3D CHARACTER EXPRESSION ANIMATION ACCORDING TO VIETNAMESE SENTENCE SEMANTICS

Do Thi Chi*, Le Son Thai
TNU - University of Information and Communication Technology

ARTICLE INFO

Received: 13/8/2022
Revised: 07/10/2022
Published: 07/10/2022

KEYWORDS

Facial expression
Emotion
Lipsync
Animation
Virtual reality

ABSTRACT

Facial expressions are the main means of communicating social information between people, being a form of nonverbal communication. In a virtual reality program or game, a compelling 3D character needs to be able to act and express emotions clearly and coherently. Animation studies show that character need to represent at least six basic emotions: happy, sad, fear, disgust, anger, surprise. However, generating expression animations for virtual characters is time-consuming and requires a lot of creativity. The main objective of the article is to generate expression animations combined with lip-synchronization of 3D characters according to the semantics of Vietnamese sentences. Our method is based on the blendshape weights of the 3D face model. The input text after emotion prediction will be passed to lip sync and emotion generator to perform 3D face animation. Experimental results with 200 Vietnamese sentences are automatically classified according to six different emotions. Then, conduct a survey that predicts the emotion shown in the video. Survey participants were asked to recognize the emotions of 3D virtual faces according to each sentence of input text. Survey results show that anger is the most recognizable emotion, happiness and excitement are easily confused.

DIỄN HOẠT BIỂU CẢM NHÂN VẬT 3D THEO NGỮ NGHĨA CÂU TIẾNG VIỆT

Đỗ Thị Chi*, Lê Sơn Thái
Trường Đại học Công nghệ Thông tin và Truyền thông – ĐH Thái Nguyên

THÔNG TIN BÀI BÁO

Ngày nhận bài: 13/8/2022
Ngày hoàn thiện: 07/10/2022
Ngày đăng: 07/10/2022

TỪ KHÓA

Biểu cảm mặt
Cảm xúc
Đồng bộ môi
Diễn hoạt
Thực tại ảo

TÓM TẮT

Biểu cảm khuôn mặt là một hình thức giao tiếp phi ngôn ngữ, đây là phương tiện chính để truyền đạt thông tin giữa con người. Trong chương trình thực tại ảo hoặc trò chơi, một nhân vật 3D hấp dẫn cần có khả năng diễn xuất và thể hiện biểu cảm một cách rõ ràng, mạch lạc. Các nghiên cứu về diễn hoạt chỉ ra rằng nhân vật cần biểu diễn tối thiểu được sáu cảm xúc cơ bản: hạnh phúc, buồn, sợ hãi, chán ghét, tức giận, ngạc nhiên. Tuy nhiên, việc tạo diễn hoạt biểu cảm cho nhân vật ảo tốn nhiều thời gian và đòi hỏi sự sáng tạo cao. Với mục tiêu tạo diễn hoạt biểu cảm kết hợp với đồng bộ môi một cách tự động cho nhân vật 3D theo ngữ nghĩa câu tiếng Việt, bài báo dựa trên các trọng số blendshape của mô hình mặt 3D. Văn bản đầu vào sau khi dự đoán cảm xúc sẽ được chuyển đến bộ đồng bộ môi và tạo cảm xúc để thực hiện diễn hoạt mặt 3D. Kết quả thực nghiệm với 200 câu tiếng Việt được phân loại quan điểm tự động theo sáu cảm xúc khác nhau. Sau đó, tiến hành cuộc khảo sát để dự đoán biểu cảm mặt của nhân vật 3D. Người tham gia khảo sát được yêu cầu nhận biết cảm xúc khuôn mặt ảo 3D được tạo ra theo từng câu văn bản đầu vào. Kết quả khảo sát cho thấy, tức giận là cảm xúc dễ nhận biết nhất, hạnh phúc và vui mừng dễ bị nhầm lẫn.

DOI: https://doi.org/10.34238/tnu-jst.6361

* Corresponding author. *Email: dtchi@ictu.edu.vn*

1. Giới thiệu

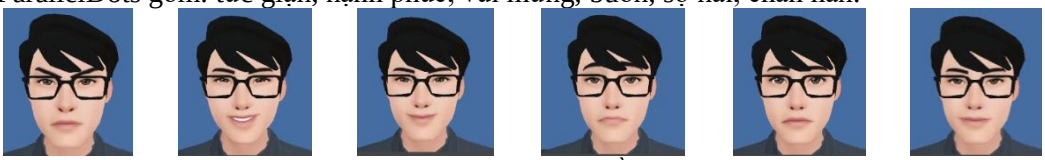
1.1. Đặt vấn đề

Ngày nay, nhân vật ảo 3D đóng vai trò quan trọng trong các ứng dụng tương tác giữa người với máy tính, trò chơi giải trí. Hình ảnh được kết xuất từ nhân vật 3D giống người thật giúp nhận thức thông tin một cách tự nhiên, trong đó các tín hiệu cảm xúc khác nhau truyền đạt các thông tin khác nhau. Biểu cảm trên khuôn mặt được sử dụng để truyền đạt nhiều ý nghĩa khác nhau trong các ngữ cảnh khác nhau, phạm vi và ý nghĩa được thể hiện từ những cảm xúc cơ bản đến phức tạp. Cảm xúc trên khuôn mặt để truyền đạt thông tin qua các bộ phận khác nhau như chuyển động của mắt, lông mày, môi, mũi v.v.. Đồng thời, những biểu cảm này kết hợp với khẩu hình khi nói tạo thành một thông điệp hoàn chỉnh trong giao tiếp.

Năm 1992, Ekman [1] đã phân loại cảm xúc thành bảy loại khác nhau: tức giận, hạnh phúc, sợ hãi, ngạc nhiên, buồn bã, ghê tởm, khinh bỉ. Sau đó, Ekman đã thêm các cảm xúc khác: thích thú, mãn nguyện, xấu hổ, tội lỗi, bối rối v.v… vào danh sách các biểu cảm của ông. Tuy nhiên, trong những nghiên cứu tiếp theo, Ekman đã nhận định sáu loại cảm xúc cơ bản: tức giận, hạnh phúc, ngạc nhiên, buồn bã, sợ hãi, ghê tởm là cơ sở cho các biểu cảm trên khuôn mặt con người. Những kết quả nghiên cứu của Ekman xoay quanh việc phát hiện ra rằng sáu cảm xúc cơ bản dường như được công nhận trên toàn cầu, ngay cả trong các nền văn hóa mà con người không thể học được biểu cảm khuôn mặt thông qua phương tiện truyền thông. Tuy nhiên, các nghiên cứu tiếp theo đã chỉ ra có nhiều cảm xúc cơ bản hơn. Năm 2017, Alan S.Cowen và Dacher Keltner [2] đã xác định được 27 loại cảm xúc khác nhau của con người. Nhóm nghiên cứu chỉ ra rằng, con người trải nghiệm cảm xúc theo một phổ và có sự pha trộn nhất định chứ không cảm nhận riêng biệt từng loại. Dựa trên các công trình nghiên cứu biểu cảm của con người, trong nội dung bài báo sử dụng bộ sáu cảm xúc cơ bản để áp dụng phân loại lớp cảm xúc cho câu văn bản tiếng Việt.

Thực tế, một câu nói được truyền tải sẽ đi kèm với một cảm xúc nhất định, sự thể hiện cảm xúc này chỉ tồn tại với con người và các loài động vật. Để máy tính có thể hiểu được cảm xúc ẩn ý sau câu nói, bài báo sử dụng thông tin ngữ nghĩa của câu văn bản đầu vào để phân loại cảm xúc [3]. Với một câu văn bản đầu vào, bộ phân tích ngữ nghĩa sẽ tìm kiếm các từ khoá có cường độ cảm xúc cao và đánh giá các trọng số của từ khoá để dự đoán cảm xúc cho văn bản [4].

Hiện nay, có nhiều công cụ phân tích cảm xúc của con người. Nghiên cứu này sử dụng bộ phân tích cảm xúc của ParallelDots [5] do có thể phát hiện chính xác cảm xúc của văn bản đầu vào; xử lý và trả kết quả trong thời gian ngắn; đảm bảo yếu tố bảo mật dữ liệu. Dựa trên những ưu điểm của bộ phân tích cảm xúc của ParallelDots, bài báo lựa chọn sử dụng công cụ phân tích cảm xúc này. Trong đó, cảm xúc chán nản thay cho ghê tởm, vui mừng thay cho ngạc nhiên. Hình 1 thể hiện sáu loại cảm xúc được bài báo sử dụng tương ứng với thang đo năng lực cảm xúc của ParallelDots gồm: tức giận, hạnh phúc, vui mừng, buồn, sợ hãi, chán nản.



Hình 1. *Sáu cảm xúc cơ bản của khuôn mặt được chọn để diễn hoạt*

Cho đến nay, có nhiều cách tiếp cận khác nhau để tạo biểu cảm mặt người. Phần lớn các nghiên cứu tiếp cận theo hướng điều khiển mô hình dựa trên nhận thức văn hoá chủ quan của con người [6] – [7]. Tuy nhiên, khả năng nhận biết chính xác các biểu cảm trên khuôn mặt nói chung còn thấp. Do đó, trong bài báo này đề xuất một mô hình tạo biểu cảm mặt người tự động trên cơ sở phân tích ngữ nghĩa của tiếng Việt. Phần tiếp theo của bài báo trình bày tổng quan một số kỹ thuật tạo biểu cảm mặt cho mô hình nhân vật 3D. Tiếp theo, trình bày chi tiết kỹ thuật tạo biểu cảm khuôn mặt và đồng bộ môi nhân vật 3D theo ngữ nghĩa tiếng Việt. Cuối cùng là kết quả thực nghiệm.

1.2. Các nghiên cứu liên quan

Một số nghiên cứu gần đây [8], [9], [10], [11] đã chỉ rõ một số phương pháp tạo biểu cảm mặt với những cảm xúc cơ bản. Nicolas [12] đã đưa ra những nghiên cứu đột phá trong lĩnh vực tạo biểu cảm khuôn mặt. Nhóm nghiên cứu đã giới thiệu phương pháp tạo biểu cảm để tương tác giữa người và máy tính dựa trên các tham số định nghĩa khuôn mặt MPEG-4 (FDPs) và các tham số diễn hoạt mặt (FAPs). Ưu điểm chính của phương pháp này là khả năng phân tích và tổng hợp biểu cảm linh hoạt.

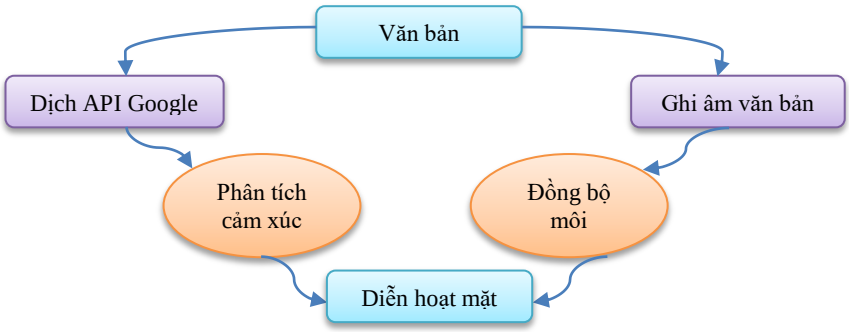
Tang [10] đã tạo nhân vật ảo nói có cảm xúc (EAVA) bằng cách chuyển văn bản thành giọng nói cảm xúc thông qua kỹ thuật đồng bộ giọng nói cảm xúc. Nghiên cứu tập trung vào dựng mô hình mặt 3D, diễn hoạt biểu cảm mặt, đồng bộ giọng nói cảm xúc, kết hợp chuyển động của môi và biểu cảm khuôn mặt. Để tìm ra sự khác nhau giữa các trạng thái cảm xúc, nhóm nghiên cứu đã sử dụng các quy tắc để xác định các thông số giữa cảm xúc và ngữ điệu trung lập. Với kết quả đánh giá chủ quan, cảm xúc tiêu cực (buồn, sợ hãi, tức giận) được đồng bộ tốt hơn cảm xúc tích cực (hạnh phúc).

Dahmani [8] sử dụng bộ mã hoá tự động biến thể (VAE) để đồng bộ biểu cảm với âm thanh giọng nói. Kết quả thực nghiệm cho thấy tỷ lệ nhận dạng tốt với hầu hết các cảm xúc. Tuy nhiên, khó phân biệt giữa cảm xúc buồn và sợ hãi.

Liu [13] xây dựng mô hình tạo biểu cảm theo phương pháp tiếp cận tâm sinh lý [14], [15] dựa trên nhận thức của con người. Quá trình tạo biểu cảm gồm hai bước chính: Định nghĩa các chuyển động của khuôn mặt dưới dạng đơn vị hành động (Aus) [16] để xây dựng một không gian dữ liệu; Tạo một loạt các biểu cảm mặt tương ứng với sáu cảm xúc cơ bản và những cảm xúc phức tạp thông qua tương quan chéo. Tận dụng khả năng dự đoán của một tập lớn các biểu cảm trong số 25 cảm xúc cơ bản và phức tạp vào không gian dữ liệu, nhóm nghiên cứu đã chứng minh được khả năng tạo biểu cảm cho nhân vật ảo.

Phần lớn các nghiên cứu diễn hoạt biểu cảm khuôn mặt được thực hiện theo mô hình điều khiển dữ liệu dựa trên nhận thức văn hoá chủ quan của con người. Kết quả nghiên cứu của các phương pháp trên cho thấy khả năng nhận biết các biểu cảm trên khuôn mặt được tạo ra chưa đồng nhất hoàn toàn. Hai mô hình tiếp cận gần đây nhất là VAE [8] và DBLSTM [17], tuy nhiên kỹ thuật tính toán phức tạp và có nhiều thách thức trong quá trình thực hiện. Ngược lại, phương pháp của chúng tôi dễ thực hiện do quy trình thiết lập các cấu hình blendshape để tổng hợp biểu cảm khuôn mặt được thực hiện một cách tự động, do đó quá trình thực hiện diễn hoạt cảm xúc khuôn mặt trên nhân vật ảo tính toán ít hơn và loại bỏ yếu tố tác nhân chủ quan của con người.

2. Phương pháp nghiên cứu



Hình 2. *Quy trình tạo diễn hoạt biểu cảm khuôn mặt với dữ liệu đầu vào, các bước trung gian và kết quả đầu ra*

Phần này trình bày chi tiết phương pháp tạo diễn hoạt biểu cảm nhân vật ảo dựa trên phân tích ngữ nghĩa câu tiếng Việt được thể hiện trong Hình 2. Đầu vào là văn bản tiếng Việt. Bài báo sử dụng bộ phân tích ngữ nghĩa tự động của văn bản để dự đoán cảm xúc cho khuôn mặt. Đầu tiên,

câu văn bản tiếng Việt được chuyển sang âm thanh, đồng thời được dịch sang tiếng Anh thông qua bộ dịch thuật API Google. Sau đó, văn bản dịch được chuyển đến bộ phân tích ngữ nghĩa API ParallelDots để xác định cảm xúc của văn bản. Tiến hành tạo diễn hoạt môi cho nhân vật dựa trên âm thanh được ghi trước đó. Cuối cùng, kết hợp diễn hoạt môi với cảm xúc đã được dự đoán để tạo ra diễn hoạt cho khuôn mặt.

2.1. Phân tích ngữ nghĩa

Văn bản tiếng Việt ban đầu sau khi được dịch sang tiếng Anh qua bộ dịch API Google sẽ được chuyển sang bộ phân tích ngữ nghĩa. Bảng 1 là kết quả một số mẫu câu tiếng Việt sau khi được dịch sang tiếng Anh sử dụng API Google.

Bảng 1. *Một số mẫu câu văn bản đầu vào tiếng Việt*

Câu văn bản	Nội dung	Văn bản được dịch
A	Ai đã làm hỏng xe của tôi?	Who damaged my car?
B	Thật vui khi gặp lại bạn	Good to see you again
C	Xin lỗi, tôi không cố ý	Sorry, I didn't mean to
D	Tôi rất mong chờ lễ hội chè sắp tới	I'm looking forward to the upcoming tea festival
E	Bạn có chắc cây cầu này an toàn	Are you sure this bridge is safe?
F	Kỳ nghỉ lễ lần này thật dài và tẻ nhạt	This holiday season is long and boring

Bài báo sử dụng bộ phân tích cảm xúc API ParallelDots [5] để dự đoán cảm xúc cho văn bản. Mỗi câu văn bản đầu vào sẽ được phân thành các lớp cảm xúc khác nhau. Mỗi lớp cảm xúc thu được kết quả phân tích gồm hai thành phần: Nhãn cảm xúc và tỷ lệ phần trăm trọng số cảm xúc. Với những trọng số cảm xúc thu được, bài báo chọn lớp cảm xúc có giá trị phần trăm cao nhất làm cảm xúc cho câu văn bản, giá trị phần trăm được chọn sử dụng làm hệ số tỷ lệ cho các blendshape của nhân vật ảo khi tạo biểu cảm.

Ví dụ, với câu văn bản "Ai đã làm hỏng xe của tôi?". Câu văn bản thu được sau khi sử dụng bộ dịch Google: "Who damaged my car?". Tiếp theo, sử dụng bộ API ParallelDots để phân tích cảm xúc của câu văn bản dịch. Kết quả phân tích gồm: **Hạnh phúc**: 0,31%; **Tức giận**: 72,67%; **Vui mừng**: 0,78%; **Buồn**: 18,12%; **Sợ hãi**: 6,21%; **Chán nản**: 1,91%. Với kết quả dự đoán trên, bài báo chọn lọc tức giận làm cảm xúc đầu vào cho câu văn bản. Bảng 2 thể hiện kết quả dự đoán cảm xúc có trọng số cao nhất của một số câu tiếng Việt sau khi phân tích bởi API ParallelDots.

Bảng 2. *Tỷ lệ phần trăm trọng số cảm xúc cao nhất của câu văn bản đầu vào*

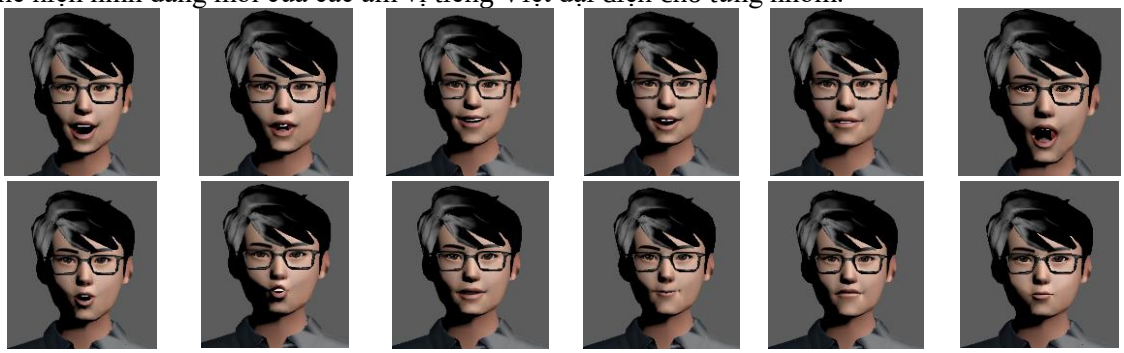
Câu văn bản	Cảm xúc	Tỷ lệ % cao nhất
A	Tức giận	72,67
B	Hạnh phúc	45,05
C	Buồn	41,06
D	Vui mừng	34,11
E	Sợ hãi	49,56
F	Chán nản	82,07

Bài báo sử dụng thư viện ứng dụng khách và SDK để tích hợp API Komprehend trong chương trình ứng dụng. Kết nối API ParallelDots với chương trình được thiết lập bằng cách sử dụng thư viện Komprehend. Để sử dụng dữ liệu trong Unity, bài báo tạo kết nối TCP giữa ứng dụng máy chủ và khách bằng ngôn ngữ C#.

2.2. Đồng bộ môi với giọng nói

Quá trình diễn hoạt môi dựa trên tệp âm thanh được ghi lại từ văn bản đầu vào. Các hình dáng môi được xây dựng dựa trên các chuyên gia về diễn hoạt. Với sự khác biệt của tiếng Việt so với các ngôn ngữ khác, chúng tôi xây dựng bộ dịch hoạt của tiếng Việt gồm: 12 nguyên âm đơn, 139

nguyên âm ghép, 17 phụ âm đơn và 11 phụ âm ghép. Trong 28 phụ âm, chúng tôi diễn hoạt 4 phụ âm b/m/p/ph/v vì chỉ có 5 phụ âm này có sự tác động nhiều của môi. Với mục tiêu tối ưu bộ dịch hoạt, các thanh điệu được sử dụng trong tiếng Việt được loại bỏ khi diễn hoạt. Điều này có ảnh hưởng tới hình ảnh khi diễn hoạt môi, nhưng sự ảnh hưởng này là không lớn. Bên cạnh đó, khi cố gắng xây dựng bộ dịch hoạt tiếng Việt với đầy đủ các dấu sẽ phát sinh sự bùng nổ tổ hợp. Với những âm vị có hình dáng chuyển động của môi giống nhau sẽ được gộp thành 1 nhóm. Hình 3 thể hiện hình dáng môi của các âm vị tiếng Việt đại diện cho từng nhóm.



Hình 3. *Hình đại diện cho các nhóm âm vị tiếng Việt*

Quá trình diễn hoạt một câu là sự kết hợp diễn hoạt từng từ đơn trong câu đó. Quá trình diễn hoạt một từ là sự kết hợp của việc chuyển trạng thái từ tiền âm vị tới hậu âm vị. Bài báo sử dụng kỹ thuật nội suy để tính toán quá trình chuyển đổi giữa các âm vị và các trạng thái giữa theo thời gian. Các âm vị được sắp xếp trên dòng thời gian, trạng thái diễn hoạt của nhân vật tại thời gian t là nội suy của 2 âm vị gần nhất, trước và sau thời gian t. Giả sử, thứ tự các âm vị là $p_0, p_1, \ldots, p_n$ diễn ra tại các thời điểm $t_0, t_1, \ldots, t_n$. Ta xác định được vị trí của t trên dòng thời gian với điều kiện: $t_i \le t \le t_{i+1}$. Trạng thái p_t mô tả hình dáng miệng tại thời gian t được xác định dựa trên hàm nội suy tuyến tính (1).

$$p_t = \frac{t - t_i}{t_{i+1} - t_i} p_{i+1} + \frac{t_{i+1} - t}{t_{i+1} - t_i} p_i \quad (1)$$

Tổng hợp bước này chúng ta thu được hình ảnh nhân vật 3D diễn hoạt nói tiếng Việt với đầu vào là tệp âm thanh đã được gán nhãn.

Với câu văn bản đầu vào sẽ thu được tệp diễn hoạt môi, mỗi tệp sẽ tương ứng với một cảm xúc được phân lớp trong sáu cảm xúc cơ bản. Tiếp tục tạo biểu cảm cho từng tệp với cảm xúc có trọng số cao nhất.

2.3. Tạo diễn hoạt biểu cảm khuôn mặt sử dụng blendshape

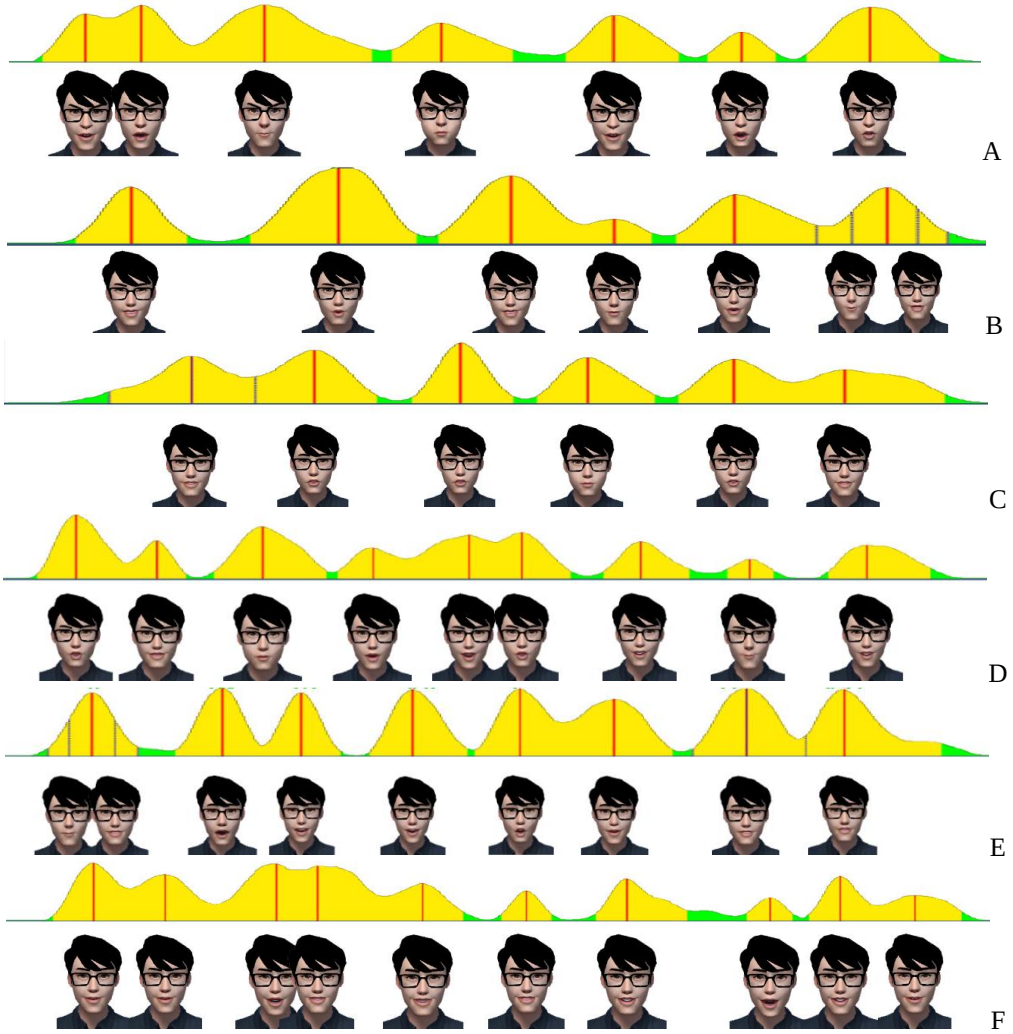
Sử dụng các tham số cấu hình blendshape để tạo sáu biểu cảm cơ bản trên khuôn mặt. Mỗi biểu cảm kiểm soát các bộ phận nhất định của khuôn mặt như mắt, lông mày, miệng. Các bộ phận được xác định bởi các thông số blendshape có giá trị từ 0 đến 1, sự thay đổi giá trị của các thông số blendshape sẽ tạo ra các cảm xúc khác nhau trên khuôn mặt. Ví dụ, khi tức giận thì lông mày nhíu lại, khi hạnh phúc thì miệng cười rộng. Bằng cách kết hợp các cấu hình blendshape trên khuôn mặt sẽ tạo ra được nhóm gồm sáu cảm xúc cơ bản được thể hiện trong Hình 1. Mô hình blendshape được xác định bởi công thức sau:

$$V = \sum_{i=1}^{n} e * V_i \quad (2)$$

Trong đó:

- V: một đỉnh của mô hình
- e: trọng số cảm xúc được dự đoán theo ParallelDots, $0 \le e \le 1$
- V_i: vị trí của đỉnh trong hình dạng mặt i

Kết quả tạo diễn hoạt biểu cảm kết hợp với đồng bộ môi của sáu câu tiếng Việt (Bảng 1) được thể hiện trong Hình 4. Mỗi câu được thể hiện bởi đồ thị dao động âm thanh và hình ảnh khuôn mặt được trích xuất tương ứng với vị trí âm vị của tệp âm thanh. Đường thẳng màu đỏ trên sóng âm thanh đại diện cho vị trí âm vị của từng từ trong tệp âm thanh.

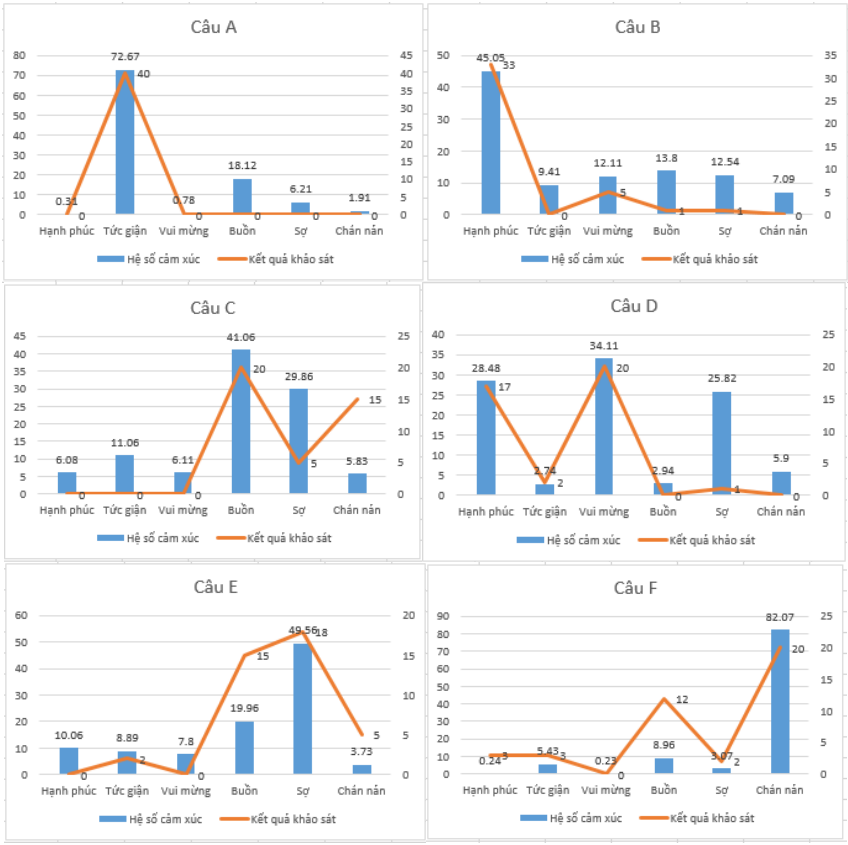


Hình 4. *Diễn hoạt biểu cảm của sáu câu văn bản ở Bảng 1 tương ứng với A, B, C, D, E, F*

3. Kết quả và bàn luận

Để đánh giá độ chính xác của khuôn mặt được tạo ra, chúng tôi tiến hành thực nghiệm trên tập dữ liệu gồm 200 câu văn bản tiếng Việt với các cảm xúc khác nhau. Quá trình khảo sát thực hiện với đối tượng là 20 người dùng mới và 20 chuyên gia diễn hoạt, mỗi người sẽ nhận dạng cảm xúc khuôn mặt nhân vật 3D khi diễn hoạt tập dữ liệu trên, bằng cách dự đoán cảm xúc của mỗi câu văn bản sẽ tương ứng với cảm xúc nào trong sáu cảm xúc cơ bản: Hạnh phúc, tức giận, vui mừng, buồn, sợ hãi, chán nản. Hình 5 là kết quả khảo sát của sáu câu văn bản được minh hoạ trong Bảng 1.

Kết quả khảo sát nhận dạng cho thấy, số lượng người nhận dạng cảm xúc đạt giá trị nhiều nhất trên từng văn bản trùng với kết quả dự đoán của API ParallelDots. Trong đó, hai cảm xúc Hạnh phúc và Tức giận là nhận biết dễ dàng và chính xác nhất. Trong khi đó, người dùng hay nhầm lẫn giữa Vui mừng với Hạnh phúc, Buồn với Sợ. Sự nhầm lẫn này là do yếu tố tương đồng về các đặc điểm nhân trắc học trên khuôn mặt. Kết quả khảo sát cũng cho thấy, khi hệ số cảm xúc theo ParallelDots càng cao thì kết quả nhận dạng cảm xúc của người tham gia khảo sát càng chính xác.



Hình 5. *Kết quả khảo sát nhận dạng cảm xúc với kỹ thuật đề xuất. Trong đó, biểu đồ cột là hệ số cảm xúc dự đoán theo API ParallelDots, biểu đồ đường là số lượng người nhận dạng từng cảm xúc của nhân vật ảo*

4. Kết luận

Bài báo trình bày kỹ thuật tạo diễn hoạt biểu cảm khuôn mặt kết hợp với đồng bộ môi một cách tự động cho nhân vật 3D theo ngữ nghĩa câu tiếng Việt. Bằng cách khai thác phân tích ngữ nghĩa câu văn bản để xác định lớp cảm xúc và trọng số cảm xúc. Các trọng số đó kết hợp với các cấu hình blendshape được dự đoán trước để tạo ra biểu cảm khuôn mặt cho câu văn bản. Sau đó, tiến hành cuộc khảo sát với 40 người tham gia để nhận biết cảm xúc được tạo ra của tập câu văn bản đầu vào sẽ thuộc cảm xúc nào trong sáu cảm xúc: hạnh phúc, tức giận, vui mừng, buồn, sợ, chán nản. Kết quả khảo sát cho thấy việc tạo diễn hoạt mặt không ảnh hưởng đáng kể đến việc nhận biết cảm xúc. Điều đó cho thấy, kỹ thuật đề xuất của bài báo là đáng tin cậy. Bài báo sử dụng bộ dịch API Google, do đó câu văn bản đầu vào còn bị phụ thuộc vào quá trình chuyển đổi ngôn ngữ của bộ dịch. Đồng thời, quá trình diễn hoạt tự động còn bị phụ thuộc vào bộ phân tích cảm xúc của API ParallelDots. Bên cạnh đó, quá trình sinh biểu cảm tự động bị ảnh hưởng bởi thiết kế khuôn mặt đầu vào của nhân vật dẫn đến các biểu cảm sinh ra còn chưa phong phú. Hướng phát triển cho nghiên cứu này là áp dụng với tập dữ liệu đầu vào lớn hơn và đa dạng các cảm xúc thay vì giới hạn trong sáu cảm xúc hiện tại.

Lời cám ơn

Bài báo được hoàn thành với sự trợ giúp của đề tài nghiên cứu với mã số: ĐH2022-TN07-01.

TÀI LIỆU THAM KHẢO/ REFERENCES

[1] P. Ekman, “Are there basic emotions?” *Psychological Review*, vol. 99, no. 3, pp. 550-553, 1992.

[2] A. S. Cowen and D. Keltner, "Self – report captures 27 distinct categories of emotion bridged by continuous gradients," *Psychological and Cognitive sciences*, vol. 114, no. 38, pp. E7900-E7909, 2017.
[3] T. Gungor and K. Celik, "A comprehensive analysis of using semantic information in text categorization," in *2013 IEEE INISTA*, 2013, pp. 1–5.
[4] C. Goddard, *Semantic analysis: A practical introduction*. Oxford University Press, 1998.
[5] ParallelDots, "Emotion analysis API," 2020. [Online]. Available: https://www.paralleldots.com/emotion-analysis. [Accessed March 30, 2021].
[6] C. Chen, C. Crivelli, O. G. B. Garrod, P. G. Schyns, J. -M. Fernandez-Dols, and R. E. Jack, "Distinct facial expressions represent pain and pleasure across cultures," *Proceedings of the National Academy of Sciences*, vol. 115, no. 43, pp. E10013–E10021, 2018.
[7] H. Yu, O. G. B. Garrod, and P. G. Schyns, "Perception-driven facial expression synthesis," *Computers & Graphics*, vol. 36, no. 3, pp. 152–162, 2012.
[8] S. Dahmani, V. Colotte, V. Girard, and S. Ouni, "Conditional variational auto-encoder for text-driven expressive audio visual speech synthesis," in *INTERSPEECH 2019-20th Annual Conference of the International Speech Communication Association*, Graz, Austria, 2019.
[9] T. Karras, T. Aila, S. Laine, A. Herva, and J. Lehtinen, "Audio-driven facial animation by joint end-to-end learning of pose and emotion," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, pp. 1– 12, 2017.
[10] H. Tang, Y. Fu, J. Tu, T. S. Huang, and M. Hasegawa-Johnson, "Eava: a 3d emotive audio-visual avatar," in *2008 IEEE Workshop on Applications of Computer Vision*, 2008, pp. 1–6.
[11] C. Chen, L. B. Hensel, Y. Duan, R. A. A. Ince, O. G. B. Garrod, J. Beskow, R. E. Jack, and P. G. Schyns, "Equipping social robots with culturally sensitive facial expressions of emotion using data-driven methods," in *14th IEEE International Conference on Automatic Face & Gesture Recognition*, 2019, pp. 1–8.
[12] N. Tsapatsoulis, A. Raouzaiou, S. D. Kollias, R. I. D. Cowie, and E. Douglas-Cowie, *Emotion recognition and synthesis based on mpeg-4 faps*, John Wiley and Sons, 2002.
[13] M. Liu, Y. Duan, R. A. A. Ince, C. Chen, O. G. B. 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.
[14] R. E. Jack and P. G. Schyns, "Toward a social psychophysics of face communication," *Annual Review of Psychology*, vol. 68, pp. 269–297, 2017.
[15] A. A. Jr and J. Lovell, "Stimulus features in signal detection," *The Journal of the Acoustical Society of America*, vol. 49, no. 6B, pp. 1751–1756, 1971.
[16] P. Ekman, "Measuring facial movement," *Environmental Psychology and Nonverbal Behavior*, vol. 1 pp. 56–75, 1976.
[17] X. Li, Z. Wu, H. M. Meng, J. Jia, X. Lou, and L. Cai, "Expressive speech driven talking avatar synthesis with dblstm using limited amount of emotional bimodal data," in *Interspeech*, San Francisco, USA, September 8–12, 2016, pp. 1477–1481.