Binocular Photometric Stereo Acquisition and Reconstruction for 3D Talking Head Applications

Chaoyang Wang[12], *Lijuan Wang*[1], *Yasuyuki Matsushita*[1], *Bojun Huang*[1], *Magnetro Chen*[1], *Frank K. Soong*[1]

[1] Microsoft Research Asia, Beijing, China
[2] Department of Computer Science, Shanghai Jiao Tong University, Shanghai, China

wangchaoyang@sjtu.edu.cn, {lijuanw, yasumat, bojhuang, machen, frankkps}@microsoft.com

Abstract

In order to render a high quality, versatile 3D talking head, a stable, high frame rate AV data acquisition system is constructed. It can capture 3D position, surface orientation and albedo texture of the talking head video images along with the corresponding speech signals. The system consists of a computer controlled LED lighting subsystem; high speed stereo cameras; a microphone; and a computer for synchronous recording of multi-stream AV data. The visual image data collected is processed through a binocular photometric stereo 3D reconstruction pipeline. The pipeline automatically segments out the face; computes the depth map with binocular stereo; computes the normal map with photometric stereo; generates albedo texture; and finally constructs a high-detailed 3d model with depth and normal cues as constraints. By using the data collected with the built system, we can capture high quality dynamic facial performance, synchronized with the subject's uttered speech.

Index Terms: talking head, binocular photometric stereo, facial performance capture

1. Introduction

A Realistic talking head has a wide range of applications, including video games, movie characters, assisted language teachers and virtual guides. While cartoon avatars are relatively easier to build, human-like realistic avatars seen in games and movies are much harder to build as any unnatural deformations make the resultant output fall into the "uncanny valley" of human rejection. Image-based facial animation techniques achieve great realism in synthesized videos by combining different facial parts of recorded 2D images [1-5]. However, it is challenging to freely change the head pose or to render different facial expressions. 2.5D talking head [6] wraps face images around a smooth geometric model, which can achieve both 3D-feel and realism to some extent. But when the head moves in large angles, the artifacts become obvious due to inaccurate facial geometry. To build a real versatile 3D talking head, high quality dynamic facial performance capture is the essential first step. In this paper, we describe a prototype system for high-fidelity 3D talking head recording. We begin by briefly reviewing previous studies that relate to our work.

Marker based capture

One conventional approach to facial performance capture is to track a set of sparse hand-placed markers attached onto a face using a single or multiple video cameras [7-10]. This approach provides robust tracking of very expressive performances; however, it is by nature limited in resolution and laborious in putting on markers. Furthermore, the markers need to be digitally removed if a facial texture is required.

Structured light capture

Similar to marker-based capture, structured light systems project known patterns on a face for densely measuring the shape [11] [12]. Acquiring the face color, however, becomes non-trivial, since uniform illumination must be temporally interleaved with the structured light, which consequently reduces the temporal resolution.

Passive capture

Beeler *et al*. [13] [14] show the possibility of reconstructing pore-scale facial geometry using a high-resolution multi-view dense stereo matching. Because of its necessity of high-resolution, it is not straightforward to apply it to temporally-dense capture. In addition, the multi-camera system requires careful calibration, and the computational cost of high-quality stereo matching is high.

Photometric stereo

Instead of directly measuring position or depth, photometric stereo estimates surface orientations by measuring the shading variations of a surface under different illuminations [15] [16] [20]. A 3D shape up to a scale can be obtained via integration of the obtained surface normal field. Photometric stereo typically uses a simple reflectance model, *e.g.*, Lambertian model, which makes it computation friendly. Another advantage of photometric stereo is that it provides an albedo map for realistic texture rendering [20].

Binocular photometric stereo

Combining advantages from both depth and normal sensors has been recently studied [18] [19][27]. The combination approach achieves high quality 3D reconstruction by fusing the coarse base geometry estimated by the depth sensor, and high resolution details obtained via by photometric stereo. In addition, light calibration can be automated by using cues from base geometry [17] [21].

We follow this direction in designing our system because of these advantages. Unlike previous works that are restricted to static targets, we address the issues of measuring dynamic targets by developing a practical running system. The challenges are (1) the system requires full control over hardware and software to trigger lights and camera shutters in a synchronized manner, (2) data transmission and storage increase dramatically in high frame rate capture, and (3) high frame rate results in a short exposure time, which may degrade the image quality as less photons being captured. Moreover, since the lip movement is one of the fastest motion on a face, the camera frame rate needs to be high enough to precisely record the movement.

In this paper, we develop a high-fidelity 3D capture system for recording dynamic 3D faces based on a binocular photometric stereo approach. The system aims at recording facial expressions and articulator movement during speech at a high frame rate. Designed for 3D talking head applications, our system captures 3D dynamic facial performance, along with synchro-

nized audio. Because our system computes surface normal and albedo textures in addition to depth, we are able to render photo-realistic 3D faces with the recorded data.

2. Data acquisition

2.1. Video audio recording system

The prototype system we build consists of a set of LED lights, two high speed cameras, a microphone, and a PC for controlling multi-stream synchronization and data storage (see Fig. 1). 16 individually controllable LEDs are fixed on a rectangular frame that is attached to a 24-inch display monitor. The lighting patterns can be controlled at a frame rate of 500> [Hz]. On top of the screen, two Point Grey Flea3 cameras are horizontally placed as a stereo configuration.

During recording, an actor sits in front of the screen, making expressions or reading text prompted on the screen. The stereo cameras capture images at a frame rate of 100fps, while the lighting system repeatedly varies the light patterns using four pre-defined patterns (see Fig. 2) at the same rate. The cameras are configured to operate in a trigger mode to synchronize with the lighting, and audio recording is also synchronized with each image frame by recording its timestamp. Overall, the system is able to record four pairs of images under four different lighting conditions every 1/25 seconds together with synchronized audio. Since the camera frame rate in our current setup is 100 [fps], and we use four pairs of images for a single reconstruction, the real frame rate of our data acquisition system is 25 [fps]. Since the patterns are switched cyclically in a sliding window fashion, we can achieve a virtual frame rate of 100 [fps], which is sufficient to capture subtle motion of human expression.

2.2. 3D face reconstruction pipeline

Figure 3 shows overview of the pipeline. After audio-video recording, the captured dynamic sequence is fed to the 3D reconstruction pipeline, where the input is converted to a sequence of 3D shape models along with normal maps and albedo textures. Our method first automatically segments out the target object from its background, computes a depth map using binocular stereo, then computes the normal map by photometric stereo using the previously produced depth map as a cue for photometric calibration. Albedo texture is also generated after normal map computation. Finally, the 3D model is reconstructed by fusing the depth and normal.

3. Binocular photometric stereo: 3D face reconstruction pipeline

3.1. Binocular photometric stereo

Binocular photometric stereo is a combination of depth estimation by triangulation and normal estimation from shading variations. Binocular stereo computes the distance from a camera to a scene point by measuring the disparity of the point in the two projection positions in the cameras. Photometric stereo on the other hand, measures the intensity variations of a point under varying lightings and computes the surface normal (see Fig. 4). Binocular stereo provides coarse yet reliable depth estimates, while photometric stereo gives fine scale details in the form of surface normal.

Using both the coarse base shape estimated by binocular stereo and the high-resolution normal information, a detailed

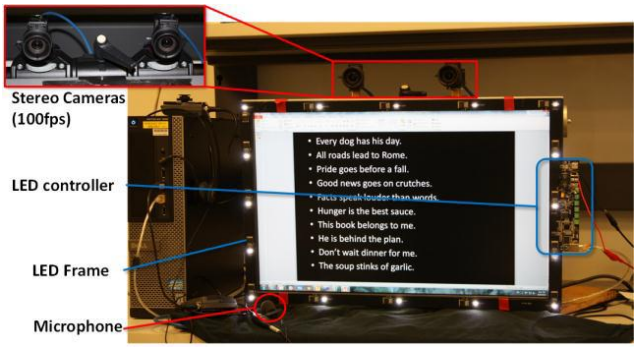

Fig. 1: A picture of our prototype system.

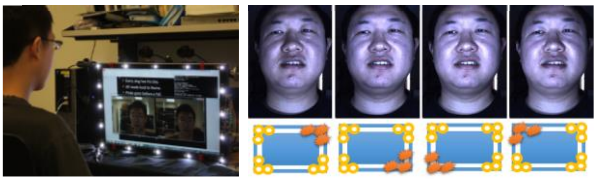

Fig. 2: Left is a recording scene. Right are four adjacent frames captured by one camera, with their corresponding LED lighting patterns at the bottom.

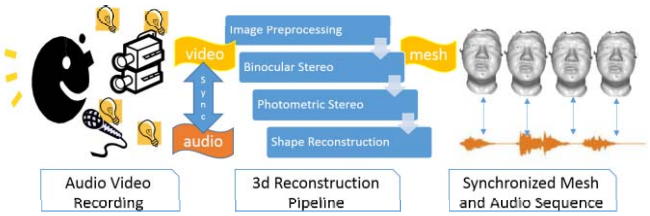

Fig. 3: Overview of our data acquisition and reconstruction system.

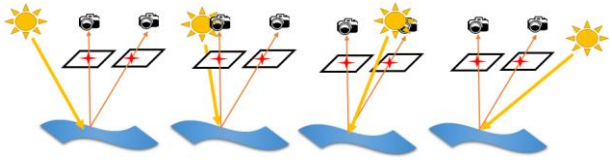

Fig. 4: An illustration of binocular photometric stereo.

shape can then be computed by fusing them. To deal with a moving object, *i.e.*, human face and articulators, the same mechanism can be applied using a high speed capture. The underlying assumption is that when the recording frame rate is much faster than the object motion speed, the surface geometry at time t is almost the same as $t + \delta t$. Therefore, by building a high speed recording setup, we are able to utilize a group of consecutive frames to reconstruct a 3D geometry at time t.

3.2. Preprocessing

Captured images are first resized by a factor of two to reduce sensor noise. They are next stereo-rectified so as to simplify a 2-dimensional correspondence problem to a 1-dimensional stereo matching problem. Each image is then converted to a gray scale image. The gray scale image is denoted as J_k^i, refers to k-th rectified image captured by camera i.

Moreover, a binary mask M_k is computed to segment out the target object from the background. General image segmentation usually requires human interaction and parameter adjustment and suffers from unstable performance under extreme illumination variations. In our particular case, we develop a simple implementation that utilizes the assumption that the target object of a scene has a larger variance in intensity under

different illuminations than that of the background. So the binary mask M_k^i is computed as

$$M_k^i(u,v) = \begin{cases} 1 & \text{if } \frac{1}{4}\sum_{j=k-1}^{j=k+2}\left(J_j^i(u,v) - \mu_j(u,v)\right)^2 > \beta \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

where (u,v) is the pixel location, μ is the mean image of four adjacent frames, and β denotes a threshold. We use $\beta = 5$ in all our experiments. Only the largest connected region in M_k^i is retained and holes in this region are filled. The segmentation mask is essential to the rest of the pipeline, since it explicitly defines the region where we assume the Lambertian reflectance and continuity of the surface.

3.3. Binocular stereo

In traditional stereo vision, two cameras, placed horizontally from one another are used to obtain two different views of a scene, in a manner similar to human binocular vision. By comparing these two images, the relative depth information can be obtained, in the form of *disparities*, which are inversely proportional to the differences in distance from a camera to scene points. Dense stereo matching can be formulated in a Markov random field (MRF) framework, where each pixel has a set of *K* labels. These labels represent *K* candidate disparities, which are the top *K* peaks in the normalized cross correlation (NCC) score. This formulation is a simpler version of Campbell *et al.*'s [22]. The optimization of the energy function assigns a label k_p to each pixel p inside the binary mask M_k^0.

$$E(k) = \lambda \sum_p \phi(k_p) + (1-\lambda) \sum_{(p,q)} \varphi(k_p, k_q), \quad (2)$$

where q denotes neighboring pixels, and λ is a weight parameter which in our experiment is set as $\lambda = 0.5$. The cost of a labeling $k = \{k_p\}$ consists of two terms: a unary potential $\phi(k_p)$ which represents photo-consistency, and a pairwise term $\varphi(k_p, k_q)$, which defines smoothness of disparities.

The images for stereo matching are created using 4 channels. Each channel corresponds to a gray image recorded under a distinct lighting. Hence unary potential is computed as 4-dimensional normalized cross correlation (NCC), with a range between 0 and 1.

The pairwise term is formulated as a normalized depth difference of two adjacent pixels:

$$\varphi(k_p, k_q) = \frac{2\left|z_{p,k_p} - z_{q,k_q}\right|}{z_{p,k_p} + z_{q,k_q}}, \quad (3)$$

where z_{p,k_p} denotes the depth of pixel p with disparity k_p.

The energy function is optimized using tree-reweighted message passing (TRW-S) [23]. In our experiment, the number of label *K* is set as 10, and maximum number of iteration of TRW-S optimization is set as 20, since in most cases the result converges after 20 iterations.

Even after optimization, the disparity map we obtain may still contain errors. These errors typically form small isolated islands off the correct surface (see the black region on the left side of the neck in Fig. 5(b) and (c)). Thus, we can detect and filter out these error pixels by applying a smoothing filter. Filtering these outliers are essential to shape reconstruction because a patch with a large error would create a spike-like artifact in the final result.

3.4. Photometric stereo

3.4.1. Uncalibrated photometric stereo

Basic assumptions of traditional photometric stereo are parallel lighting and the surface reflectance follows Lambert's law, *i.e.*, $d = \rho(\boldsymbol{n}^T\boldsymbol{l})$, where d is an observed intensity, ρ is diffuse albedo, $\boldsymbol{n} \in \mathbb{R}^3$ is a unit surface normal, and $\boldsymbol{l} \in \mathbb{R}^3$ is a unit illumination direction. Given measurements under Q distinct lighting conditions, the image formation model can be written in a matrix form using an observation matrix $\boldsymbol{D} \in \mathbb{R}^{P\times Q}$, where *P*-pixel images are cascaded as column vectors:

$$\boldsymbol{D} = \boldsymbol{S}\boldsymbol{L}. \quad (4)$$

Each row of $\boldsymbol{D}$ is a vector intensity values of the same pixel under different illumination conditions. Each row of pseudo normal matrix $\boldsymbol{S} \in \mathbb{R}^{P\times 3}$ is $\rho_p \boldsymbol{n}_p^T$, the scalar product of albedo and surface normal at pixel p, and each column of $\boldsymbol{L} \in \mathbb{R}^{3\times Q}$ refers to a lighting direction. Traditional photometric stereo obtains illumination directions $\boldsymbol{L}$ by light calibration, which is laborious and needs to be done every time before recording. In our case, however, we can perform auto-calibration by making a full use of the depth map we obtain from binocular stereo.

Uncalibrated photometric stereo [26] shows that by factorizing $\boldsymbol{D}$ using singular value decomposition (SVD), we can solve for illumination directions and normals up to a 3×3 linear ambiguity. Because of the Lambertian image formation model, $\boldsymbol{D}$ should be a rank-3 matrix. Thus the best rank-3 approximation in the least-squares sense can be obtained as:

$$\boldsymbol{D} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T, \quad \tilde{\boldsymbol{S}} = \tilde{\boldsymbol{U}}\tilde{\boldsymbol{\Sigma}}^{\frac{1}{2}}\boldsymbol{A}, \quad \tilde{\boldsymbol{L}} = \boldsymbol{A}^{-1}\tilde{\boldsymbol{\Sigma}}^{\frac{1}{2}}\tilde{\boldsymbol{V}}^T, \quad (5)$$

where $\tilde{\boldsymbol{U}}, \tilde{\boldsymbol{\Sigma}}, \tilde{\boldsymbol{V}}, \tilde{\boldsymbol{S}}, \tilde{\boldsymbol{L}}$, denotes the best rank-3 approximate of $\boldsymbol{U}, \boldsymbol{\Sigma}, \boldsymbol{V}, \boldsymbol{S}, \boldsymbol{L}$, and $\boldsymbol{A}$ is an arbitrary invertible 3×3 transformation matrix. Normal map $\boldsymbol{N} \in \mathbb{R}^{P\times 3}$ is obtained by normalizing the pseudo normal map $\boldsymbol{S}$. Thus the major problem of uncalibrated photometric stereo is resolving the ambiguity $\boldsymbol{A}$. Suppose we have a rough guess of the normal map $\boldsymbol{N}_d$ that is estimated from the depth map produced by binocular stereo. With an assumption of a uniform albedo map, we can approximate the ambiguity $\boldsymbol{A}$ by $\left(\tilde{\boldsymbol{U}}\tilde{\boldsymbol{\Sigma}}^{\frac{1}{2}}\right)^{\dagger}\boldsymbol{N}_d$, where operator † refers to Moore-Penrose pseudo inverse. The uniform albedo assumption is practical since facial skin is approximately constant.

3.4.2. Normal estimation from depth map

Since the resolution of the depth map estimated by binocular stereo is low (see Fig. 5(c)), it is difficult to directly compute surface normal by differentiation. Hence depth map Z_d is first made differentiable by applying a guided filter [24] using the normalized pseudo normal map $\tilde{\boldsymbol{S}}$ as a guidance image. The un-normalized form of surface normal at a pixel location (u,v) is then computed by:

$$N_d(u,v) = \begin{pmatrix} (Z_{u+1,v} - Z_{u,v})/(X_{u+1,v} - X_{u,v}) \\ (Z_{u,v+1} - Z_{u,v})/(Y_{u,v+1} - Y_{u,v}) \\ 1 \end{pmatrix}. \quad (6)$$

3.5. Shape reconstruction

In order to fuse the coarse base shape estimated by binocular stereo and the high-resolution normal details of photometric stereo, the shape reconstruction problem can be formulated as a Poisson equation with both metric depth and normal constraints as

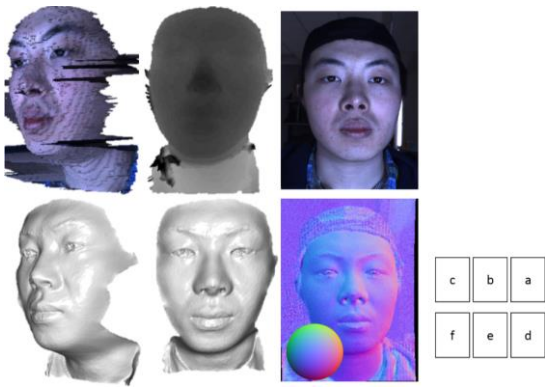

Fig. 5: (a) image captured by one camera. (b) disparity map computed by binocular stereo, in which darker color means larger disparity, thus smaller depth. (c) terrace-like surface rendered by projecting disparity map into 3D space. Notice those spike-like artifacts which are caused by gross errors in disparity map. (d) normal map produced by photometric stereo. (e)&(f) frontal and right view of the final reconstructed 3D face.

$$\begin{pmatrix} \lambda E \\ \nabla_u^2 + \nabla_v^2 \end{pmatrix} Z = \begin{pmatrix} \lambda Z_d \\ \nabla_u \left(\frac{N_x}{N_z} \circ \nabla_u X_d \right) + \nabla_v \left(\frac{N_y}{N_z} \circ \nabla_v Y_d \right) \end{pmatrix}, \quad (7)$$

where Z is the final depth map to be solved, E is an identity matrix, (X_d, Y_d, Z_d) is a 3D point location estimated by binocular stereo, and (N_x, N_y, N_z) forms the three channels of normal map. $\frac{N_x}{N_z}$ is defined as element-wise division, and operator $\circ$ represents element-wise product. In our experiment, λ is set to 0.1. The normal constraint is formulated using the second order derivatives, which performs well in tolerating noise from both normal map and depth map in the reconstruction process.

The final 3D shape is obtained from the dense point cloud $\mathcal{P}$, which is the re-projection of the depth map Z, with a domain of pixels inside the binary mask M_k^0:

$$\mathcal{P} = \left\{ \left(\frac{(u - c_x) Z_{u,v}}{f}, \frac{(v - c_y) Z_{u,v}}{f}, Z_{u,v} \right) : M_k^0(u, v) = 1 \right\} \quad (8)$$

where (c_x, c_y) is the principal point and f is the focal length of the camera.

4. Experimental results

4.1. Experiment setup

We evaluate our system by capturing facial performance sequences of multiple volunteers. As described in Section 2.1, two stereo cameras are configured as a trigger mode, capturing 640×480 RGB image sequences, with 4 [ms] shutter time, at 100 [fps] frame rate. Stereo cameras are calibrated using Camera Calibration Toolbox for Matlab [25]. The reconstruction pipeline is majorly implemented in Matlab, and thewhole process is fully automatic.

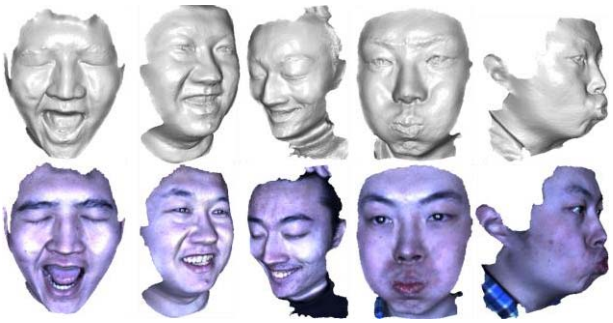

Fig. 6: 3D reconstructions of different facial expression (upper/lower figures wo/w texture).

4.2. Experiment on real-life data

Volunteers are first asked to perform exaggerated expressions, like smiling, pouting, mouth wide-open, *etc.*, as shown in Fig. 6. Facial details, such as wrinkles and small pimples, are all clearly visible on the reconstructed surface. This experiment shows that our system faithfully recovers different facial shapes and is able to record diverse facial performances.

The second experiment is for 3D talking face data acquisition. Figure 7 shows a sequence of a subject reading a prompted text sentence. The first row shows recovered shapes without textures for the purpose of showing fine details. The second row shows renderings with estimated albedo textures. The result shows that our system is able to capture rapid articulator movement (including lips and teeth) during speaking.

5. Discussion

One limitation of our current system is a lack of side and back view of the face due to the limitation of camera viewing angles. Thus, one possible extension of our system is to add more cameras using multi-view photometric stereo.

By using the data collected with our system, we are in the process of constructing a high quality, dynamic 3D talking head model, synchronized with speech. Furthermore, after collecting a sufficient amount of data, statistical talking head based on Hidden Markov Models (HMMs) can be trained to render high quality 3D talking head for any given text or speech input.

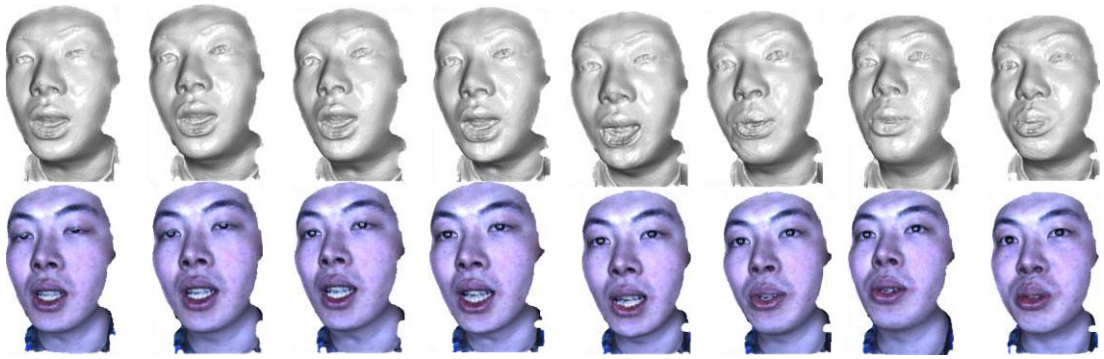

Fig. 7: 3D reconstruction of a speech animation sequence. For a complete animation sequence, see http://research.microsoft.com/en-us/projects/hd_talking_head/demo_zx5.avi

6. References

[1] Wang, L.-J., Qian, X.-J., Ma, L., Chen, Y.-N., and Soong, F., "A Real-Time Text to Audio-Visual Speech Synthesis System", in INTERSPEECH, 2338-2341, 2008.

[2] Wang, L., Qian, X.-J., Han, W., and Soong, F., "Synthesizing Photo-real Talking Head via Trajectory-guided Sample Selection", in INTERSPEECH, 446-449, 2010.

[3] Cosatto, E., and Graf, H.P., "Photo-Realistic Talking Heads from Image Samples", in IEEE Trans. Multimedia, 2(3): 152-163, 2000.

[4] Ezzat, T., Geiger, G., and Poggio, T., "Trainable Video Realistic Speech Animation," in Proc. ACM SIGGRAPH2002, San Antonio, Texas, 388-398, 2002.

[5] Wang, L.-J., Qian, Y., Scott, M.R., Chen, G., Soong, F.K., "Computer-Assisted Audiovisual Language Learning", in IEEE Computer 45(6): 38-47, 2012.

[6] Wang, L.-J., Han, W., Soong, F., "High Quality Lip-Sync Animation for 3D Photo-Realistic Talking Head", in ICASSP, 4592-4532, 2012.

[7] Bradley, D., Popa, T., Sheffer, A., Heidrich, W., and Boubekeur, T., "Markerless Garment Capture", in ACM Trans. Graphics (Proc. SIGGRAPH), 99, 2008.

[8] Furukawa, Y., and Ponce, J., "Dense 3D Motion Capture for Human Faces", in CVPR, 1674–1681, 2009.

[9] Bickel, B., Botsch, M., Angst, R., Matusik, W., Otaduy, M., Pfister, H., and Gross, M., "Multi-scale Capture of Facial Geometry and Motion", in ACM Trans. Graphics (Proc. SIGGRAPH), 33, 2007.

[10] Lin, I.-C., and Ouhyoung, M., "Mirror Mocap: Automatic and Efficient Capture of Dense 3D Facial Motion Parameters from Video", in Visual Computer 21(6): 355–372, 2005.

[11] Wand, M., Adams, B., Ovsjanikov, M., Berner, A., Bokeloh, M., Jenke, P., Guibas, L., Seidel, H.-P., and Schilling, A., "Efficient Reconstruction of Nonrigid Shape and Motion from Real-time 3D Scanner Data", in ACM Trans. Graph. 28(2): 1–15, 2009.

[12] Zhang, L., Snavely, N., Curless, B., and Seitz, S. M., "Spacetime Faces: High Resolution Capture for Modeling and Animation", in ACM Trans. Graphics 23(3): 548–558, 2004.

[13] Beeler, T., Bickel, B., Sumner, R., Beardsley, P., and Gross, M., "High-quality Single-shot Capture of Facial Geometry", in ACM Trans. Graphics (Proc. SIGGRAPH), 40, 2011.

[14] Beeler, T., Hahn, F., Bradley, D., Bickel, B., Beardsley, P., Gotsman, C., Sumner, R. W., and Gross, M., "High-quality Passive Facial Performance Capture Using Anchor Frames", in ACM Trans. Graph., 30(4):75, August 2011.

[15] Woodham, R., "Photometric Method for Determining Surface Orientation from Multiple Images", in Optical Engineering, 19(1):139–144, 1980.

[16] Kim, H., Wilburn, B., and Ben-Ezra, M., "Photometric Stereo for Dynamic Surface Orientations", in ECCV, 59–72, 2010.

[17] Hernandez, C., and Vogiatzis, G., "Self-calibrating a Realtime Monocular 3d Facial Capture System", in Proceedings International Symposium on 3D Data Processing, Visualization and Transmission (3DPVT), 2010.

[18] Nehab, D., Rusinkiewicz, S., Davis, J., and Ramamoorthi, R., "Efficiently Combining Positions and Normals for Precise 3D Geometry", in ACM Trans. on Graphics, 24(3):536–543, Aug. 2005.

[19] Wu, C.-L., Wilburn, B., Matsushita, Y., and Theobalt, C., "High-quality Shape from Multi-view Stereo and Shading under General Illumination", in CVPR, 969-976, 2011.

[20] Ma, W.-C., Hawkings, T., Peers, P., Chabert, C.-F., Weiss, M., and Debevec, P., "Rapid Acquisition of Specular and Diffuse Normal Maps from Polarized Spherical Gradient Illumination", in Rendering Techniques 2007: 18th Eurographics Workshop on Rendering, 183–194, June 2007.

[21] Joshi, N., and Kriegman, D., "Shape from Varying Illumination and Viewpoint," in ICCV, 2: 1-7, 2007.

[22] Campbell, N.D.F., Vogiatzis, G., Hernandez, C., and Cipolla, R., "Using Multiple Hypotheses to Improve Depth-maps for Multi-View Stereo", in ECCV, 766–779, 2008.

[23] Kolmogorov, V., "Convergent Tree-reweighted Message passing for energy minimization", in IEEE Trans. Pattern Anal. Mach. Intell, 28(10): 1568-1583, 2006.

[24] He, K., Sun, .J, and Tang, X. "Guided Image Filtering", in ECCV, 1-14，2010.

[25] Bouguet, J. Y., "Camera calibration toolbox for matlab", Technical report, 2007. Software available at http://www.vision.caltech.edu/bouguetj/calib doc/.

[26] H. Hayakawa. "Photometric Stereo under a Light-source with Arbitrary Motion", Journal of the Optical Society of America, 11(11):3079–3089, November 1994.

[27] Ikeuchi, K. "Determining a depth map using a dual photometric stereo", IJRR, 6(1):15–31, 1987.