Detection of Mouth Movements and Its Applications to Cross-Modal Analysis of Planning Meetings

Yingen Xiong  
Nokia Research Center  
Palo Alto, CA 94304, USA

Bing Fang and Francis Quek  
Virginia Polytechnic Institute and State University  
Blacksburg, VA 24061, USA

Abstract—Detection of meaningful meeting events is very important for cross-modal analysis of planning meetings. Many important events are related to speaker’s communication behavior. In visual-audio based speaker detection, mouth positions and movements are needed as visual information. We present our techniques to detect mouth positions and movements of a talking person in meetings. First, we build a skin color model with the Gaussian distribution. After training with skin color samples, we obtain parameters for the model. A skin color filter is created corresponding to the model with a threshold. We detect face regions for all participants in the meeting. Second, We create a mouth template and perform image matching to find candidates of the mouth in each face region. Next, according to the fact that the skin color in lip areas is different from other areas in the face region, by comparing dissimilarities of skin color between candidates and the original color model, we decide the mouth area from the candidates. Finally, we detect mouth movements by computing normalized cross-correlation coefficients of mouth area between two successive frames. A real-time system has been implemented to track speaker’s mouth positions and detection mouth movements. Applications also include video conferencing and improving human computer interaction (HCI). Examples in meeting environments and others are provided.

Keywords—multimodal meeting analysis; mouth movement; meeting event detection; speaker detection; skin color model; experiment for planning meeting; human computer interaction;

I. INTRODUCTION

Meetings are gatherings of humans for the purpose of communication. Such communication may have various purpose: planning, conflict resolution, negotiation, collaboration, confrontation, etc. One of the most important objectives of multimodal analysis of planning meetings is to understand human multimodal communicative behavior, and how witting or unwitting visual displays relate to such communication. To reach this objective must necessarily combine the psycholinguistics of multimodal human language, signal and language processing, and computer vision.

In video based multimodal analysis of planning meetings, meeting events are recorded by cameras and microphones [1]. Extraction of meaningful meeting events from video and audio signals is a key step in meeting analysis. Especially, we are interested in the events related to the speakers. Human communication is inherently multimodal. Gesture, speech, and gaze function as an integrated whole, no part of which may be fully understood in isolation [2]. The meaningful meeting events, of course, relate to speakers’ gesture, speech, gaze, and gaze target. In order to extract these meaningful events, we need to locate speakers both spatially and temporally. Mouth movements can provide not only information of speaking behavior (speaking or not speaking) but also speaker’s location signals. This paper addresses the aspect of extraction of mouth positions and movements of a talking person in planning meetings.

A. Detection of human mouths from Video

In related work on detection of facial features, there are template-based approaches [3]; wavelet-based approaches [4]. In template-based approaches, image matching is performed between a mouth template and face images to locate mouths. The main difference in this kind of approaches is how to create templates. There are mouth templates taken from real people [3], synthetic templates [5], and deformable templates [6]. Wavelet-based approaches apply different levels to find face features. In [4], the first level is used for face matching and the second level for facial feature localization. Wavelet-based approaches also need to apply templates to locate facial features.

In multimodal analysis of planning meetings, we need to detect speakers both spatially and temporally, so that the speakers’ communication behavior can be analyzed at the same time. Mouth movements are important signals for speaker localization. The fact that the visual motion of a speaker’s mouth is highly correlated with the audio data generated from the voicebox and mouth [7] has been exploited for lip/speechreading [8] and for combined audio-visual speech recognition [9]. The challenge of mouth detection for speaker localization in meetings is that the mouth shape is changing while people are speaking. We need to locate mouth positions and measure movements.

In order to detect and recognize speakers efficiently in meetings, in this paper, we present an efficient parallel approach to extract mouth positions and movements of multiple speakers in meetings. We build a skin color Gaussian model for each participant. Different skin color models are applied to deal with different skin colors. We create skin color filters related these color models with different
thresholds. The face regions for all participants can be extracted by applying the skin color filters. We create a mouth template and perform image matching to find candidates of mouth area in the face region for each participant. The final decision for the real mouth area from the candidates is made by computing dissimilarities between color models of candidates and original skin color model. The candidate which corresponds to the maximum dissimilarity is the real mouth region. We compute normalized cross-correlation coefficients of mouth area between two successive frames to express mouth movements. We demonstrate the efficiency of the approach in experiments and applications of multimodal meeting analysis.

II. SUMMARY OF OUR APPROACH

The approach includes three main parts. The first part is to create skin color models with skin color samples. We build a skin color model with the Gaussian distribution. We train this model with skin color samples collected from image sequences of video and obtain parameters of the model. The second part is to build a skin color filter and segment face regions. We create a skin color filter corresponding to the color model with a threshold. Face regions are extracted from image sequences by filtering pixels into skin color and non-skin color. Face image sequences are obtained by processing the incoming image stream. In order to handle differences of skin color from different people, we build an individual skin color model and filter for each person in the image frame. The third part is detection of mouth positions and movements. We create mouth template and perform image matching in face regions to find mouth candidates. In order to decide which candidate is a real mouth area, we compute dissimilarities of color between candidates and the original skin color model. The mouth area is the one corresponding to maximum dissimilarity of skin color. In the mean time, we obtain mouth positions in the whole image frame. The mouth movements are measured by normalized cross-correlation coefficients computed from the mouth area between two successive frames.

III. EXTRACTION OF FACE REGIONS WITH SKIN COLOR MODEL

A. Skin Color Gaussian Model

The skin color model theory is established by Yang and other researchers [10]. The skin color model theory is based on the facts: human skin color are clustered in the color space; under certain lighting conditions, a skin color distribution can be characterized by a multivariate Gaussian distribution in the normalized color space. Therefore, we can model human face with different color appearances in the normalized color space. By computing the probability of a pixel in skin color Gaussian distribution we can segment the face region.

We build the face skin color model in RGB space. Usually the RGB space in original color images includes luminance component, which makes it difficult to characterize skin color because lighting effects change the appearance of the skin. In order to reduce lighting effects, we convert original color images to chromatic color images. Suppose \(x(R, G, B)\) and \(x'(R_n, G_n, B_n)\) are pixels in the original color image and chromatic color image respectively.

\[
R_n = \frac{R}{R + G + B}, B_n = \frac{B}{R + G + B}, G_n = \frac{G}{R + G + B} \tag{1}
\]

In above, as \(R_n + B_n + G_n = 1\), there are only two independent components, so we omit the third component. For each pixel, we have a color vector \(x = (R_n, B_n)^T\). The two dimensional Gaussian distribution model is expressed as \(N(\mu, \Sigma)\) i.e.

\[
p(x) = \frac{1}{2\pi\Sigma^{1/2}} \exp\left[-\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right] \tag{2}
\]

\[
\mu = E\{x\}, \Sigma = E\{(x - \mu)(x - \mu)^T\} \tag{3}
\]

where, \(\mu\) is the mean vector and \(\Sigma\) is the covariance matrix.

Before we can use this model, we need to create samples to estimate the parameters \(\mu\) and \(\Sigma\). One of the efficient parameter estimation approaches is Maximum Likelihood Estimation (MLE).

\[
\hat{\mu} = \frac{1}{n} \sum_{k=1}^{n} x_k, \hat{\Sigma} = \frac{1}{n} \sum_{k=1}^{n} (x_k - \hat{\mu})(x_k - \hat{\mu})^T \tag{4}
\]

In experiments, we found that people with different skin colors have different parameters of skin color models. The difference for mean vectors is not so much, but for covariance matrices is big. In our meeting analysis, we build a skin color model for each participant.

In actual tracking implementation, we sample several image frames from an individual video and generate a subject-oriented skin color model. If necessary, we might even update the skin color frame by frame for a tracking video sequence. We can also build adaptive skin color models for face tracking.

B. Skin Color Filter

Once the skin color model is built, we can create a skin color filter to segment skin color regions and non-skin color regions, so that we can obtain face regions from video. In order to build skin color filter, we apply the skin color model to compute the probability of each pixel in an image being skin color. The probability of a pixel with color \(x\) is

\[
P(x) = \exp\left\{-\frac{1}{2} (x - \hat{\mu})^T \hat{\Sigma}^{-1} (x - \hat{\mu}) \right\} \tag{5}
\]

Given a probability threshold, we can create a skin color filter. As skin colors do vary between each individual subject, we shall find best threshold values for different subjects under different applications (background, illumination, etc.).
We can use samples to train the system to learn thresholds for different participants in meetings.

In implementation, we scale the skin color probability of every pixel in an image from \([0, 1]\) to \([0, 255]\), thus we can create a probability gray scale image. Furthermore, by thresholding, we create a black/white binary image to represent non-skin color regions and skin color regions. We can use samples to train the system to learn thresholds for different participants in meetings.

IV. DETECTION OF MOUTH MOVEMENTS

To detect mouth movements, we need first to extract mouth positions which are required for defining a mouth area with a window. In visual-audio speaker localization, we use mouth positions to express speaker’s spatial positions. We measure changes inside the mouth area between two successive image frames to detect movements of the mouth. In order to find a mouth on face images, we create a mouth template and perform image matching with the face images in video. The work in this section includes creating a mouth template, implementing image matching with the template, making decision for the real mouth area, and measuring mouth movements.

A. Mouth Template and Mouth Detection

We define an initial mouth template with a window and perform image matching in skin color regions to find mouth candidates. We move the mouth template in skin color regions and compute correlation coefficients which are applied as the similarity measure. The candidates are the areas corresponding to the maximum similarity.

After obtaining candidates, we need to decide the real mouth area. Since the skin color in the mouth area (lip area) is different from that of other areas, we apply skin color differences to decide which candidate to be a real mouth area. We create a skin color Gaussian model for each candidate and compute dissimilarities between candidates’ skin color models and the original skin color model. For the candidate \(i\), the dissimilarity with the original skin color model can be expressed as:

\[
 d_i(p(x) : \mu, \Sigma; p_i(x) : \mu_i, \Sigma_i) \tag{6}
\]

The candidate corresponding to the maximum dissimilarity is the mouth area.

\[
 \max_i [d_i(p(x) : \mu, \Sigma; p_i(x) : \mu_i, \Sigma_i)] \tag{7}
\]

After obtaining the real mouth area for the current frame, the next step is the mouth template update. We know the fact that his mouth shape is changing while a person is talking. If we still use the original template, we will get error, even loss tracking. We need to update the mouth template to fit the change. We assume that the change of the mouth shape is small between two successive frames. We define a new mouth template with the mouth position and area in the current image frame. With the updated template, we perform mouth detection for a new image frame.

B. Mouth Movements

For visual information of speaker localization, we do not need to know the quantity of motion for a mouth. We only need to know whether it has motion or not. If the mouth shape does not have any change, we know that this person definitely is not talking. On the other hand, if the mouth shape is changing, we can say that this person may be talking at this moment. Instead of tracking mouth movements, we apply some techniques to detect changes of the mouth shape. Here, we use normalized cross-correlation coefficients to measure the change of the mouth area between two successive frames. Suppose \(F_t\) and \(F_{t+1}\) are two frame images at time \(t\) and \(t+1\). The normalized cross-correlation coefficient \(C_t\) can be computed as below.

\[
 C_t = \frac{\sum_{(x,y) \in W} (F_t(x,y) - \overline{F_t})(F_{t+1}(x,y) - \overline{F_{t+1}})}{(\sum_{(x,y) \in W} (F_t(x,y) - \overline{F_t})^2)\sum_{(x,y) \in W} (F_{t+1}(x,y) - \overline{F_{t+1}})^2} \tag{8}
\]

where \(W\) is a windowing function in \(F_t\) (typically a rectangle). In order to account for the detection error of mouth positions, the minimal \(C_t\) is found by translating \(F_{t+1}\) over a small search radius \(r\) (yielding \(C'_t\)).

Figure 1. Meeting Room Configuration and Video

V. EXPERIMENTS AND APPLICATIONS

A. Experimental Setup and Meeting Room Configuration

Figure 1 (a) shows the original experimental setup and configuration of the meeting room. In our planning meeting experiments, there are eight participants labeled A B C D E F G H in the meeting. Ten movie cameras labeled C1 C2 C3 C4 C5 C6 C7 C8 C9 C10 are installed to record the meeting events. T1 and T2 are two table microphones. Each camera is installed in a fixed position on the ceiling of the meeting room, so that each camera can see certain participants at the same time. We can set camera pairs to capture 3D data. In this meeting room, we also installed a Vicon motion capture system (which did not display on this figure) to provide...
ground truth data, so that we can evaluate our results. Eight Vicon infra red cameras are installed in the fixed positions on the ceiling of the meeting room, so that they can track Vicon markers installed on targets and provide us motion data and positions of targets. With this configuration, we capture whole data for cross-modal analysis of planning meetings. For this experiment, we only have five participants C, D, E, F, G in the meeting. The room configuration is same.

For the detection of participants’ mouths, we only need three videos taken by cameras C1, C5, and C7 respectively shown in Figure 1 (b), (c), and (d). From camera C1, C5 and C7, we can see the front faces of participants E, F and G, C and D respectively.

B. Experimental Data Processing Workflow

We create a parallel procedure to implement the detection process. In this procedure, we create a skin color model and filter for each participant, so that we can handle the differences of skin color from different participants. We use skin color samples extracted from each participant to train his/her skin color model off line. After the system learns all parameters from samples, we create skin color filters and perform face segmentation, mouth detection, and mouth movement computation for all participants on line.

C. Results and Analysis

We apply our approach to the dataset AFIT-01-07-05 captured a meeting scenario which five people were discussing a topic of Foreign Material Exploitation (FME) in our meeting room described in Section V-A. The dataset had comprised 74,805 frames (41.6 minutes). It is captured in synchronized stereo with 10 calibrated cameras. For the detection of mouth movements, we only need three videos taken by cameras C1, C5, and C7 for the five participants C, D, E, F, and G. Figure 1 (b), (c), and (d) show pictures from these three videos respectively.

The dataset captured five people in a meeting to discuss a topic of FME. Figure 3 (on the top) shows the audio signal for the speech. We use three videos taken from C1, C5, and C7 to detect the five subjects’ mouth movements. Since we need the subjects’ front faces for the detection, we set C1 for subject E, C5 for subjects F and G, and C7 for subjects C and D.

First, we build a skin color model for each participant. We collect skin color samples from their face areas and apply these samples to train the models to obtain parameters. For subject E, the model parameters are

$$\hat{\mu}_E = \begin{pmatrix} 95.00 \\ 75.27 \end{pmatrix}, \quad \hat{\Sigma}_E = \begin{pmatrix} 33.98 & -10.06 \\ -10.06 & 96.67 \end{pmatrix}.$$  

Based on this skin color model, we create a skin color filter for subject E. The threshold of the skin color filter is 120. In the same way, we build skin color models and filters to other subjects, for subject C:

$$\hat{\mu}_C = \begin{pmatrix} 100.90 \\ 71.56 \end{pmatrix}, \quad \hat{\Sigma}_C = \begin{pmatrix} -24.88 & 137.44 \\ -24.88 & -10.06 \end{pmatrix},$$

and the threshold 120, for subject D:

$$\hat{\mu}_D = \begin{pmatrix} 96.12 \\ 76.35 \end{pmatrix}, \quad \hat{\Sigma}_D = \begin{pmatrix} 24.25 & -8.42 \\ -8.42 & 43.38 \end{pmatrix},$$

and threshold 100, for subject F:

$$\hat{\mu}_F = \begin{pmatrix} 97.58 \\ 69.70 \end{pmatrix}, \quad \hat{\Sigma}_F = \begin{pmatrix} 14.45 & 2.02 \\ 2.02 & 376.48 \end{pmatrix},$$

and threshold 130, for subject G:

$$\hat{\mu}_G = \begin{pmatrix} 92.32 \\ 73.67 \end{pmatrix}, \quad \hat{\Sigma}_G = \begin{pmatrix} 13.53 & 4.08 \\ 4.08 & 238.26 \end{pmatrix}.$$  

and threshold 150.

With these skin color models and filters, we segment face regions for all subjects. We create a mouth template for each subject and extract mouth positions from each frame of the video. With the positions, we can define mouth areas for each subject. At last, we compute normalized cross-correlation coefficients to measure mouth movements. We have processed the whole dataset and obtained mouth movements for all five subjects. We also applied the results to the project of speaker localization and recognition. Since the dataset is too long (74,805 frames), we do not show whole results here. As an example, we choose a segment from the video and display its results.

![Figure 2. Mouth Positions for All Five Subjects](image-url)
With the positions of mouths, we apply a window to define mouth area and measure the movements of the mouths with normalized cross-correlation coefficients between two successive frames. Figure 3 shows the results which we extracted from videos. There are two columns in this figure. From top to bottom of the first column, they are speech signal and normalized cross-correlation coefficients for subjects C and D. In the second column, they are normalized cross-correlation coefficients for subjects E, F, and G.

From the results, we can see some facts. Subject E has a lot of mouth movements in the whole segment. This means that he speaks a lot. This is true because he is a leader of this meeting. The second person is subject G. He is explaining an operation process in this segment. Before 2100 frames, he speaks a lot. For subject C, before 2300 frames, she is silence. After that, she explains one possible operation. Subject D has not spoken for the whole segment. He just listens to others’ talking and gives eye gazes to other subjects since he has head movements which we can see in Figure 2. Subject F has mouth movements before 2000 frames because he is speaking with subject G.

We obtained the results for the whole dataset (AFIT-01-07-05). The results are checked using the Final Cut Pro system. We found that the results are satisfying except some cases which the mouths are occluded by subjects’ hands and the cases which subjects turn around their heads and the mouths can not be seen. We also applied the results to the project of speaker localization as visual signals and we obtained good results.

VI. CONCLUSION

In this paper, we presented our techniques to extract mouth positions and movements of a talking person in planning meetings. Our approach can locate speakers’ mouths both spatially and temporally. We created a parallel procedure for the approach to extract all participants’ mouth positions and movements which can provide not only speaking information but also position signals for speaker localization. In this parallel procedure, we built a skin color model for each participant in meetings to handle different skin colors and variations of background and illumination. We also created a skin color filter for each participant to segment his/her face region. Image matching is performed in the face region with a mouth template to find mouths for all participants. Mouth movements are characterized by normalized cross-correlation coefficients of mouth area between two successive frames. We demonstrated applications in multimodal analysis of planning meetings. We applied our approach to dataset captured meeting scenarios. In this paper, we presented an exemplar result. The approach detected all correct results except for the cases which the mouths are occluded by subjects’ hands and the heads turn away from cameras so that no mouth can be seen. The results have been also used in the project of visual-audio speaker localization.

REFERENCES


