A TECHNIQUE TO SIMULATE THE MOVEMENT OF A 3D HUMAN HEAD MODEL ACCORDING TO VIETNAMESE DIALOGUE

Tran Thanh Phuong[1], Ngo Duc Vinh[2], Ha Manh Toan[3], Do Nang Toan[3*], Nong Minh Ngoc[4]
[1]Lac Hong University, [2]Hanoi University of Industry,
[3]Institute of Information Technology - Vietnam Academy of Science and Technology, [4]Thai Nguyen University

ARTICLE INFO

Received: 07/7/2023
Revised: 30/8/2023
Published: 31/8/2023

KEYWORDS

Virtual reality
Morphing
3D head
Vietnamese dialogue
Handling syllables

ABSTRACT

In the field of virtual reality, the problem of simulating gestures, 3D head states is a topic that has received much attention because of its strong applicability in many real-life problems, such as building personal objects in movies, in systems of virtual broadcasters, virtual guides... The article is interested in the problem of simulating the movements of the 3D human head model following Vietnamese dialogue and the implementation contents with a sequential nature from 3D head modeling problems to control algorithms, transform the model. Specifically, the article presented how to build a 3D human head model, how to calculate transformation vectors based on individual syllable representations as well as model transformation over time, based on analyzing the content of the input Vietnamese dialogue to create the desired animation. The experimental results have proved the effectiveness of the proposed proposals and are the basis for moving towards more complete human simulation virtual reality applications.

MỘT KỸ THUẬT MÔ PHỎNG CỬ ĐỘNG CỦA MÔ HÌNH ĐẦU NGƯỜI 3D THEO LỜI THOẠI TIẾNG VIỆT

Trần Thanh Phương[1], Ngô Đức Vĩnh[2], Hà Mạnh Toàn[3], Đỗ Năng Toàn[3*], Nông Minh Ngọc[4]
[1]Trường Đại học Lạc Hồng, [2]Trường Đại học Công nghiệp Hà Nội,
[3]Viện Công nghệ thông tin - Viện Hàn lâm Khoa học & Công nghệ Việt Nam, [4]Đại học Thái Nguyên

THÔNG TIN BÀI BÁO

Ngày nhận bài: 07/7/2023
Ngày hoàn thiện: 30/8/2023
Ngày đăng: 31/8/2023

TỪ KHÓA

Thực tại ảo
Biến dạng
Đầu người 3D
Lời thoại Việt Nam
Xử lý âm tiết

TÓM TẮT

Trong lĩnh vực thực tại ảo, vấn đề mô phỏng cử chỉ, trạng thái đầu người 3D là một chủ đề nhận được nhiều sự quan tâm bởi khả năng ứng dụng mạnh mẽ trong nhiều bài toán thực tế, chẳng hạn như xây dựng các nhân vật trong phim ảnh, trong các hệ thống phát thanh viên ảo, hướng dẫn viên ảo… Bài báo quan tâm đến vấn đề mô phỏng cử động của mô hình đầu người 3D theo lời thoại tiếng Việt và các nội dung thực hiện có tính chất xâu chuỗi từ những vấn đề về mô hình hóa đầu người 3D đến những thuật toán điều khiển, biến đổi mô hình. Cụ thể, bài báo đã trình bày cách thức xây dựng mô hình mẫu đầu người 3D, cách thức tính toán các vector biến đổi dựa vào các biểu diễn âm tiết riêng lẻ cũng như việc biến đổi mô hình theo thời gian, trên cơ sở phân tích nội dung lời thoại tiếng Việt đầu vào để tạo ra được hoạt cảnh mong muốn. Các kết quả thực nghiệm đã chứng tỏ được sự hiệu quả của những đề xuất được đặt ra và là cơ sở cho việc tiến tới những ứng dụng thực tại ảo mô phỏng con người hoàn thiện hơn.

DOI: https://doi.org/10.34238/tnu-jst.8297

* Corresponding author. *Email: donangtoan@gmail.com*

1. Giới thiệu

Nghiên cứu mô phỏng con người ảo 3D là một bài toán quan trọng có ý nghĩa trong lĩnh vực thực tại ảo, trong đó vấn đề mô phỏng cử động của mô hình đầu người 3D rất được quan tâm.

Đặc biệt gần đây, chủ đề mô phỏng cử động của đầu người 3D dựa trên nguồn âm thanh tiếng nói đưa vào nhằm tạo ra video đầu người nói chân thực đang là chủ đề thu hút nhiều nghiên cứu, đã có nhiều công trình đề xuất nhằm đồng bộ hóa âm thanh với chuyển động của môi [1] – [6]. Tuy nhiên, sự gắn kết giữa âm thanh và chuyển động của đầu người chưa được giải quyết. Trong khi đó, chuyển động tự nhiên và nhịp nhàng của đầu cũng là một trong những yếu tố chính để tạo ra các video đầu người nói sát thực [7], [8].

Chúng ta biết rằng, các cử chỉ, trạng thái biểu hiện khuôn mặt trong một hệ thống mô phỏng con người ảo 3D được tạo ra từ sự chuyển động của hệ thống các cơ nét mặt, kết quả là dẫn đến những biến dạng nhất định trong thời gian ngắn của những bộ phận khuôn mặt như da, lông mày, mũi, mí mắt, và môi [9]. Về mặt bản chất, mô hình 3D có thể được hiểu là một tập điểm trong không gian 3 chiều và những thông tin về mối quan hệ giữa các điểm đó. Như vậy, việc biến đổi mô hình đầu người 3D cũng sẽ xoay quanh cấu trúc dữ liệu này.

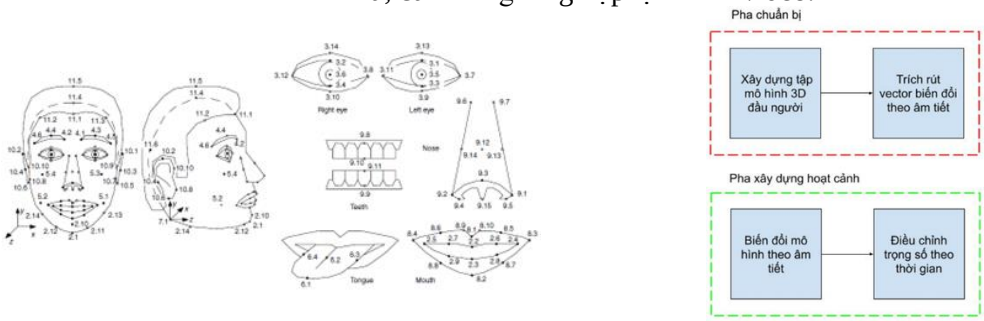
Ở đây, xét trên việc biến đổi các thành phần trên khuôn mặt nhằm mục tiêu tạo ra những sự thay đổi trên mô hình sao cho phù hợp với lời thoại, mô hình đầu người 3D có thể coi là không có sự thay đổi về liên kết đặc trưng giữa các điểm và chất liệu ảnh. Những sự thay đổi chỉ là những sự dịch chuyển có giới hạn và có sự ràng buộc theo những tiêu chí nhất định của các điểm trong mô hình đầu người 3D. Như vậy, vấn đề biến đổi mô hình sẽ được giới hạn trong nhiệm vụ biến đổi tọa độ của tập điểm trong mô hình đầu người 3D với những tham số nhất định.

Một đặc tả quan trọng được đưa ra cho vấn đề tạo hoạt cảnh đầu 3D là tập điểm trong chuẩn MPEG-4 [9]. MPEG-4 đặc tả các đặc điểm đầu người với 84 điểm đặc trưng (Hình 1a) đi kèm với các tham số hoạt hóa. Tập hợp các tham số này tương ứng với các cử chỉ của đầu người mà kết quả là gây ra sự biến dạng của mô hình đầu người so với trạng thái cân bằng trung tính. Từ việc xây dựng quá trình biến dạng của mô hình đầu người với các tham số hoạt hóa kết hợp với nhãn thời gian, ta có thể tạo ra hoạt cảnh của đầu người tương ứng. Vị trí các điểm đặc trưng được thiết lập tương ứng với các mốc trên đầu người như là miệng, mũi, mắt,… và chúng được sắp xếp theo từng nhóm.

Một số nghiên cứu đã sử dụng tập điểm MPEG-4 trong các nghiên cứu của mình. Chẳng hạn, nhóm nghiên cứu của Turban [10] đã thực hiện áp dụng chuẩn MPEG-4 vào dữ liệu các bản quét 3D chất lượng cao từ video RGB-D Kinect. Nghiên cứu của nhóm hướng đến việc xây dựng các mô hình khuôn mặt chân thực và có thể áp dụng vào tạo chuyển động cho mô hình đầu người. Một nghiên cứu khác là của nhóm Itimad [11] với việc xây dựng một mô hình mặt 3D dựa theo chuẩn MPEG-4 và có chức năng đồng bộ hóa chuyển động môi phù hợp với dữ liệu âm thanh đầu vào. Nghiên cứu cũng áp dụng hàm Raised Cosine Deformation vào dữ liệu hình học khuôn mặt và triển khai ngôn ngữ EEMML (Emotional Eye Movements Markup Language) để hướng tới tính chân thực cho hoạt cảnh khuôn mặt. Năm 2015, nhóm nghiên cứu của Zhao [12] đã công bố một công trình về tổng hợp hoạt cảnh khuôn mặt 3D có cảm xúc dựa trên mô hình FCRBM (Factored Conditional Restricted Boltzmann Machines), từ chuỗi nhãn cảm xúc và tham số FAPs (Facial Action Parameters) ban đầu, kỹ thuật lấy mẫu Gibbs được áp dụng để lấy chuỗi FAPs tương ứng và sinh ra hoạt ảnh khuôn mặt 3D dựa theo đặc tả MPEG-4.

Gần đây, một nghiên cứu nhằm đồng bộ biểu cảm khuôn mặt với tiếng nói đã được đề xuất bởi nhóm Dahmani [13] vào năm 2019. Các tác giả đã áp dụng tiếp cận học sâu vào việc xây dựng một không gian nhúng. Từ dữ liệu văn bản đầu vào, các đặc tả xác định các tham số cảm xúc và áp dụng để điều chỉnh các tham số về nhãn thời gian, âm thanh cũng như hình ảnh. Tiếp đến vào năm 2020, nhóm nghiên cứu của Liu [14] đã nghiên cứu quá trình sinh biểu cảm gồm hai bước dựa trên tiếp cận tâm sinh lý. Cụ thể, các tác giả đã thực hiện xây dựng không gian dữ liệu dựa trên các chuyển động đơn vị của khuôn mặt và tạo ra các trạng thái biểu cảm khác nhau.

Suzhen Wang [15] và đồng nghiệp đã đề xuất một bộ khung gồm các bước sau: Thực hiện tách chuyển động của đầu khỏi toàn khung hình chuyển động phụ thuộc âm thanh và dự báo chuyển động của đầu tương ứng với âm thanh, sau đó bộ tạo trường chuyển động sẽ sinh các điểm điều khiển chính từ âm thanh và tư thế đầu, cuối cùng tổng hợp lại thành video.



(a) *(b)*

Hình 1. *(a) Các điểm của chuẩn MPEG-4; (b) Sơ đồ chung của phương pháp đề xuất*

Vào năm 2022, nhóm nghiên cứu của Chi [16] đã trình bày một kỹ thuật đồng bộ môi với lời nói trên cơ sở tạo hoạt cảnh khuôn mặt sử dụng "blendshape" và thử nghiệm trên nền tảng Unity sử dụng ngôn ngữ lập trình C#.

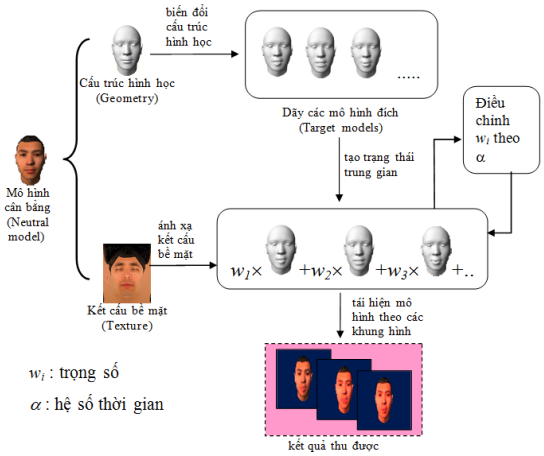
Bài báo này đề cập đến việc biến đổi mô hình đầu người 3D tại các thành phần trên khuôn mặt nhằm mục tiêu tạo ra những sự thay đổi trên mô hình sao cho phù hợp với lời thoại đầu vào trên cơ sở đã được phân tích ra thành các âm tiết. Để làm được điều này, chúng tôi đã xây dựng một tập các mô hình mẫu đầu người 3D và sử dụng nó làm cơ sở để trích các vector biến đổi mô hình theo các biểu diễn âm tiết riêng lẻ. Với một lời thoại đầu vào được đại diện bởi tập chuỗi các âm tiết có nhãn thời gian, mô hình sẽ được biến đổi một cách tương ứng có trọng số để tạo ra được hoạt cảnh đầu ra tương ứng để thể hiện cử động của mô hình theo lời thoại.

Phần còn lại của bài báo được cấu trúc như sau: phần 2 là các nội dung kỹ thuật đề xuất từ xây dựng mô hình đến các kỹ thuật biến đổi, phần 3 là thử nghiệm và cuối cùng là kết luận.

2. Kỹ thuật đề xuất

2.1. Sơ đồ chung

Chúng tôi đề xuất một bộ khung thực hiện gồm hai pha: pha chuẩn bị và pha xây dựng hoạt cảnh (Hình 1b).



Hình 2. *Lược đồ tạo hoạt ảnh trong chương trình*

Pha chuẩn bị: Đầu tiên thực hiện xây dựng tập các mô hình mẫu đầu người 3D. Tập này gồm một mô hình cân bằng và các mô hình biểu diễn trạng thái đầu người 3D theo mỗi âm tiết tương ứng. Với mỗi cặp: mô hình cân bằng và một mô hình ứng với một âm tiết, ta sẽ trích rút được một tập các vector biến đổi một mô hình cân bằng sang một mô hình biểu diễn âm tiết tương ứng. Mỗi tập vector biến đổi này sẽ được lưu trữ lại thành file như một cơ sở dữ liệu để phục vụ cho việc biến đổi các mô hình khác sau này.

Pha xây dựng hoạt cảnh (hình 2): Trước tiên, mô hình cân bằng sẽ được nạp vào. Sau đó, mô hình này được biến đổi để tạo ra một tập các mô hình theo các âm tiết của lời thoại dựa vào các tập vector biến đổi để có được một dãy chỉ số mô hình xuất hiện trong hoạt ảnh. Cuối cùng, sử dụng kỹ thuật morphing để tạo ra sự chuyển biến trạng thái mềm mại giữa các mô hình để cho ra một hoạt cảnh.

2.2. Tập mô hình đầu 3D

Các mô hình đầu 3D được xây dựng làm dữ liệu cơ sở cho việc biến đổi tương ứng với lời thoại. Tập mô hình này bao gồm mô hình cân bằng và các mô hình biểu diễn trạng thái đầu người 3D theo mỗi âm tiết tương ứng.

Mô hình cân bằng thể hiện trạng thái đầu người ở tình trạng trung tính, cụ thể định nghĩa như bBảng 1 dưới đây.

Bảng 1. *Bảng các trạng thái đầu người ở tình trạng trung tính*

Tên bộ phận	Trạng thái
Cơ mặt	Tất cả cơ mặt đều thư giãn
Mắt	Các mí mắt tiếp tuyến với tròng đen Đồng tử (con ngươi) bằng 1/3 đường kính của tròng đen
Miệng	Các môi khép lại Đường giữa hai môi nằm ngang và có cùng độ cao với khoé môi. Miệng đóng và răng trên chạm vào răng dưới Lưỡi bằng, nằm ngang với đầu lưỡi chạm vào đường biên giữa răng trên và răng dưới

Trong phạm vi nghiên cứu của bài báo, mô hình đầu người 3D được mô hình hóa dưới dạng là một tổng thể các bộ phận. Mỗi bộ phận là một đối tượng 3D, biểu diễn bởi một tập các điểm được phân bố theo trật tự nhất định, bao gồm một dạng hình học của đối tượng trong không gian 3 chiều, và một kết cấu bề mặt phủ lên bề mặt ngoài của dạng hình học đó. Cụ thể ở đây, dạng hình học của mỗi đối tượng là các lưới đa giác hình thành từ tập điểm trên. Khi xây dựng, mặt phẳng khuôn mặt song song với mặt phẳng Oxy và ánh mắt theo hướng trục Z. Các mô hình tương ứng với các âm tiết sẽ có cấu trúc dữ liệu giống với mô hình cân bằng nhưng có các trạng thái tọa độ các điểm trên bề mặt lưới khác biệt dựa theo mỗi biểu diễn khuôn mặt ứng với âm tiết.

2.3. Trích rút vector biến đổi theo âm tiết

Việc biến đổi mô hình từ trạng thái cân bằng sang trạng thái đích ứng với mỗi âm tiết thực tế là biến đổi các điểm trên mô hình theo mỗi vector tương ứng để tạo ra một biến thể của khuôn mặt. Để có được một mô hình đích mô phỏng một phát âm, những bộ phận sau trong mô hình sẽ được biến đổi: da mặt, hàm răng dưới, hàm răng trên, và lưỡi.

Giả sử đã có N là tập điểm của mô hình cân bằng và T tập điểm của mô hình đích.

$$N = \left(x_{11}, y_{11}, z_{11}, \dots x_{mn_m}, y_{mn_m}, z_{mn_m}\right) \in R^{3\times m\times n} \quad (1)$$

$$T = \left(x_{11}, y_{11}, z_{11}, \dots x_{mn_m}, y_{mn_m}, z_{mn_m}\right) \in R^{3\times m\times n} \quad (2)$$

Ta cần tính ra tập vector V để biến đổi từ N sang T, cùng với tập chỉ số I tương ứng. Công thức tính như sau:

$$\begin{cases} V_{ik} = T_{ij} - N_{ij} \\ I_{ik} = j \end{cases} với \begin{cases} j \in [1, n_i] / N_{ij} \neq T_{ij} \\ k \in [1, p_i] \end{cases} \quad (3)$$

Trong đó:

- m: số bộ phận của mô hình chịu sự biến đổi.
- ni: số điểm trên bộ phận thứ i.
- pi: số vector biến đổi của bộ phận thứ i.

Dưới đây là thuật toán để tính vector biến đổi cho một phép biến đổi từ mô hình cân bằng sang một mô hình ứng với một âm tiết nào đó.

Thuật toán trích rút vector biến đổi:

Input:

- NeutralModel: Mô hình cân bằng.
- TargetModel: Mô hình được biến đổi theo âm tiết.

Output:

- vector: Tập vector biến đổi theo một âm tiết nào đó.
- vec_index: Tập các chỉ số của điểm trên mô hình có vector biến đổi tương ứng trong tập vector.

Các bước thực hiện:

```
For each part ∈(Skin, LowerTeeth, UpperTeeth, Tongue)
    Begin
        {Tính tập vector biến đổi cho bộ phận part}
        num_of_vectorpart  = 0
        For i = 1 to num_of_point(part)
        Begin
            {Tính vector biến đổi tại điểm i }
            If (xi(NeutralModel)  ≠ xi(TargetModel)and
                    yi(NeutralModel)  ≠ yi(TargetModel)and
                            zi(NeutralModel)  ≠ zi(TargetModel))
            Begin
                vectori = (xi(TargetModel) - xi(NeutralModel),
                            yi(TargetModel) - yi(NeutralModel) ,
                        zi(TargetModel) - zi(NeutralModel))
                vec_indexnum_of_vector   = i
                increase(num_of_vector)
            End
        End
    End
```

Tập các vector này sẽ được lưu thành file. Với mỗi âm tiết, ta sẽ có một file chứa một tập vector biến đổi cho âm tiết tương ứng. Các file vector biến đổi sẽ được lưu theo cấu trúc như sau:

n_1

$index_{11}$ x_{11} y_{11} z_{11}

$index_{12}$ x_{12} y_{12} z_{12}

……

$index_{1n}$ x_{1n} y_{1n} z_{1n}

…….

…….

nm

$index_{m1}$ x_{m1} y_{m1} z_{m1}

$index_{m2}$ x_{12} y_{m2} z_{m2}

…….

$index_{mn}$ x_{mn} y_{mn} z_{mn}

Trong đó :

- m: là số bộ phận trên mô hình chịu sự biến đổi.
- ni: số vector biến đổi cho bộ phận thứ i.
- $index_{ij}$: chỉ số của điểm trên bộ phận thứ i của mô hình ứng với vector biến đổi thứ j.

2.4. Biến đổi mô hình theo âm tiết

Khi đã có tập các file vector biến đổi, ta có thể sử dụng để biến đổi mọi mô hình cùng dạng với mô hình mẫu theo các âm tiết tương ứng. Quá trình biến đổi mô hình theo một âm tiết được thực hiện như sau.

Đầu tiên, ta nạp vào mô hình cân bằng cần biến đổi. Sau đó, đọc các vector từ file lưu trữ tập vector cho âm tiết tương ứng. Tiếp theo, các bộ phận Da mặt, Hàm răng dưới, Hàm răng trên và Lưỡi của mô hình đích sẽ được biến đổi bằng các dịch chuyển các điểm theo các vector biến đổi tương ứng. Lưu ý là chỉ có một số điểm được lưu trữ vector biến đổi mới được dịch chuyển.

Bây giờ, giả sử ta đã có N là tập điểm của mô hình cân bằng, V tập vector biến đổi, và I là tập chỉ số vector.

$$N = \left(x_{11}, y_{11}, z_{11}, \dots x_{mn_m}, y_{mn_m}, z_{mn_m}\right) \in R^{3\times m\times n} \quad (4)$$

$$T = \left(x_{11}, y_{11}, z_{11}, \dots x_{mn_m}, y_{mn_m}, z_{mn_m}\right) \in R^{3\times m\times n} \quad (5)$$

$$I = \{k \in [1, n_i]/N_{ik} \neq T_{ik}\} \quad (6)$$

Trong đó:

- m: số bộ phận của mô hình chịu sự biến đổi.
- n_i: Số điểm trên bộ phận thứ i.

Ta cần tính ra tập điểm T của mô hình đích. Công thức tính như sau:

$$\begin{cases} T_{ij} = N_{ij} - N_{ik} \\ j = I_{ik} \end{cases} với \begin{cases} i \in [1, m] \\ k \in [1, n_i]/N_{ik} \neq T_{ik} \\ j \in [1, n_i] \end{cases} \quad (7)$$

Dưới đây là thuật toán để biến đổi mô hình theo mỗi âm tiết.

Thuật toán biến đổi mô hình:

Input:

- NeutralModel: Mô hình cân bằng.
- vector: Tập vector biến đổi theo một âm tiết nào đó.
- vec_index: Tập các chỉ số của điểm trên mô hình có vector biến đổi tương ứng trong tập vector.

Output:

- TargetModel: Mô hình theo âm tiết nào đó.

Các bước thực hiện:

```
For each part ∈(Skin, LowerTeeth, UpperTeeth, Tongue)
Begin
    {Biến đổi bộ phận part}
    For i = 1 to num_of_vector(part)
    Begin
        {Dịch chuyển điểm trên mô hình ứng với vector biến đổi}
        pointvec_index (i) (TargetModel)= pointvec_index(i)(NeutralModel) + vectori
    End
End
```

2.5. Điều chỉnh trọng số theo thời gian

Khi thực hiện phép biến đổi theo dãy n hình S_1.. S_n (S_i là một đối tượng hình tương ứng bộ phận của mô hình), thực chất là một chuỗi các phép biến đổi liên tục cho từng cặp hình (S_1,S_2), (S_2,S_3), ..., (S_i,S_{i+1}), (S_{i+1},S_{i+2}), ..., (S_{n-1},S_n).

Mỗi hình S_i có một trọng số wi kèm theo. Trọng số này sẽ được thay đổi theo thời gian. Ví dụ, khi quá trình morphing tiến đến phép morphing giữa cặp mô hình S_i và S_{i+1}, thì các w sẽ được điều chỉnh theo công thức (8) dưới đây.

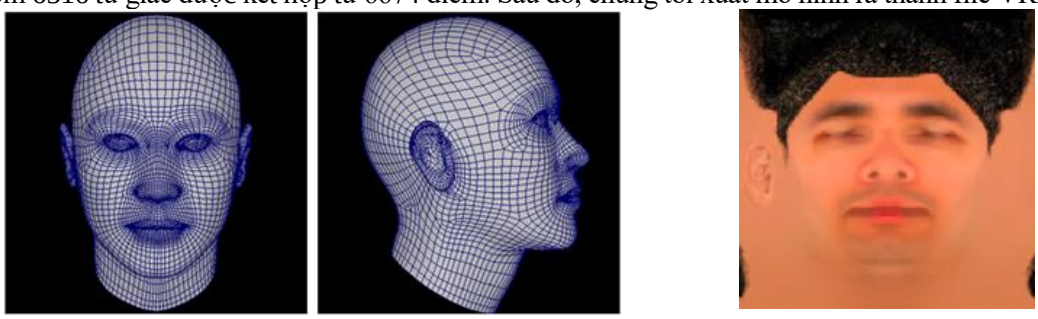
$$w_{j=1..n}\begin{cases} 1-\alpha, j=i \\ \alpha, j=i+1 \\ 0, j\neq i, j\neq i+1 \end{cases} \quad (8)$$

Trong đó, α∈[0,1] là một hệ số biến đổi theo thời gian giữa mỗi cặp khung hình cơ bản.

3. Thử nghiệm và đánh giá kết quả

Kỹ thuật được cài đặt thử nghiệm bằng ngôn ngữ lập trình Java 3D trên nền DirectX. Đồng thời kết hợp sử dụng ngôn ngữ mô hình hoá thực tại ảo VRML để lưu trữ các mô hình (Hình 4).

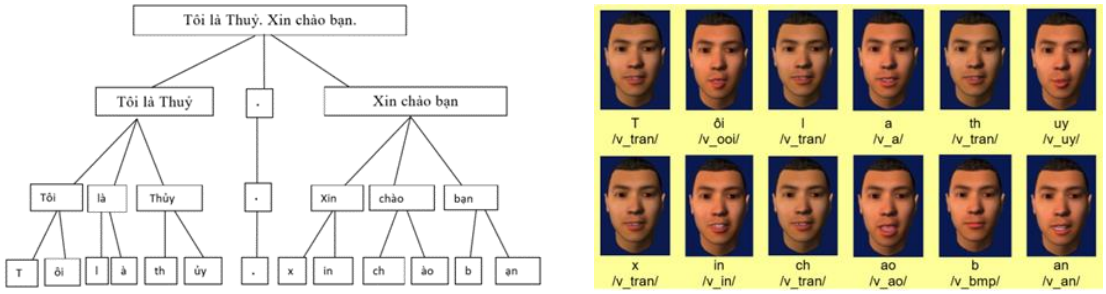
Mô hình của chúng tôi được xây dựng bằng phần mềm FaceGen Modeller, có các lưới đa giác gồm 6516 tứ giác được kết hợp từ 6674 điểm. Sau đó, chúng tôi xuất mô hình ra thành file VRML.



(a) *(b)*

Hình 3. *(a) Cấu trúc hình học của mô hình, (b) Kết cấu bề mặt của bộ phận da mặt*

Với một đoạn văn bản lời thoại đầu vào, nội dung sẽ được phân tích theo nhiều cấp dưới dạng cấu trúc cây. Đoạn văn sẽ được phân tích ra từng câu, từng tiếng và từng âm tiết. Hình 5 thể hiện quá trình thực hiện với câu "Tôi là Thủy. Xin chào bạn.".



(a) *(b)*

Hình 5. *(a) Cấu trúc cây phân tích văn bản, (b) Ánh xạ âm tiết với mô hình*

Từ những nội dung kỹ thuật được triển khai, chúng tôi tiến hành xây dựng một hệ thống phần mềm hoàn chỉnh mô phỏng cử động của mô hình đầu người 3D theo lời thoại tiếng Việt. Trên cơ sở những mô hình đã được xây dựng trước và được tải sẵn sàng vào dữ liệu của hệ thống lúc khởi động, phần mềm sẽ thực hiện từng bước phân tích những đoạn văn bản lời thoại đầu vào để trên cơ sở đó ánh xạ tương ứng những âm tiết với những thông tin đặc trưng trong cơ sở dữ liệu của

hệ thống phần mềm. Từ đó, phần mềm thực hiện nội suy và biểu diễn những hình ảnh đầu người theo chuỗi các âm tiết đã được phân tích tương ứng.

(a) *(b)*

Hình 6. *(a) Giao diện chương trình sau khi tải mô hình đầu người 3D và nhập văn bản, (b) Giao diện chương trình khi thực hiện mô phỏng biểu cảm theo lời thoại*

Hình 6 mô tả giao diện chương trình phần mềm mô phỏng cử động của mô hình đầu người 3D theo lời thoại tiếng Việt. Phần mềm đã xây dựng hoàn chỉnh các giao diện điều khiển tương ứng với các bước thực hiện của nghiên cứu đề xuất, từ bước nhập văn bản tiếng Việt, phân tích ra các âm tiết với nhãn thời gian, ánh xạ ra các mô hình cơ sở và thực hiện nội suy dữ liệu hoạt cảnh đầu người 3D. Phần mềm được xây dựng với ngôn ngữ Java trên công cụ phát triển Netbeans. Các hệ thống điều khiển được xây dựng chi tiết với từng chức năng tương ứng với các bước chính trong nghiên cứu đề xuất. Nội dung lời thoại có thể được nhập trực tiếp trên khung văn bản của cửa sổ cho phép thử nghiệm với những câu thoại tiếng Việt bất kỳ. Sau đó, nội dung lời thoại sẽ được phân tích và ánh xạ tương ứng với nhãn thời gian. Thông tin âm tiết cùng với nhãn thời gian sẽ được cập nhật trực tiếp trên khung văn bản của phần mềm. Đồng thời, để quan sát được quá trình mô phỏng, người dùng sẽ sử dụng nút bấm ngay bên dưới khung hiển thị mô hình đầu người.

4. Kết luận

Việc mô phỏng cử động của mô hình đầu người 3D theo lời thoại là một vấn đề quan trọng trong việc tạo ra bản sao kỹ thuật số con người ảo tương ứng tái tạo lại con người thật với những nội dung đa phương tiện. Bài báo đã trình bày một quy trình để thực hiện việc mô phỏng cử động của mô hình đầu người 3D theo lời thoại bắt đầu từ những yêu cầu cho việc thiết kế mô hình đầu 3D, việc tính các đặc trưng chuyển động theo lời thoại đến các kỹ thuật nội suy mô hình phục vụ sinh hoạt cảnh theo thời gian. Điều này đã đáp ứng được nhu cầu xây dựng bản sao kỹ thuật số con người ảo tương ứng tái tạo lại con người thật với những nội dung đa phương tiện trong nhiều trường hợp ví dụ như phát thanh viên hay những vị trí chuyên trả lời giải đáp các câu hỏi của khách hàng. Những trường hợp đối tượng như vậy không yêu cầu có nhiều những cử động của cơ thể, tay chân mà chủ yếu là những cử động, biểu cảm, thần thái của khuôn mặt khi trình bày. Trong thời gian sắp tới, nhóm nghiên cứu sẽ tiếp tục mở rộng dựa trên những nội dung đã thực hiện được để hoàn thiện hơn việc mô phỏng con người trong nhiều lĩnh vực ứng dụng khác nhau, chẳng hạn kết hợp mô phỏng sự biến đổi của thanh quản, phối vào trong những đối tượng ca sĩ ảo ứng dụng trong thanh nhạc…

Lời cảm ơn

Nghiên cứu này được tài trợ bởi đề tài NVCC02.02/23-23 “Hỗ trợ hoạt động nghiên cứu khoa học cho nghiên cứu viên cao cấp năm 2023, NVCC02.02/23-23”.

TÀI LIỆU THAM KHẢO/REFERENCES

[1] L. Chen, R. K. Maddox, Z. Duan, and C. Xu, "Hierarchical cross-modal talking face generation with dynamic pixel-wise loss," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 7832-7841.

[2] H. Zhou, Y. Liu, Z. Liu, P. Luo, and X. Wang, "Talking face generation by adversarially disentangled audio-visual representation," in *Proceedings of the AAAI conference on artificial intelligence*, 2019, vol. 33, no. 01, pp. 9299-9306.

[3] Y. Song, J. Zhu, D. Li, A. Wang, and H. Qi, "Talking face generation by conditional recurrent adversarial network," in *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, 2019, pp. 919-925.

[4] V. Konstantinos, P. Stavros, and P. Maja, "Realistic speech-driven facial animation with gans," *International Journal of Computer Vision*, vol. 128, pp. 1398-1413, 2020.

[5] K. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, and C. Jawahar, "A lip sync expert is all you need for speech to lip generation in the wild," in *Proceedings of the 28th ACM international conference on multimedia*, 2020, pp. 484-492.

[6] Y. Zhou, X. Han, E. Shechtman, J. Echevarria, E. Kalogerakis, and D.Li, "Makelttalk: speaker-aware talking-head animation," *ACM Transactions On Graphics*, vol. 39, no. 6, pp. 1-15, 2020.

[7] Y. Ding, C. Pelachaud, and T. Artieres, "Modeling multimodal behaviors from speech prosody," in *Proceedings of 13th International Conference on Intelligent Virtual Agents*, Springer, 2013, pp. 217-228.

[8] L. Chen, C. Guofeng, L. Celong, L. Zhong, K. Ziyi, X. Yi, and X. Chenliang, "Talking-head generation with rhythmic head motion," in *European Conference on Computer Vision*, Springer, 2020, pp. 35-51.

[9] M. Fratarcangeli and M. Schaerf, "Realistic modeling of animatable faces in MPEG-4," *in Proceedings of 17th Annual Conference on Computer Animation and Social Agents*, 2004, pp. 285-297.

[10] L. Turban, D. Girard, N. Kose, and J.-L. Dugelay, "From Kinect video to realistic and animatable MPEG-4 face model: A complete framework," in *2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, 2015, pp. 1-6.

[11] I. R. Ali, H. Kolivand, and M. H. Alkawaz, "Lip syncing method for realistic expressive 3D face model," *Multimedia Tools Applications*, vol. 77, pp. 5323-5366, 2018.

[12] Y. Zhao, D. Jiang, and H. Sahli, "3D emotional facial animation synthesis with factored conditional Restricted Boltzmann Machines," in *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2015, pp. 797-803.

[13] S. Dahmani, V. Colotte, V. Girard, and S. Ouni, "Conditional variational auto-encoder for text-driven expressive audiovisual speech synthesis," in *INTERSPEECH 2019-20th Annual Conference of the International Speech Communication Association*, 2019, pp. 2598-2602.

[14] M. Liu, Y. Duan, R. A. Ince, C. Chen, O. G. Garrod, P. G. Schyns, and R. E. Jack, "Building a generative space of facial expressions of emotions using psychological data-driven methods," in *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*, 2020, pp. 1-3.

[15] S. Wang, L. Li, Y. Ding, C. Fan, and X. Yu, "Audio2Head: Audio-driven One-shot Talking-head Generation with Natural Head Motion," *arXiv e-prints*, no. arXiv: 2107.09293, 2021.

[16] C. D. Thi and T. L. Son, "3D character expression animation according to Vietnamese sentence semantics," *TNU Journal of Science and Technology*, vol. 227, no. 16, pp. 20 - 28, 2022.