A REAL-TIME 3D HEAD MESH MODELING AND EXPRESSIVE ARTICULATORY ANIMATION SYSTEM

Jun Yu, Zeng-fu Wang

Department of Automation, University of Science and Technology of China, Hefei, China

harryjun@ustc.edu.cn, zfwang@ustc.edu.cn

ABSTRACT

In view of animated human computer interfaces, this paper proposes a 3D head mesh modeling and expressive articulatory animation system. The appearance mesh model is first reconstructed from multi-view visible images using inter-regional cooperative optimization and depth super-resolution, and the universal internal articulatory mesh is then integrated with the reconstructed appearance mesh by interpolation. After establishing the head mesh model, the anatomical and biomechanical characteristics of articulators are combined to synthesize articulatory animation. The evaluations demonstrate the system can build a realistic and vivid virtual head for animated interface in real-time.

Index Terms— Head model reconstruction, stereo matching, image super-resolution, articulatory animation

1. INTRODUCTION

Reconstructing the mesh model of an individual, i.e., head mesh modeling, and synthesizing articulatory animation are crucial for building an animated interface [1-4], and have been widely used in human computer interaction.

The head mesh modeling for appearance can be performed by scanning [5], visible images [6]. Visible image based method is the main motivation of the present work because it is not time-consuming and only needs one or few of cameras. Several data acquisition techniques have been used for invisible internal articulators [7-9]. Magnetic Resonance Imaging (MRI) can capture dense data, while it is hard to be applied in real-time for the high computational complexity. X-ray can capture dense data in real-time, but collecting data will be harmful to human health. Electro-Magnetic Articulography (EMA) can capture data in real-time, but the data are sparse. By the data acquisition mentioned above, the realism of a virtual head can be increased with an open mouth. Articulatory animation can be performed as follows. A parameterized model [10] is built using given meaningful shape parameters to control the motion of a head mesh model. Anatomical model [11-17] simulates articulatory dynamics from the view of anatomy. Data-driven model [18] learns the variations of appearance from an articulatory dataset, and renders desired animation by the parameters learned. It can generate fine-detailed animation for appearance. However, the training data of internal articulators are rather limited for the invisibility. This difficulty makes the data-driven model unable to simulate various articulatory deformations.

The proposed research includes two parts: head mesh modeling and articulatory animation. In the first part, Matting Laplacian Matrix is first used to apply the super-resolution for generating high-resolution depth images of appearance based on a stereo visible image matching approach. The reconstruction result of appearance from visible image and that of internal articulators from MRI are then fused to produce a complete head mesh model. In the second part, an anatomical model and a biomechanical model are built to deform appearance and tongue by given muscle activations directly. Although several methods employ simple rules to move a tongue model effectively, it is important to model the dynamics of internal articulators at the level of muscles for high realism. The benefits can be especially obtained when the virtual head sticking tongue out, e.g., demonstrating a particular pronunciation.

2. 3D HEAD MESH MODELING

2.1. Appearance Mesh Modeling

Wang et al. [19] proposed a stereo matching algorithm by inter-regional cooperative optimization. Suppose $R_1,\cdots,R_n$ be segmented regions, the energy function of all regions is decomposed into the sum of sub-target energy functions as:

$$E(x) = E_1(x) + \cdots + E_i(x) + \cdots + E_m(x) \quad (1)$$

Where $E_i(x)$ is the energy function of the *ith* region R_i.

The energy function of a region and its adjacent regions are first minimized simultaneously, and then the results are propagated via iterative calculation as:

$$E_i^{(k)}(x) = \min\left((1-\lambda_i^{(k)})E_i^{(k-1)}(x) + \lambda_i^{(k)}\sum_{j\neq i}\omega_{ij}E_j^{(k-1)}(x)\right),\ i,j = 1,...,m \quad (2)$$

Where $E_j(x)$ is the energy function of the *jth* region R_j, R_j is an adjacent region of R_i, λ_i, ω_{ij} are the weights.

However, the resolution of the depth image obtained above is limited, and should be increased. Suppose the

depth image pixels be independent and the distribution of noise is Gaussian with the variance σ^2, an observation model for depth super-resolution can be given as:

$$p(g/f) = \left(1/2\pi\sigma^2\right)^{N^2/2} \exp\left\{-\|g - DHf\|_2^2 / 2\sigma^2\right\} \quad (3)$$

Where *f*, *g* are the high-resolution depth image of original scene, observed low-resolution depth image (*N*×*N*). The matrix *H* represents a blurring filter. The matrix *D* represents the downsampling operator. The model means *g* is a blurred and downsampled version of *f*.

The following minimization is applied to obtain *f*:

$$f^* = \arg\min_f \ln p(g/f) = \arg\min_f \|g - DHf\|_2^2 \quad (4)$$

A high-resolution depth image cannot be obtained just by using the information in the observed low-resolution depth image. In other words, it is simple to obtain the high-resolution camera image of the same scene [20], and the depth discontinuities in a scene often occur with color or brightness changes within the associated camera image of the same scene [21]. So the information in the high-resolution camera image is helpful to obtain the high-resolution depth image of the same scene. Then *f* can be expressed with camera image *I* in an image window *w* as:

$$f_i \approx aI_i + b, \forall i \in w \quad (5)$$

The parameters *a*, *b* are obtained as follows [21]:

$$J(f) = \arg\min_{a,b} \sum_j \left(\sum_{i \in w_j} \left(f_i - a_j c_i - b_j \right)^2 + \varepsilon a_j^2 \right) = f^T L f \quad (6)$$

Where w_j is an image window around pixel *j*, *L* is a Matting Laplacian Matrix. Then $J(f)$ is added to Equation (4) as the regularization term:

$$T(f) = \|g - DHf\|_2^2 + \lambda f^T L f \quad (7)$$

By solving Equation (7) with a local optimization method, the high-resolution depth image is obtained, and the appearance mesh model is obtained afterwards.

2.2. Internal Articulatory Mesh Modeling

The universal internal articulatory mesh is first constructed based on the MRI data captured on the sagittal plane of a person [22]. The obtained meshes consist of oral cavity, mandible, palate, pharynx, teeth and tongue. Because the MRI data capturing of internal articulators is expensive, it is applied only once, and the captured universal internal articulatory mesh is integrated with the specific appearance mesh for obtaining the corresponding specific internal articulatory mesh. The integration process is as follows. Several dominant vertex pairs of the specific appearance mesh and the universal internal articulatory mesh are selected. The global motion of the universal internal articulatory mesh are obtained by the dominant vertex pairs based on the anatomical knowledge [23] and the method reported in [24]. Then $U_i = V_i - V_i^0$ (V_i^0, V_i are the coordinates of the *ith* dominant vertex of the universal internal articulatory mesh before and after displacement), are obtained by the global motion, and used to construct an interpolation function as:

$$\begin{aligned} &f(V) = \sum_i c_i e^{(\|V - V_i\|/64)} + N \cdot V + T \\ &s.t.\ U_i = f(V_i), \sum_i c_i = 0, \sum_i c_i V_i^T = 0 \end{aligned} \quad (8)$$

Where c_i, N, T are the parameters to be solved.

After solving Equation (8), the displacement of the *jth* other vertex of the universal internal articulatory mesh is calculated, and then the head mesh can be obtained.

3. ARTICULATOR ANIMATION

Based on the head mesh model discussed above, articulatory animation is synthesized for appearance and internal articulators respectively.

3.1. Animation of appearance

Anatomical model simulates facial motion by the contraction of muscles and the motion of skeleton, and then displays the appearance by the deformed skin. It is suitable for facial animation intuitively. Our anatomical model for appearance includes three parts: skeleton, skin, muscle. The *skeleton* includes the skull and mandible. The former is generally passive, and the latter has the rotation when opening mouth and the translation when stretching lips. The *skin* is approximated by an elastic mesh [11][12], and connected with muscle by two classes of spring. The first class is used to simulate the elasticity of skin. The second class is used to ensure that the skin not to be split. The *muscle* is modeled by the Waters model [13][14], which divides muscles into linear, sheet, sphincter ones. The first two are used for tension, the last is used for shrinkage.

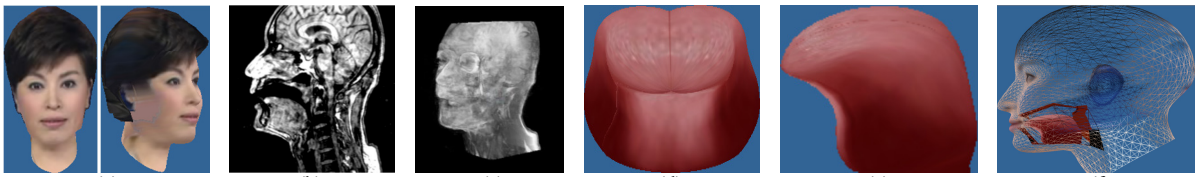
In summary, the forces on the skin mesh, including the elasticity between skin and muscle, the contraction of muscle, the drawing of jaw and the restriction of skull, are first computed. Each vertex of the skin mesh is then displaced by the forces, and the appearance is synthesized.

3.2. Animation of Internal Articulators

For internal articulators, the movements of the palate and upper teeth are passive, and the jaw and lower teeth only have the movements of up-down rotation. So they can be abbreviated, and we focus on the modeling of the tongue whose motion is more complicated and non-rigid. According to the anatomical knowledge [25], the tongue model is separated into connective tissues and muscles.

The connective tissue is modeled as the Mooney-rivlin material [26], which shows the isotropic, quasi-incompressible, non-linear, hyperelastic properties.

Although the muscle is similar to the connective tissue, it is endowed with an additional term which embodies the active and passive properties of muscle fibers [27]. The strain energy function for a muscle is given by:



(a) (b) (c) (d) (e) (f)

Fig. 1. (a) Skin mesh and hair mesh. (b) The MRI slice on midsagittal plane. (c) 3D data by combining all MRI slices. (d) Front of tongue mesh. (e) Profile of tongue mesh. (f) Profile of head mesh model after integration.

$$U = U_I(I_1, I_2) + U_J(J) + U_f\left(\lambda_f, A\right) \tag{9}$$

Where I_1, I_2 are deviatoric isotropic invariants of the strain, J is the Jacobian of the deforming gradient. $U(I_1, I_2), U(J)$ $U_f(\lambda_f, A)$ represents the active and passive non-linear mechanical behaviors of the muscle fiber during contracting, and is modeled by Hill's three element model [28], which includes a contractile element (CE), a serial elastic element (SEE) and a parallel elastic element (PE). CE is responsible for generating active force, while PE and SSE represent the passive mechanical behavior.

U_f is the strain energy function stored in the muscle fiber, and given by:

$$U_f\left(\lambda_f, A\right) = \int_1^{\lambda_{f:\max}} \left[\sigma_{PE}\left(\lambda_f\right) + \sigma_{SEE}\left(\lambda_s\right)\right] d\lambda_f \tag{10}$$

Where $\lambda_{f:\max}$ is the maximum of fiber stretch ration λ_f, λ_s is the fiber stretch ratio in SEE, σ_{PE} is the stress produced in PE, and σ_{SEE} is the stress produced in SEE.

By deriving Equation (9), the Cauchy stress is calculated as:

$$\sigma = \left(\left(U'_{I1} + U'_{I2}\bar{I}_1^C\right)\bar{B} - U'_{I2}\bar{B}^2 - \left(U'_{I1}\bar{I}_1^C + 2U'_{I2}\bar{I}_2^C\right)I/3\right)\cdot 2/J + U'_J J + U'_f\left(\lambda_f\left(n \otimes n\right) - \lambda_f I/3\right)/J \tag{11}$$

Where $\bar{B} = J^{-2/3}FF^T$ is the left Cauchy strain tensor with the volume change eliminated, $U'_{Ii} = \partial U_I / \partial I_i$, $U'_J = \partial U_J / \partial J$, n is the muscle fiber direction.

According to the biomechanical model discussed above, finite-element method is applied to simulate the tongue mesh deformation by activating muscles.

4. EXPERIMENS

Experiments are conducted using a workstation with AMD Athlon (tm) II X4 640 3.01G, memory 2G, NVIDIA GT200. The GPU+CPU framework [17] is used for acceleration.

4.1. Head Mesh Modeling

Fig. 1 shows the a head mesh modeling result. The result looks quite like the person in visible images, thus providing a solid foundation for following articulatory animation.

The mesh modeling for appearance is evaluated by visible images. The test data are the captured visible images with the resolution 352×288 from 51 persons, and used in our appearance mesh modeling algorithm and the method in [19] respectively. The evaluation of 3D head mesh modeling quality has not been more formal than observing how the mesh model projection result fits to its corresponding head shape and facial feature shape in visible images. Therefore, an index is defined as follows:

$$Q = \sum_{i=1}^{N}\sum_{j=1}^{M_i} abs\left(y_{mod}^{(i,j)} - y_{org}^{(i,j)}\right) \Big/ \left(N \cdot M_i\right) \tag{12}$$

Where N is the number of visible images, M_i is the pixel number in the head region in the *ith* visible image, $y_{org}^{(i,j)}$ is the *jth* pixel value in that region, $y_{mod}^{(i,j)}$ is the *jth* pixel value in the head region in the *ith* mesh projection image.

The mesh modeling for appearance is also evaluated by captured 3D face data for 70 persons using a Minolta Vivid 910. Fig. 3 shows a ground truth depth map and the estimated depth map. Only the skin part is used to alleviate the influence of noise. The estimated depth appears to reflect the shape of the true depth. Then they are used to calculate the average and standard deviation of the error. The unit is millimeter (mm). Table 1 shows the proposed approach can reconstruct mesh model from visible images accurately. It verifies the combination of stereo matching and depth super-resolution can generate more high-quality and high-resolution depth images of appearance by comparing with the stereo matching method in [19].

Table I The performance comparison of appearance mesh modeling.

	Mean Q	Average error	Standard deviation of error	Average time each frame takes
Our appearance mesh modeling algorithm	2.3	2.97mm	0.89mm	0.070s
The stereo matching method in [19]	3.8	4.28mm	1.29mm	0.043s

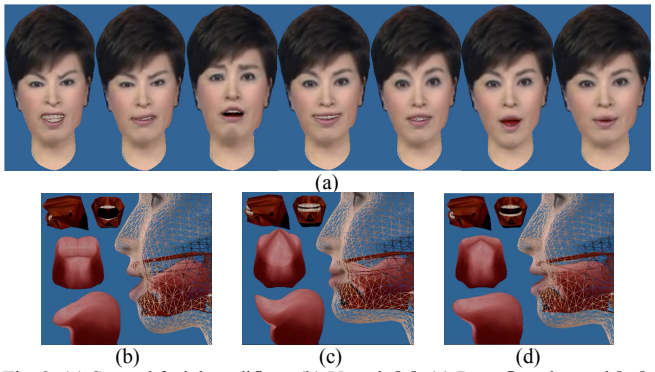
The mesh modeling for internal articulators is evaluated. The appearance meshes reconstructed from the captured visible images of 51 persons are used to estimate the corresponding internal articulatory meshes. The internal articulatory meshes of the same persons captured from MRI are used as the baselines for comparison. Then the estimated meshes are compared with the ground truth to calculate the average and standard deviation of the error. Table 2 shows the estimated internal articulatory meshes approximate to the ground truth nicely.

Table II The performance comparison of internal articulatory mesh modeling.

	Average error	Standard deviation of error	Average time each frame takes
Our internal articulatory mesh modeling algorithm	3.46mm	1.64mm	0.002s

4.2. Articulatory Animation

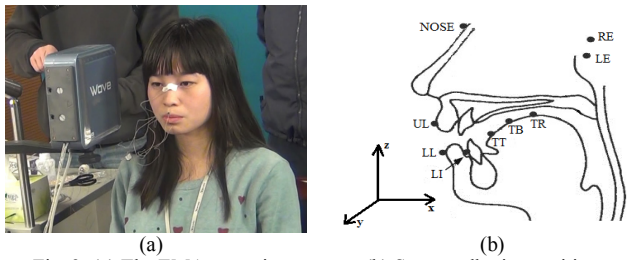
Fig. 2 shows the articulatory animation results of several modifiers and phonemes. As can be seen from it, the system can generate a virtual head which has a human-like appearance, and produce realistic articulatory animation.



(a) (b) (c) (d)

Fig. 2. (a) Several facial modifiers. (b) Vowel: [a]. (c) Retroflexed vowel [yr]. (d) Consonant in [za].

The EMA is used to trace the 3D trajectories of articulatory movements of speakers, which are used as the baseline for evaluation. The sensors are attached onto the positions (Fig. 3) as Tongue Rear (TR), Tongue Blade (TB), Tongue Tip (TT), Lower Incisor (LI), Lower Lip (LL) and Upper Lip (UL). Another 3 sensors (NOSE, left ear (LE) and right ear (RE)) are used as references to remove head movements. The articulatory data are sampled at 200 Hz.



(a) (b)

Fig. 3. (a) The EMA capturing process. (b) Sensor adhering positions.

After selecting several points on the head mesh model corresponding to EMA sensor adhering positions, we set the muscle activation values by the electromyography (EMG) data [29] and captured EMA data, and then the synthesized trajectories on the points corresponding to EMA sensor adhering positions are compared with EMA data (ground truth). Fig. 4 illustrates the comparison results of TR, TB and TT. Moreover, the synthesized trajectories are compared with the captured EMA data on 21 phonemes and 19 words to calculate the mean absolute error. Table 3 gives the average and standard deviation of the error.

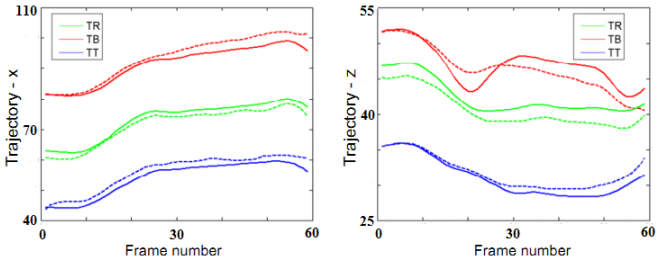


Fig. 4. Comparisons between synthesized animation (dashed line) and ground-truth (solid line) of the vowel /u/.

The appearance part of animation can also be evaluated by the captured videos. The test data are facial motion parameters extracted from 13 videos in the MPEG-4 testing database and 110 videos in the Cohn-Kanade database [31] by a facial motion tracker [32]. Then the animation videos are compared with the captured videos by the index in Equation (12). Fig. 4 and Table 3 show the synthesis results are good approximations to the ground truth, and demonstrate the animation model can synthesize realistic articulatory animation by activating corresponding muscles.

Table III The performance of internal articulatory animation.

	Mean Q	Average error	Standard deviation of error	Average time each frame takes
Our algorithm	2.6	1.57mm	0.23mm	0.06s

Moreover, the running time in Table 1, Table 2, Table 3 show the proposed system can be executed in real-time.

5. CONCLUSION

A real-time head mesh modeling and articulatory animation system is proposed for appearance and internal articulators. Stereo image matching with depth super-resolution, and MRI data are combined for head mesh modeling. Anatomical and biomechanical characteristics are combined for articulatory animation. In future, a multi-modal animation system will be developed.

6. ACKNOWLEDGEMENT

This work is supported by the National Natural Science Foundation of China (No. 61572450, No. 61303150), the Open Project Program of the State KeyLab of CAD&CG, Zhejiang University (No. A1501), the Fundamental Research Funds for the Central Universities (WK2350000002), the Open Funding Project of State Key Laboratory of Virtual Reality Technology and Systems, Beihang University (No. BUAA-VR-16KF-12), the Open Funding Project of State Key Laboratory of Novel Software Technology, Nanjing University (No. KFKT2016B08).

7. REFERENCES

[1] H. Li, J. H. Yu, et al., Realtime Facial Animation with On-the-fly Correctives, TOG, 32(4): 235-243, 2013.

[2] C. Cao, Y. Lin, W. S. Lin, K. Zhou, 3D Shape Regression for Real-time Facial Animation, ACM TOG, 32(4): 149-158, 2013.

[3] C. Cao, Q. M. Hou, K. Zhou, Displaced Dynamic Expression Regression for Real-time Facial Tracking and Animation, SIGGRAPH, 2014: 796-812.

[4] S. Suwajanakorn, I. Kemelmacher-Shlizerman, and S. M. Seitz, Total Moving Face Reconstruction, ECCV, 2014: 796-812.

[5] K. Waters, D. Terzopoulos, Modeling and animating faces using scanned data, Visualization and Computer Animation, 2(1): 123-128, 1991.

[6] L. Won-Sook, Fast head modeling for animation, Journal of Image and Vision Computing, 4(3): 355-364, 2000.

[7] L. Wang, H. Chen, S. Li, et al., Phoneme-level articulatory animation in pronunciation training, *SC*, 54(7): 74-86, 2012.

[8] K. M. Hiiemae, J. B. Palmer, Tongue movements in feeding and speech, *CROBM*, 14(6): 413-429, 2003.

[9] Q. Fang, A Study On Construction And Control Of A Three-Dimensional Physiological Articulatory Model For Speech Production, PHD Thesis, JAIST, 2009.

[10] F. I. Parke, K. Waters, Computer facial animation. Boston: Wellesley, 1996.

[11] Y. C. Lee, D. Terzopoulos, K. Waters, Realistic modeling for facial animation, *SIGGRAPH*, 1995: 55-62.

[12] W. M. Wang, et al., A physically-based modeling and simulation framework for facial animation, *ICIG*, 2009: 521-526.

[13] K. Waters, A muscle model for animating three dimensional facial expression, *Computer Graphics*, 22(4): 17-24, 1987.

[14] S. Marcos, J. G. Garcia Bermejo, E. Zalama, A realistic facial animation suitable for human-robot interfacing, *ICIRS*, 2008: 3810-3815.

[15] E. Sifakis, A. Selle, A. Robinson-Mosher, R. Fedkiw, Simulating speech with a physics-based facial muscle model, *SCA*, 2006: 261-270.

[16] R. M. Koch, et al., Simulating facial surgery using finite element models, *Annual Conference on CGIT*, 1996: 421-428.

[17] M. Bro-Nielsen, Finite element modeling in surgery simulation, *Proceedings of the IEEE*, 86(3): 490-503, 1998.

[18] V. Blanz, C. Basso, T. Poggio, T. Vetter, Reanimating faces in images and video, *Computer Graphics*, 22(3): 641-650, 2003.

[19] Z. F. Wang, Z. G. Zheng, A region based stereo matching algorithm using cooperative optimization, CVPR, 2008: 701-708.

[20] Q. Yang, R. Yang, J. Davis, and D. Nister, Spatial-depth super resolution for range images, CVPR, 2008: 153-160.

[21] A. Levin, D. Lischinshi, and Y. Weiss, A closed form solution to natural image matting, CVPR, 2006: 274-281.

[22] Jonathan Richard Shewchuk, Tetrahedral Mesh Generation by Delaunay Refinement, ACM Symposium on Computational Geometry, 1998: 86-95.

[23] P. Ekman, W. V. Friesen, Manual for the Facial Action Coding System, Palo Alto, CA: Psychologists Press, 1978.

[24] M. Kampmann, Automatic 3-D face mode adaption for model-based coding of videophone sequences, *TCSVT*, 12(3): 172-182, 2002.

[25] K. Miyawaki, A study on the musculature of the human tongue, *Annual Bulletin of the Research Institute of Logopedics and Phoniatrics*, 8: 23-50, 1974..

[26] M. Mooney, A theory of large elastic deformation, *J. Appl. Phys.*, 11(9): 582-592, 1940..

[27] C. Y. Tang, G. Zhang, C. P. Tsui, A 3d skeletal muscle model coupled with active contraction of muscle fibres and hyperelastic behavior, *Journal of biomechanics*, 42(7): 865-872, 2009..

[28] M. Kojic1, S. Mijailovic, N. Zdravkovic1, Modelling of muscle behaviour by the finite element method using Hill's three-element model, *International Journal for Numerical Methods in Engineering*, 43(5): 941-953, 1998.

[29] P. F. MacNeilage, G. N. Sholes, An electromyographic study of the tongue during vowel production, *JSHR*, 7(3): 229, 1964.

[30] M. Muller, Information Retrieval for Music and Motion. Springer Berlin Heidelberg, 2007.

[31] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, I. Matthews, The Extended Cohn-Kande Dataset (CK+): A complete facial expression dataset for action unit and emotion-specified expression, Workshop on CVPR, 2010, pp. 217-224.

[32] F. Dornaika, F. Davoine, Simultaneous facial action tracking and expression recognition in the presence of head motion, *International Journal of Computer Vision*, 76(3): 257-281, 2008.