracking distance threshold, ε_L, is set to zero. While this increases the iterations it is considered acceptable since the region is much smaller, and hence, computationally less expensive than the face tracking implementation.

Lip region scaling is determined by the scaling of the face. This is accomplished by determining the height and width ratios between the face and lip bounds prior to scaling invocation. If this ratio is altered during the scaling invocation the lip bounds are adjusted accordingly.

Finally, an example of the lip tracking implementation can be seen in Fig. 3. This displays the tracking results for a subject with minimal spatial velocity and near the camera (increased lip ROI size); note the successful tracking regardless of the mouth being open or closed. Compare this with the example of Fig. 3 (b), in which the reduced size of the lip candidate ROI and the subject's spatial velocity results in the algorithm selecting a local mode that does not correspond to the lip model ROI; however, as can be seen in Fig.3 (c), the lip ROI has nearly recovered the target once the subject's velocity decreased. In an attempt to reduce these errors, the lip ROI, prior to lip tracking, was shifted from its original position according to the new location of the face. However, if the subject's head rotated, the lips were more

readily lost. Therefore, until a camera with pan, tilt, and zoom (PTZ) capabilities can be implemented, the assumption is made that the user will be made aware of the system's limitations.



(a) (b) (c)

Fig. 3. Example of lip tracking success

2.6 Tracking Contributions

Since the mean shift tracking algorithm is based on PDFs, it is robust against not only occlusions and clutter, but variations in color as well. This is due not only to the bin sizes of the histograms but also to the fact that the algorithm maximizes the similarity measure (Bhattacharyya coefficient) about the mode with no restrictions of locating a "perfect" match.

This robustness to color variations is exploited in nearly all mean shift tracking systems to date [18 - 20, 22, 23]. Specifically, these systems utilize a generic face mask as the model for not only detecting a face but tracking it as well; all of which have shown some degree of success. However, the downside of using one generic face model within the tracker is the increase in the size of the histograms. This is due to the fact that the Bhattacharyya coefficient never approaches 1 (model and candidate are identical). Because the expected Bhattacharyya coefficient is low the threshold to initialize the system must also be decreased. Therefore, an increased resolution of the spatial gradient density estimation must be performed to minimize the false positives of objects similar in color that are not faces. Or, equivalently, the approximated PDF must increase in resolution. The only way to increase this estimation is to increase the number of histogram bins.

However, our implementation uses generic face models only to detect the face. As an additional contribution to refining the tracking algorithm, we are able to extract the face bounds of a subject using both hard (predetermined) and soft (adaptive) thresholds, concentration signals for their simplicity in both storage and computational requirements, and an optimal color space (sHSI). This enables us to create a run-time facial model in which we can expect to find a near "perfect" match. This drastically reduces the required histogram size, reducing the storage and computations required. It should be noted, however, that because of the reduced histogram size, we require a larger value of the Bhattacharyya coefficient for initial detection than that of other systems. This is why we search using three generic masks, as opposed to their one.

Comparing the storage requirements of our system to those previously mentioned, we have reduced the storage by no less than ½. While our system implements a 16x16 histogram (256 bins), the method presented in [22], which uses color as well as optical flow, and therefore an increase in complexity, implements an 8x8x8 histogram (512 bins); or twice the storage requirements. Furthermore, their computational costs are increased due to the optical flow calculations as well as the implementation of a 3-dimensional kernel. The work of Zhang et al. [23] reduced the computations by using only color (RGB) but this required increasing the histogram size to 11x11x11 (1331 bins), or approximately 5 times the storage requirements of the proposed system.

Additional contributions include the ROI scaling algorithm as well as the MS vector scalar. The ROI scaling algorithm, while still requiring additional work, provides an alternative to the computationally expensive approaches of [16, 19, 20, 23]. Finally, the MS vector scalar provided an approximate 36% reduction in computations by decreasing the required iterations to achieve ROI convergence.

3. RESULTS

All testing was performed using Matlab 2006a running on a desktop computer with 1GB of memory and a 2.93 GHz Intel Celeron CPU. The video dataset consists of 325 fifteen-second videos from [24], with 360x240 frame resolution, comprising 86 subjects, each of which is filmed from 4 angles while the vehicle is in motion. The motion of the vehicle provides an ever-changing background due to changes in the exterior scenery; furthermore, on several occasions the head of an individual riding in the backseat will enter the frame. Additionally, lighting conditions within the car continually change, providing frequent scenes in which direct sunlight will illuminate the subject's face for varying lengths of time.

Additional videos include three sequences from television to test the ability of the system to detect and track highly filtered, light-controlled subjects with frequent scene changes, as well as the camera (scene) changes and backlit sets. These videos have a frame resolution of 320x240, ranged in duration from 2 to 12 seconds, and contained four total subjects. Finally, nine videos were created to test facial occlusion, the scale algorithm, and system confusion. These nine videos provide three additional test subjects and two additional environments with frame resolutions of 320x240 and durations ranging from 12 to 63 seconds. In total, 337 videos totaling approximately 85 minutes, consisting of five environments and 93 subjects, were analyzed.

In analyzing performance of the system in tracking these videos, we note a considerable decrease in lip tracking success (44.3%) over that of facial tracking (77.8%). This is the result of the reduced size of the lip ROI, the dynamic range of the lips, and a relatively static color range. Specifically, the histogram (approximated PDF) of the lip ROI will contain minimal pixel data as the region is small. Furthermore, the histogram will contain little variation from that of any surrounding region of the lip ROI. Specifically, the lip ROI will contain approximately four classes of color: skin, the lips, the dark region of the mouth opening, and possibly teeth. Within each of these classes, one would expect minimal variation in pixel range for such a small ROI, resulting in a histogram comprised of four regions of predominant bin containment.

As a result of the low success rate, the lip localization algorithm may be abandoned. Furthermore, because of the numerous local modes (due to the small ROI), the lip tracking component of the proposed system will be replaced in future system development.

In contrast, facial tracking was much more successful (77.8%) due to the number of facial features as well as the range of colors within the ROI. Furthermore, the face typically differs from that of the background in color and features and, as a result, there is a reduction in the number of local modes. This provides robustness to even large spatial displacements of the target. Also, because of the increased size of the facial ROI, not only is additional information obtained, but outlier effects are minimized due to sheer volume of pixels.

Three examples corresponding to system success can be seen in Fig. 4. Note the method implemented to overlay the ROIs can result in the lip ROI being copied into the face ROI; this can be seen in (c) and in no way affected the test results. The figure displays the first detected frame on the left and the last frame on the right. Of the three videos, both (a) and (b) lost no frames in detecting and localizing, whereas (c) required 5 frames to locate the face and lips. Of interest is the successful tracking of the lips in (a), where hair has partially occluded the lips for the previous 6 seconds. Additionally, it should be noted that the lips were lost for 3 seconds within this video due to the subject's hand covering her mouth to remove her hair from her face.

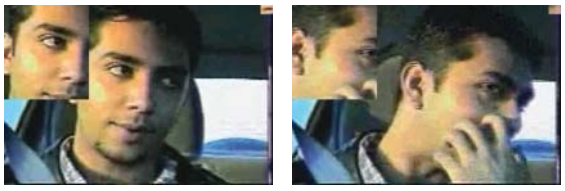
Three examples of system failure are found in Fig. 5. The video in (a) required 409 frames (leaving just 36 frames of video) for face and lip localization, due to the contrast in lighting between the subject's left and right face. Note that the face ROI comprises only the high intensity region of the face and the detected "lips" is actually the right nostril. For the videos (b) and (c) we see that the face and lips were initially successfully located. For the subject in (c), his exaggerated facial and body movements resulted in the lip and face ROI frequently lagging his current location. Subject (b) fully covered his mouth and dusted his mustache several times, resulting in lip ROI losing track.

(a)

(b)

(c)

Fig. 4. Examples of system success



(a)

(b)

(c)

Fig. 5. Examples of system failure

4. CONCLUSION

This paper presented a method of tracking both the face and lips using color-based features. With bounds provided by the localization algorithm, custom model was generated to enable tracking by the mean shift algorithm. This model resulted in a dramatic decrease in storage and processing requirement over similar system. Additional processing reductions were realized utilizing the proposed scaling algorithm allowing the model to adjust to a subject moving in 3d space. Finally, a method of scaling the MS vector was proposed that reduced the required MS algorithm iterations by approximately 36%.

Preliminary testing suggests an increase in image resolution would greatly increase the system performance by providing model and candidate ROIs that are much more informative. The addition of a PTZ camera would allow the scale algorithm to control the zoom, thereby increasing the resolution of the candidate ROIs.

REFERENCES

[1] Summerfield, Q., "Lipreading and Audio-Visual Speech Perception," Philosophical Transactions of the Royal Society of London: Biological Sciences, Vol. 335, Page(s): 71-78 (1992).

[2] Beaumesnil, B., Luthon, F., "Real Time Tracking for 3D Realistic Lip Animation," 18th International Conference on Pattern Recognition, Vol. 1, Page(s): 219-222 (2006).

[3] Da Silveira, L. G., Facon, J., Borges, D.L., "Visual Speech Recognition: A Solution From Feature Extraction To Words Classification," Brazilian Symposium on Computer Graphics and Image Processing, Page(s): 399-405 (2003)

[4] Delmas, P., Lievin, M., "From Face Features Analysis to Automatic Lip Reading," 7th International Conference on Control, Automation, Robotics and Vision, Vol. 3, Page(s): 1421-25 (2002)

[5] Gomez, E., Travieso, C.M., Briceno, J.C., Ferrer, M.A., "Biometric Identification System by Lip Shape," Proc. 36th Annual 2002 International Carnahan Conference on Security Technology, Page(s): 39-42 (2002)

[6] Kumar, K.; Tsuhan C., Stern, R.M., "Profile View Lip Reading," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Vol. 4, Page(s): 429-432 (2007)

[7] Kumatani, K., Stiefelhagen, R., "State Synchronous Modeling on Phone Boundary for Audio Visual Speech Recognition and Application to Multi-View Face Images," IEEE International Conference on Acoustics, Speech and Signal Processing, Vol. 4, Page(s): 417-420 (2007)

[8] Crow, B., Zhang, X., "Face and Lip Localization in Unconstrained Imagery," Proc. 10th IASTED International Conference on Signal and Image Processing, (2008)

[9] Birchfield, S., " Elliptical Head Tracking Using Intensity Gradients and Color Histograms," IEEE Conference on Computer Vision and Pattern Recognition, Santa Barbara, CA (1998)

[10] Yao, Z., Li, H., "Tracking a Detected Face with Dynamic Programming," Computer Vision and Pattern Recognition Workshop, Page(s): 63 (2004)

[11] Jiang, H., Drew, M., "A Predictive Contour Inertia Snake Model for General Video Tracking," International Conference on Image Processing, Vol. 3, Page(s): 413-416 (2002)

[12] An, K.H., Yoo, D.H., Jung, S.U., Chung, M.J., "Robust Multi-View Face Tracking," IEEE/RSJ International Conference on Intelligent Robots and Systems, Page(s): 1905-10 (2005)

[13] Oliver, N., Pentland, A., Berard, F., "LAFTER: A Real-Time Face And Lips Tracker With Facial Expression Recognition," IEEE Conference on Computer Vision and Pattern Recognition, Page(s): 123-129 (1997)

[14] Tang, Z., Miao, Z., "Fast Background Subtraction and Shadow Elimination Using Improved Gaussian Mixture Model," IEEE International Workshop on Haptic, Audio, and Visual Environments and Games (HAVE 2007), Page(s): 38-41 (2007)

[15] Fukunaga, K., Hostetler, L., "The Estimation of the Gradient of a Density Function, with Applications in Pattern Recognition," IEEE Transactions on Information Theory, Vol. 21, Page(s): 32-40 (1975)

[16] Bradski, G., "Computer Vision Face Tracking for Use in a Perceptual User Interface," Intel Technology Journal, Q2 (last accessed Aug. 17, 2007),

[17] Liang, D., Huang, Q., Jiang, S., Yao, H., Gao, W., "Mean-Shift Blob Tracking with Adaptive Feature Selection and Scale Adaptation," IEEE International Conference on Image Processing, Vol. 3, Page(s): 369-372 (2007)

[18] Yang, C., Duraiswami, R., Davis, L., "Efficient Mean-Shift Tracking via a New Similarity Measure," Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 1, Page(s): 176-183 (2005)
[19] Comaniciu, D. & Ramesh, V., "Robust Detection and Tracking of Human Faces with an Active Camera," Proc. Third IEEE International Workshop on Visual Surveillance, Page(s): 11-18 (2000)
[20] Comaniciu, D., Ramesh, V., Meer, P., "Kernel-Based Object Tracking," In: IEEE Transactions on Pattern Analysis and Machine Intelligence," Vol. 25, Page(s): 564-577 (2003)
[21] Zhang, X., Montoya, H.A., Crow, B., "Finding Lips in Unconstrained Imagery for Improved Automatic Speech Recognition," Proc. 9th International Conference on Visual Information Systems, Page(s): 185-192 (2007)
[22] Oshima, N., Saitoh, T., Konishi, R., "Real Time Mean Shift Tracking using Optical Flow Distribution," SICE-ICASE International Joint Conference, (2006)
[23] Zhang, X., Qiao, H., Liu, Z., "Multi-Information Fusion for Scale Selection in Robot Tracking," Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems, Page(s): 2828-32 (2006)
[24] Lee, B., Hasegawa-Johnson, M., Goudeseune, C., Kamdar, S., Borys, S., Liu, M., Huang, T., "AVICAR: Audio-Visual Speech Corpus in a Car Environment," INTERSPEECH2004-ICSLP, (2004)