# CAMERA AND MICROPHONE ARRAY FOR 3D AUDIOVISUAL FACE DATA COLLECTION

*Yuxiao Hu, Hao Tang, and Thomas S. Huang*

University of Illinois at Urbana-Champaign

## ABSTRACT

This paper proposes a novel camera/microphone array system capable of capturing dynamic facial expression video with synchronized speech and reconstructing realistic 3D face models from the data. Both hardware and software issues including camera calibration, video/audio synchronization, facial marker tracking and 3D shape reconstruction are considered. To our best knowledge, this system is the first camera/microphone array system that is able to capture high-resolution facial expression video with synchronized speech. The system can be used to collect dynamic 3D audiovisual face data for many multimedia applications.

***Index Terms—*** Camera/microphone array, 3D face model, 3D shape reconstruction

## 1. INTRODUCTION

Recently, more and more researchers are working on multimodal/multi-sample face recognition as opposed to traditional static image-based face recognition, which demands for databases of human faces in video sequences with speech [1]. On the other hand, multi-view/non-frontal view face analysis tasks also require 3D face data so the researchers can get arbitrary views of a human face during the process. According to these common and important research interests, building a synchronized camera/microphone array to collect multimodal 3D face video is necessary. There are already some face databases which provide 3D shape and texture information about human faces, including BU-3DFE [2], 3D-RMA [3], GavabDB [4], YorkDB [5], XM2VTS [6], FRGC [7] and [8], etc. These databases are either collected by Cyberware 3D scanners or some techniques based on structured lights. Temporal information about facial movements are absent so they can not be used in researches which require dynamic facial information such as affect recognition and non-frontal view lip reading. The most related work is done by Intel Research China, in which they built a system including 7 cameras and 12 microphones [9]. The drawback of this database is that the video signals of the cameras are not synchronized so that typical stereo techniques cannot be applied to recover the 3D information.

In this paper, we propose a camera/microphone array system capable of capturing synchronized multimodal face video with speech and reconstructing 3D face models according to facial feature points. In this system, multi-channel video/audios are first captured by a camera/microphone array. Then, 2D facial markers are tracked for each video channel. Finally, based on the corresponded 2D facial feature points in different views and stereo geometry, the 3D face shapes are reconstructed for each frame and aligned with the audio signals. Our experiments justify that our system achieves synchronized multi-view face video/audio captured data and the camera calibration, facial marker tracking algorithms are accurate enough to recover 3D face shape efficiently. We systematically investigate both hardware and software issues during system construction, including physical setup, camera calibration and synchronization, data flow control, facial marker tracking, and 3D shape reconstruction. Comparing to existing face data collection systems, our system has the following highlights,

1) Robustness: Our system is able to capture synchronized multi-view face video in real-time speed (30 fps) and speech audio (44.1KHz sampling rate). There is no out-of-sync and dropped frames;

2) Flexibility: Based on carefully designed hardware structure and software implementation, the system can be flexibly extended to any number of cameras and microphones without burdensome adjustments;

3) Complete Solution: Besides the data capture system, we also provide necessary tools for camera calibration, color space conversion, 2D facial feature tracking and 3D face shape recovery.

## 2. SYSTEM FRAMEWORK

### 2.1. Physical Setup

The physical setup of the system is illustrated in Figure 1. The subject is recorded by an array of 5 cameras and 2 microphones regularly located on a 1m radius half-circle. Figure 1 illustrates how the hardware pieces are connected with one other. In our system, we connect 2 Cameras to one IEEE 1394 hub and 3 cameras to another hub. All these cameras are synchronized by MultiSync software. Two microphones are also connected to the PC through an external sound card.

This work was funded in part by the U.S. Government VACE program and in part by the National Science Foundation Grant CCF 04-26627.

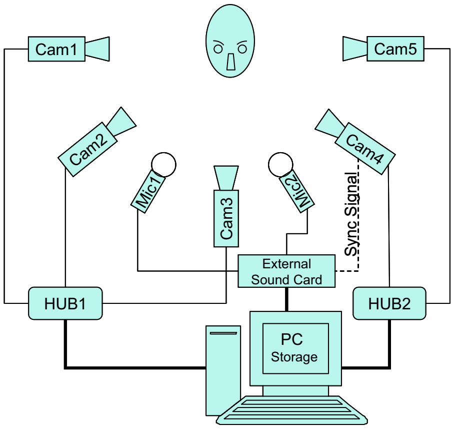

**Fig. 1**. *Physical setup and hardware configuration.*

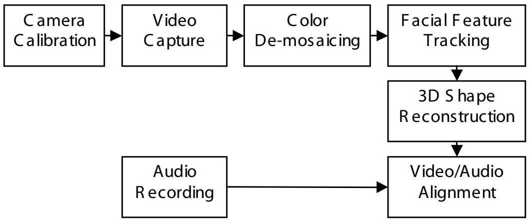

**Fig. 2**. *Data Collection Schema.*

All the sound channels are automatically synchronized by the sound card. There is one synchronization cable connecting one of the cameras to the sound card. Once video capturing is started, trigger signals from the synchronization cable will be recorded in one of the sound channels so that the video and audio signals can be aligned later.

### 2.2. Data Collection Schema

Based on the above hardware configuration and physical setup, the process of data collection and post process are presented in Figure 2. After the cameras/microphones are calibrated, the raw multi-view dynamic face video are captured and saved under the control of our capture program. Here the video is in raw format because the color of the video pixels is encoded with Bayer pattern to reduce the CPU load and transfer bandwidth. In order to track the facial feature points based on their colors, we need to de-mosaic the raw avi video to rgb color format video. If the facial feature points in different views are all detected/tracked successfully, based on the calibrated cameras and multi-view geometry, the 3D shape of the face is reconstructed and aligned with the audio signals frame by frame in offline.

## 3. SYSTEM HARDWARE CONFIGURATION

In this section, we detailed the hardware configuration of the proposed data collection system and the reasoning.

### 3.1. Cameras

For video collection, 5 OEM-style IEEE-1394 board level Dragonfly2 cameras from Point Grey Research Inc. are utilized. The camera has a maximum resolution of $640 \times 480$ pixels, and a maximum frame rate of 60 Hz. In our case, a single camera working in 30 Hz frame rate will occupy $640 \times 480 \times 8 \times 30 = 80Mbps$ data bandwidth.

### 3.2. Storage

One IEEE-1394 bus has the maximal 400 Mbps transfer speed, considering the overhead of synchronization and other costs, typically only less than 80% of the peak transfer speed can be achieved, so that each bus is able to process the data stream of three cameras. In order to transfer all the five video channels' data to the PC, we also need the IEEE 1394 PCI interface adapter which provides up to 133 MBps (about 1 Gbps) data bandwidth. Typically only 50% of this PCI bus bandwidth is effective for peripheral devices, which is still enough in our case. Based on our experiments, the hard drive access speed is the bottleneck of the capturing system. There are two different solutions. One is to use more memory as cache to compensate for the latency of the hard drive, the other is to use a faster disk configuration such as RAID0 to double the hard disks speed. In our current implementation, 4GB memory buffer are used.

### 3.3. Audio Interface and Microphones

A MOTU 8pre firewire audio interface external sound card is used to record multi-track synchronized speech signals. The interface provides 8 mic inputs with preamps. We use input port 1 for video-audio synchronization. That is, the camera shutter signal is converted into a pulse, sent to input port 1, and recorded into a separate sound track. The rest of the input ports are used for microphones. We currently use two condenser microphones, namely, Project Studio B1 microphones, for high-quality speech recording at 44.1 kHz sampling rate (CD qualilty).

## 4. SYSTEM SOFTWARE DEVELOPMENT

Several software tools are also developed to support the hardware and post-process the data.

### 4.1. Camera Calibration

Camera calibration determines the internal geometric and optical characteristics (intrinsic parameters) as well as the 3D position and orientation of a camera with respect to a certain world coordinate system (extrinsic parameters). We use the Camera Calibration Toolbox for Matlab to calibrate our cameras [10]. With this toolbox, the first step is to find a closed-form solution to the calibration problem using the direct linear transformation (DLT) method. The second step is to perform nonlinear parameter estimation using the DLT solution obtained by the first step as initialization.

### 4.2. Facial Marker Tracking

One fundamentally important tool is to locate facial feature points. This is done with the help of artificial markers physically attached to the face of the actor. We first track the 2D positions of the markers using a neighbor search algorithm. Due to occlusions and/or lighting changes, we may occasionally fail to track one or more markers by this low-cost algorithm. We then use marker collocation statistics as a means to estimate the positions of the markers that are lost during normal tracking. The idea follows from [11]. The positions and movements of the facial markers are highly correlated. We collect the marker collocation statistics during the period when all the facial markers are accurately tracked and use this marker collocation statistics to help recover the positions of those markers that are lost during normal tracking.

### 4.3. 3D Shape Reconstruction

Another important tool that we develop is stereo triangulation for 3D reconstruction. Specifically, we want to recover the 3D positions of the facial feature points of the actor recorded in order to provide facial shape information related to different facial expressions and speech gestures. The idea of stereo triangulation is pretty straightforward: With a calibrated camera, we can determine a ray in 3D space on which an object point must lie. With two calibrated cameras posed from different views, we can determine the 3D position of the object point by intersecting the two rays.

Suppose that we have two camera coordinate systems and a certain world coordinate system. Let $\overline{X}_1 = [X_1, Y_1, Z_1]^T$, $\overline{X}_2 = [X_2, Y_2, Z_2]^T$, and $\overline{X}_w = [X_w, Y_w, Z_w]^T$ be the coordinates of the same point $P$ in space with respect to camera 1 coordinate system, camera 2 coordinate system, and world coordinate system, respectively. Then $\overline{X}_1$ and $\overline{X}_w$, $\overline{X}_2$ and $\overline{X}_w$ are related by the following rigid transformation:

$$\overline{X}_1 = R_1 \overline{X}_w + T_1 \tag{1}$$

$$\overline{X}_2 = R_2 \overline{X}_w + T_2 \tag{2}$$

where $R_1, R_2, T_1, T_2$ are the rotation matrices and translation vectors characterizing the rigid transformation from world coordinate system to camera 1 coordinate system and camera 2 coordinate system, respectively.

Let $\overline{x}_1 = [x_1, y_1, z_1]^T = \overline{X}_1 / Z_1$ and $\overline{x}_2 = [x_2, y_2, z_2]^T = \overline{X}_2 / Z_2$ be the normalized coordinates of the central (pinhole) projection of the point $P$ onto the two image planes of camera 1 and camera 2, respectively. Our goal is to compute the 3D coordinates $\overline{X}_1$ and $\overline{X}_2$ from 2D image projections $\overline{x}_1$ and $\overline{x}_2$ by stereo triangulation.

From Equation 1 and Equation 2, we can show that

$${R_1}^{-1}(\overline{X}_1 - T_1) = {R_2}^{-1}(\overline{X}_2 - T_2) \tag{3}$$

That is

$${R_1}^T \overline{X}_1 - {R_1}^T T_1 = {R_2}^T \overline{X}_2 - {R_2}^T T_2 \tag{4}$$

Equation 4 can be rewritten as

$$Z_1 {R_1}^T \overline{x}_1 - Z_2 {R_2}^T \overline{x}_2 = {R_1}^T T_1 - {R_2}^T T_2 \tag{5}$$

In matrix form, we have

$$[{R_1}^T \overline{x}_1 - {R_2}^T \overline{x}_2] \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} = {R_1}^T T_1 - {R_2}^T T_2 \tag{6}$$

Let $A = [{R_1}^T \overline{x}_1 - {R_2}^T \overline{x}_2]$, $\overline{z} = [Z_1, Z_2]^T$, and $b = {R_1}^T T_1 - {R_2}^T T_2$. $A$ is a $3 \times 2$ matrix. Thus, Equation 6 is an overconstrained linear equation $Az = b$ whose least-squares solution is as follows:

$$\begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} = (A^T A)^{-1} A^T ({R_1}^T T_1 - {R_2}^T T_2) \tag{7}$$

Finally, $\overline{X}_1$ and $\overline{X}_2$, the 3D coordinates of $P$ with respect to camera 1 coordinate system and camera 2 coordinate system respectively, are obtained by

$$\overline{X}_1 = Z_1 \overline{x}_1, \overline{X}_2 = Z_2 \overline{x}_2 \tag{8}$$

## 5. EXPERIMENTS AND EVALUATIONS

We have conducted several experiments to evaluate our system with respect to the software tools.

### 5.1. Evaluation 1: Calibration Error

For each camera, we collected 12 images of a planar checkerboard posed at different orientations. The corners of the grids of the checkerboard pattern in each image were automatically extracted and were used to calibrate both the intrinsic and extrinsic parameters of the cameras. In order to verify the calibration result, we reproject the grids onto the original images and computed the reprojection error. Figure 4 displays the reprojection error of one of the cameras in the system. We observe that the average error is 0.16279 and 0.13482 pixels in the x and y directions, respectively.

### 5.2. Evaluation 2: Facial marker tracking

We verified the tracking tool on real facial motion data. During recording, 9 facial markers are attached to the key points of the face of an actor. The actor demonstrated different facial expressions and speech gestures as well as head movements. In order for the marker collocation model to collect sufficient statistics, we used $N = 1000$ frames of data for training. In order to illustrate that the marker collocation model can recover tracking errors, we added a flying artificial bar to the video data to simulate occlusions. Figure 3 shows a snapshot of the tracking results with and without artificial occlusion.

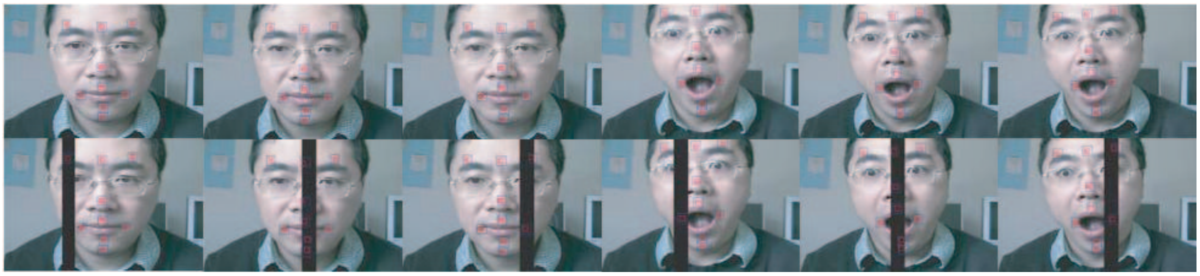

**Fig. 3**. *A snapshot of the tracking results with and without artificial occlusion.*

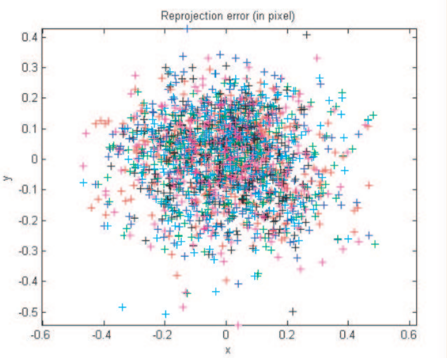

**Fig. 4**. *Reprejection error (in pixel) of one of the cameras.*

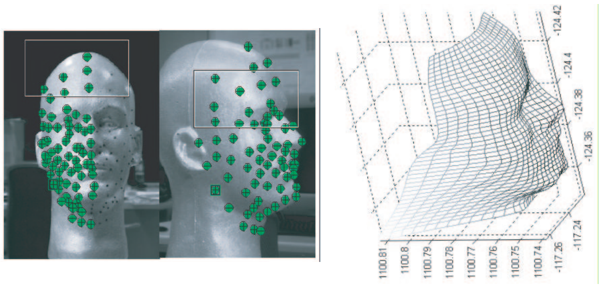

**Fig. 5**. *3D facial shape reconstruction.*

### 5.3. Evaluation 3: 3D shape reconstruction

We performed an experiment to testify stereo triangulation for 3D shape reconstruction. We took a two-view snapshot of a foam head model using our system. The head model has color facial markers attached at the key positions of the face. By scanning the model using a Cyberware 3D scanner we obtained the ground truth of the 3D coordinates of the markers. We manually labeled the correspondences of the markers in the two images, as shown in Figure 5 (left), and obtained two sets of 2D coordinates of those markers. By stereo triangulation, we recovered the 3D coordinates of the labeled facial markers with respect to a camera coordinate. Figure 5 (right) illustrates the interpolated and smoothed meshes of the recovered 3D positions of the facial markers. Comparing to the ground truth, the average reconstruction error is 3.16 mm.

## 6. CONCLUSION

3D multimodal face data are required for many human computer interface tasks including multimodal speech recognition, facial expression/speech co-articulation and lip reading, etc. In order to build a dynamic 3D face database, a camera/microphone array system is designed and implemented. The system is able to capture synchronized face video and reconstruct the face model according to the tracked facial feature points. Both hardware and software issues are systematically investigated. Preliminary experiment results justify that the proposed framework is able to capture synchronized multi-view face images in real-time with decent reconstruction accuracy. Besides the data capture system, necessary tools for camera calibration, color space conversion, 2D facial feature tracking and 3D face shape recovery are also provided for testing and extension purpose. For more details about our data, codes and algorithms, please refer to: ifp.uiuc.edu/˜hu3/3DDate/project.htm.

## 7. REFERENCES

[1] K. Chang, K. Bowyer, P. Flynn, "An evaluation of multimodal 2D+3D face biometrics," IEEE Trans. on Pattern Analysis and Machine Intelligence, V. 27, I. 4, Apr. 2005 pp. 619 - 624.

[2] L. Yin, X. Wei, Y. Sun, J. Wang, and M. Rosato, "A 3D Facial Expression Database For Facial Behavior Research," 7th International Conference on Automatic Face and Gesture Recognition (FG2006), Apr. 2006. pp. 211 - 216.

[3] http://www.sic.rma.ac.be/˜ beumier/DB/3d_rma.html

[4] http://gavab.escet.urjc.es/articulos/GavabDB.pdf

[5] http://www-users.cs.york.ac.uk/˜ tomh/3DFaceDatabase.html

[6] http://www.ee.surrey.ac.uk/Research/VSSP/xm2vtsdb

[7] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek, "Overview of the face recognition grand challenge," CVPR2005, June 2005.

[8] Yuxiao Hu et. al., "Building Large Scale 3D Face Database for Face Analysis," MCAM 2007, pp. 343-350.

[9] L. Liang, Y. Luo, F. Huang, and A. Nefian, "A multistream audio-video large-vocabulary Mandarin Chinese speech database," ICME 2004, June, 2004, Taiwan, China.

[10] http://www.vision.caltech.edu/bouguetj/calib_doc/index.html.

[11] J. Barker, "Tracking Facial Markers with an Adaptive Marker Collocation Model," ICASSP 2005, Philidelphia, pp. 665-669.