ĐẠI HỌC QUỐC GIA TP. HCM

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

HOÀNG MINH THANH

OPENHUMAN: MÔ HÌNH TỔNG HỢP CỬ CHỈ HỘI THOẠI ĐA PHƯƠNG THỨC DỰA TRÊN VĂN BẢN VÀ ÂM THANH

LUẬN VĂN THẠC SĨ

TP. Hồ Chí Minh – Năm 2024

ĐẠI HỌC QUỐC GIA TP. HCM

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

HOÀNG MINH THANH

OPENHUMAN: MÔ HÌNH TỔNG HỢP CỬ CHỈ HỘI THOẠI ĐA PHƯƠNG THỨC DỰA TRÊN VĂN BẢN VÀ ÂM THANH

Ngành: Khoa học máy tính

Mã số Ngành: 8480101

NGƯỜI HƯỚNG DẪN KHOA HỌC

HDC: PSG. TS. LÝ QUỐC NGỌC

TP. Hồ Chí Minh – Năm 2024

LỜI CAM ĐOAN

Tôi cam đoan luận văn thạc sĩ ngành Khoa học máy tính, với đề tài OpenHuman: Mô hình tổng hợp cử chỉ hội thoại đa phương thức dựa trên văn bản và âm thanh là công trình khoa học do tôi thực hiện dưới sự hướng dẫn của TS. Lý Quốc Ngọc.

Những kết quả nghiên cứu của luận văn hoàn toàn trung thực và chính xác.

TP. Hồ Chí Minh, ngày... tháng... năm...

HỌC VIÊN THỰC HIỆN

(Ký và ghi rõ họ tên)

LỜI CẢM ƠN

Chúng tôi xin chân thành cảm ơn thầy Lý Quốc Ngọc đã tận tình hướng dẫn, truyền đạt kiến thức và kinh nghiệm, và đưa ra các giải pháp cho chúng tôi trong suốt quá trình thực hiện đề tài luận văn tốt nghiệp này.OpenHuman

Xin gửi lời cảm ơn đến quý thầy cô Khoa Công Nghệ Thông Tin trường Đại Học Khoa Học Tự Nhiên - Đại Học Quốc Gia Thành Phố Hồ Chí Minh, những người đã truyền đạt kiến thức quý báu cho chúng tôi trong thời gian học tập vừa qua.

Đồng thời cảm ơn các nhà khoa học đã nghiên cứu về đề tài mà chúng tôi đã trích dẫn để có thể có những kiến thức hoàn thiện luận văn của chúng tôi.

Sau cùng chúng tôi xin gửi lời cảm ơn đến gia đình, bạn bè,.. những người luôn động viên, giúp đỡ chúng tôi trong quá trình làm luận văn.

Một lần nữa, xin chân thành cảm ơn !

MỤC LỤC

Chương 1

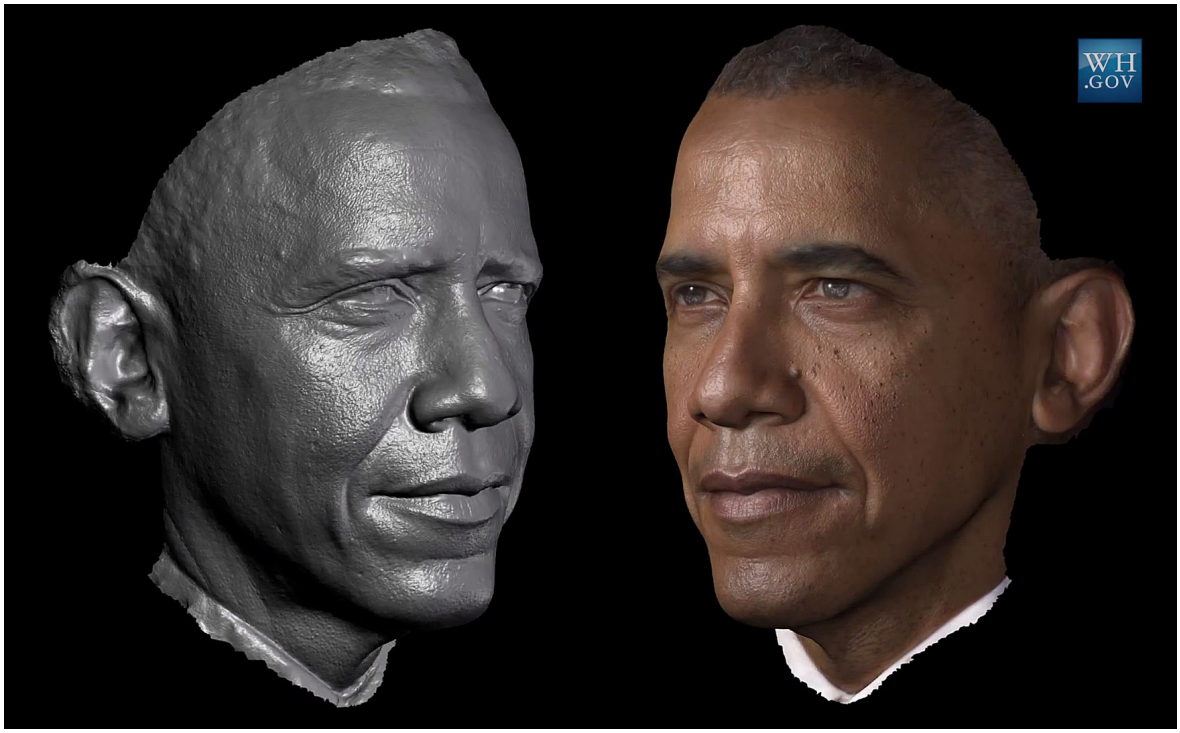
Giới thiệu

1 Bối cảnh chung

Hình 1: Công nghệ CGI với người kỹ thuật số siêu thật

Mỗi ngày, trên thế giới có hàng tỷ người nhìn vào màn hình RGB, kết quả hiển thị trên màn hình là đầu ra của mọi hệ thống phần mềm, nên việc hiển thị từng pixel trên màn hình và cách để mô phỏng lại hình ảnh trên một cách chân thực nhất được các nhà khoa học về đồ họa máy tính (Computer Graphic) nghiên cứu từ những năm 1960s và đặc biệt là việc mô phỏng lại con người Hình 1.

Ngày nay, công nghệ đồ họa máy tính đã hoàn toàn có thể mô phỏng nhiều vật



Hình 2: Minh họa công trình tái tạo khuôn mặt tổng thống Obama

giống đến mức siêu thực (realistic) các vật phức tạp như nước, đường xá, bánh mỳ,... và thậm chí là cả cơ thể và khuôn mặt con người với độ chi tiết đến từng lông tơ, nốt mụn và vân mắt. Vào năm 2015, bằng kỹ thuật quyét 3 chiều ghi lại toàn bộ các góc của khuôn mặt, sự phản chiếu ánh sáng, trong công trình [16] nhà khoa học đồ họa máy tính đã có thể tái tạo toàn bộ khuôn mặt của tổng thống Obama trên máy tính với độ chính xác cao và gần như không thể phân biệt Hình 2.

Trí tuệ nhân tạo thể hiện kết quả vượt bậc những năm gần đây không chỉ trong nghiên cứu mà con trong ứng dụng thực tế, tiêu biểu như ứng dụng ChatGPT, Midjouney và sự phát triển cả theo chiều dọc và chiều ngang trong việc ứng dụng trí tuệ nhân tạo với nhiều lĩnh vực khác nhau. Mặc dù đồ họa máy tính đã có thể xây dựng khuôn mặt người siêu thật, việc sinh cử chỉ lại phụ thuộc vào việc Chụp chuyển động (Motion Capture) từ các cảm biến và gặp rất nhiều khó khăn khi xây dựng một hệ thống trí tuệ nhân tạo để học từ dữ liệu.

Các hệ thống trí tuệ nhân tạo hiện nay đã có thể tạo văn bản và âm thanh tiệm cận như con người, nhưng một trong những trở ngại lớn nhất để xây dựng con người kỹ thuật số hiện nay chính là việc sinh cử chỉ. Chính vì vậy mà mục tiêu của luận văn là

xây dựng một hệ thống sinh cử chỉ hội thoại dựa trên cảm xúc và ngữ nghĩa với dữ liệu đầu vào gồm cả văn bản và giọng nói.

2 Động lực nghiên cứu

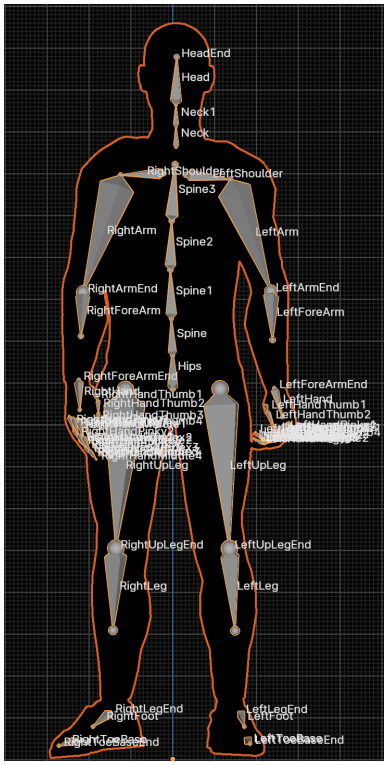
Tổng hợp cử chỉ hội thoại giúp ích cho rất nhiều lĩnh vực như hoạt ảnh, dựng phim, trò chơi điện tử, giáo dục và những ứng dụng thực tại ảo. Việc tổng hợp cử chỉ chuyển động được thực hiện theo cách truyền thống là thuê các diễn viên sử dụng các tracker và bố trí các hệ thống cảm biến xung quanh thu nhận chuyển động để đạt được độ chính xác nhân thực nhất. Tuy nhiên, các chuyển động thu được sau đó chỉ được phát lại và không có sự biến chuyển giữa các hành động hay chuyển động, chính vì vậy, việc áp dụng trí tuệ nhân tạo để có thể học các chuyển động từ dữ liệu thu nhận và sau đó có thể sinh ra dữ liệu mới sẽ là một cuộc cách mạng trong ngành công nghiệp Motion Capture.

Vào năm 2011, một nhóm tác giả [2] đã chứng minh rằng có sự liên hệ giữa giọng nói và cử chỉ con người, đây chính là tiền đề để cho thấy chúng ta có thể dùng dữ liệu âm thanh để có thể dùng để học và biểu diễn được cử chỉ con người. Với sự thành công của các mô hình ngôn ngữ tự nhiên trong việc xử lý ngôn ngữ văn bản, với sự chính xác siêu thật trong việc mô phỏng gương mặt con người trong lĩnh vực Đồ họa máy tính và với sự chính xác và dễ dàng từ việc tổng hợp giọng nói con người hiện nay. Thì việc ứng dụng trí tuệ nhân tạo để sinh cử chỉ con người là một trong những điểm nghẽn cổ chai duy nhất trong việc phát triển một trợ lý ảo để trao đổi và tương tác với con người.

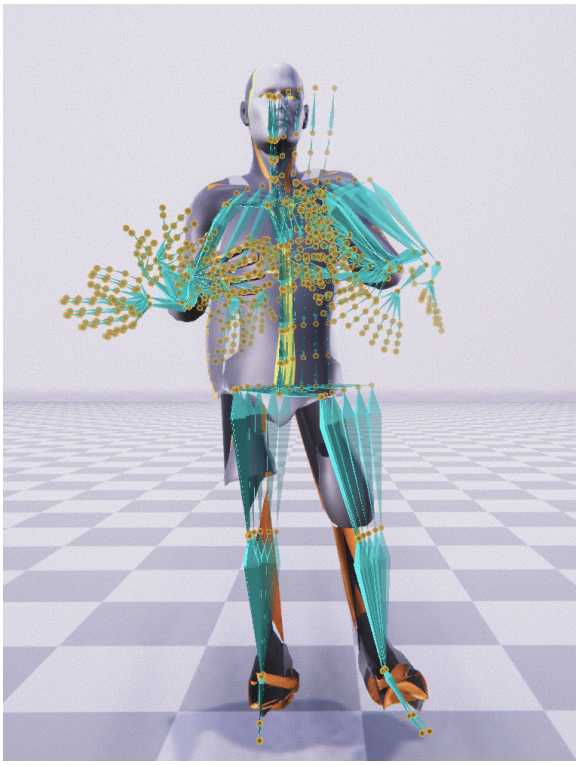
3 Phát biểu bài toán

Mục tiêu của việc sinh cử chỉ (gesture generation) bằng phương pháp học máy là tạo ra các cử chỉ tự nhiên, chân thật như con người (human-likeness) và đồng thời phù hợp với ngữ cảnh. Sinh cử chỉ là bài toán hồi quy (regression), với đầu vào là một chuỗi cử chỉ cho trước và kết quả đầu ra là chuỗi cử chỉ tiếp tục với cử chỉ trước đó.

Mỗi khung hình chuyển động của một nhân vật hay khung xương (skeleton) bao gồm dữ liệu về toạ độ vị trí và vận tốc theo thời gian. Ở đây dữ liệu của chúng tôi của



(a) Khung xương và tên của các khớp của một khung xương trong mỗi khung hình.



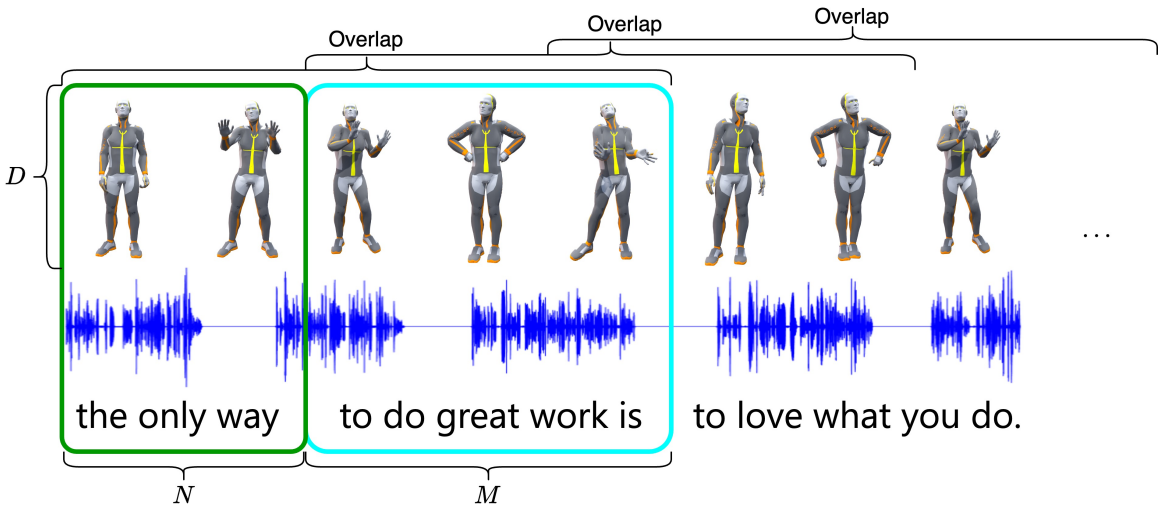
(b) Chuỗi chuyển động của cử chỉ bao gồm 6 cử chỉ quá khứ và 6 cử chỉ tương lai.

một khung xương với mỗi khung hình (frame) bao gồm:

$$\mathbf{g} = \left[\mathbf{p}_{\text{root}}, \mathbf{r}_{\text{root}}, \mathbf{p}'_{\text{root}}, \mathbf{r}'_{\text{root}}, \mathbf{p}_{\text{joins}}, \mathbf{r}_{\text{joins}}, \mathbf{p}'_{\text{joins}}, \mathbf{r}'_{\text{joins}}, \mathbf{d}_{\text{gaze}}\right] \tag{1}$$

Trong đó với mỗi $\mathbf{g} \in \mathbb{R}^{1141}$ bao gồm:

- $\mathbf{p}_{\text{root}} \in \mathbb{R}^3$: Toạ độ của điểm gốc
- $\mathbf{r}_{\text{root}} \in \mathbb{R}^4$: Góc quay của điểm gốc
- $\mathbf{p}'_{\text{root}} \in \mathbb{R}^3$: Vận tốc thay đổi của toạ độ gốc
- $\mathbf{r}'_{\text{root}} \in \mathbb{R}^3$: Vận tốc thay đổi của góc quay gốc
- $\mathbf{p}_{\text{joins}} \in \mathbb{R}^{3n_{\text{join}}}$: Toạ độ của các khung xương
- $\mathbf{r}_{\text{joins}} \in \mathbb{R}^{6n_{\text{join}}}$: Góc quay của các khung xương theo mặc phẳng X và Y
- $\mathbf{p}'_{\text{joins}} \in \mathbb{R}^{3n_{\text{join}}}$: Vận tốc thay đổi của toạ độ các khung xương
- $\mathbf{r}'_{\text{joins}} \in \mathbb{R}^{3n_{\text{join}}}$: Vận tốc thay đổi của góc quay các khung xương



Hình 3: Minh hoạ một chuỗi cử chỉ, ta lấy N frame đầu làm cử chỉ khởi tạo s (seed gesture) và M khung hình còn lại làm cử chỉ để học

- $\mathbf{d}_{\text{gaze}} \in \mathbb{R}^3$: Là hướng nhìn

Tổng cộng có 75 khớp (joins) hay $n_{\text{join}} = 75$, với mỗi khung hình (frame) ta sẽ có một vector gồm 1141 chiều. Tập dữ liệu là tập nhiều chuỗi cử chỉ với độ dài tuỳ ý, từ mỗi cử chỉ độ dài tuỳ tý ta sẽ cắt thành các đoạn $N + M$ khung hình, $g \in \mathbb{R}^{(N+M)\times 1141}$, trong đó cử chỉ $\mathbf{s} \in \mathbb{R}^{N\times 1141}$ đầu tiên là cử chỉ khởi tạo (seed gesture), $\mathbf{x} \in \mathbb{R}^{M\times 1141}$ cử chỉ tiếp theo cho việc dự đoán.

Dữ liệu âm thanh $\mathbf{a}_{\text{raw}} \in \mathbb{R}^{\text{length}}$ là chuỗi âm thanh thô được đọc ở sample rate 16000, sau đó được cắt thành $\mathbf{a} \in \mathbb{R}^{64000}$ tương ứng với 4 giây.

Quá trình xử lý dữ liệu được trình bày đầy đủ ở phụ lục **??**.

4 Các khó khăn cần giải quyết

Có rất nhiều khó khăn trong việc xây dựng một mô hình có thể học được các đăng trưng cử chỉ hội thoại như con người

Thứ nhất, *dữ liệu không đủ nhiều và chất lượng*, chi phí để tạo ra một bộ dữ liệu trong ngành công nghiệp Motion Capture có chất lượng và quy mô lớn để ứng dụng vào trong thực tế là rất lớn.

Thứ hai, *sự thiếu đồng nhất của về ngữ cảnh của các loại dữ liệu*, các bộ dữ liệu về văn bản thường rất nhiều hơn so với giọng nói và cũng không thể biết được văn

bản đó được tạo bởi người nào, sự đồng bộ giữa giọng nói và cảm xúc lúc nói cũng thiếu trong việc tạo tập dữ liệu. Ngoài ra, các dữ liệu văn bản lại được nói bởi rất nhiều người khác nhau và ở nhiều nội dung và chủ đề khác nhau.

Thứ ba, *sự phân bố không cân xứng về dữ liệu giữa các loại đăng trưng cần học*. Các dữ liệu dùng cho nghiên cứu cử chỉ hiện nay thường tập trung vào ngôn ngữ Tiếng Anh, các cử chỉ có sự phân bố không cân xứng giữa khi nói, khi hỏi, khi im lặng.

Thứ tư, *chi phí tính toán với nhiều loại dữ liệu của mô hình rất lớn*. Với đầu vào của mô hình gồm rất nhiều loại dữ liệu khác nhau như văn bản, tiếng nói và điểm 3D, nên cần rất nhiều lớp encode cho từng loại dữ liệu cũng là rào cản không nhỏ bởi chi phí tính toán khi huấn luyện và khi suy luận. Nếu giảm thông tin dữ liệu đầu vào cũng sẽ giảm kết quả suy luận của mô hình khi sinh cử chỉ.

Cuối cùng, *các bước xử lý phải yêu cầu tuần tự*, cách hiệu quả nhất để con người tương tác với máy tính là thông qua giọng nói và nhập từ bàn phím, tuy nhiên việc xử lý được văn bản và giọng nói để làm đầu vào cho mô hình phải thực hiện tuần tự, độ trễ để có thể suy luận trong sản phẩm thực tế cũng là một vấn đề bởi người dùng không thể đợi lâu để nhận được kết quả cử chỉ và từ cử chỉ đó biểu diễn lên máy tính bằng kỹ thuật đồ họa máy tính.

5 Đóng góp dự kiến

- Dựa trên tập dữ liệu có sẵn, chúng tôi chuyển âm thanh của dữ liệu thành văn bản, và dùng văn bản đó để làm dữ liệu huấn luyện mới.
- Dựa trên mô hình cơ bản DiffuseStyleGesture, chúng tôi mở rộng thêm đặc trưng văn bản trong quá trình khử nhiễu có điều kiện.
- Chúng tôi sử dụng Unity để render, trích xuất dữ liệu và trực quan hoá kết quả sinh cử chỉ.
- Chúng tôi xây dựng hệ thống kết xuất, và minh hoạ chương trình bằng Unity

Chương 2

Các công trình liên quan

Bài toán sinh cử chỉ là cũng tương tự như các bài toán khác đều đã nghiên cứu và phát triển song hành cùng các phương pháp học máy truyền thống cũng như các phương pháp học sâu hiện đại ngày nay, gồm các nhóm phương pháp dựa trên luật và các phương pháp dựa trên dữ liệu.

1 Mối quan hệ giữa cử chỉ và lời nói

Cử chỉ được chia thành 6 nhóm chính theo ngôn ngữ học [6], [19] cử chỉ thích nghi (adaptors), cử chỉ biểu tượng (emblems), cử chỉ chỉ định (deictics), cử chỉ biểu trưng (iconics), cử chỉ ẩn dụ (metaphorics) và cử chỉ nhấn mạnh (beat).

Trong đó, cử chỉ nhấn mạnh không liên quan trực tiếp đến ngữ nghĩa lời nói [10] nhưng rất quan trọng để tạo sự hài hòa về nhịp điệu giữa lời nói và cử chỉ [19] . Tuy nhiên, lời nói và cử chỉ nhấn mạnh không đồng bộ hoàn toàn về mặt nhịp điệu [15], nên việc học mối liên hệ thời gian giữa chúng gặp khó khăn [3], [11], [21].

Cử chỉ liên quan đến các cấp độ khác nhau của thông tin lời nói [19]. Ví dụ, cử chỉ biểu tượng như cầm ngược cái ngón cái thường đi kèm với ngữ nghĩa cấp cao như tốt hay tuyệt vời, trong khi cử chỉ nhấn mạnh thường xuất hiện cùng với sự nhấn mạnh âm thanh cấp thấp. Nhiều nghiên cứu trước đây chỉ sử dụng các đặc trưng được trích xuất từ lớp cuối cùng của bộ mã hóa âm thanh để tổng hợp cử chỉ [1], [3], [12], [18], [22]. Tuy nhiên, cách thiết lập này có thể khuyến khích bộ mã hóa trộn lẫn thông tin lời nói ở nhiều cấp độ khác nhau vào cùng một đặc trưng, gây ra sự mơ hồ và tăng độ khó khăn trong việc khai thác các dấu hiệu nhịp điệu và ngữ nghĩa rõ ràng.

Như được chỉ ra trong các tài liệu ngôn ngữ học [10] [17] [20], cử chỉ được sử dụng trong cuộc hội thoại hàng ngày có thể được chia thành một số lượng hạn chế các đơn vị ngữ nghĩa với các biến thể chuyển động khác nhau. Chúng tôi giả định rằng các đơn vị ngữ nghĩa này, thường được gọi là từ ngữ, liên quan đến các đặc trưng cấp cao của âm thanh lời nói, trong khi các biến thể chuyển động được xác định bởi các đặc trưng âm thanh cấp thấp. Do đó, chúng tôi tách rời các đặc trưng cấp cao và cấp thấp từ các lớp khác nhau của bộ mã hóa âm thanh và học các ánh xạ giữa chúng và các từ ngữ cử chỉ và các biến thể chuyển động, tương ứng. Các thử nghiệm chứng minh cơ chế này thành công trong việc tách rời các đặc trưng ở nhiều cấp độ của cả lời nói và chuyển động và tổng hợp các cử chỉ phù hợp về mặt ngữ nghĩa và có phong cách.

2 Tổng quan các phương pháp cho bài toán sinh cử chỉ

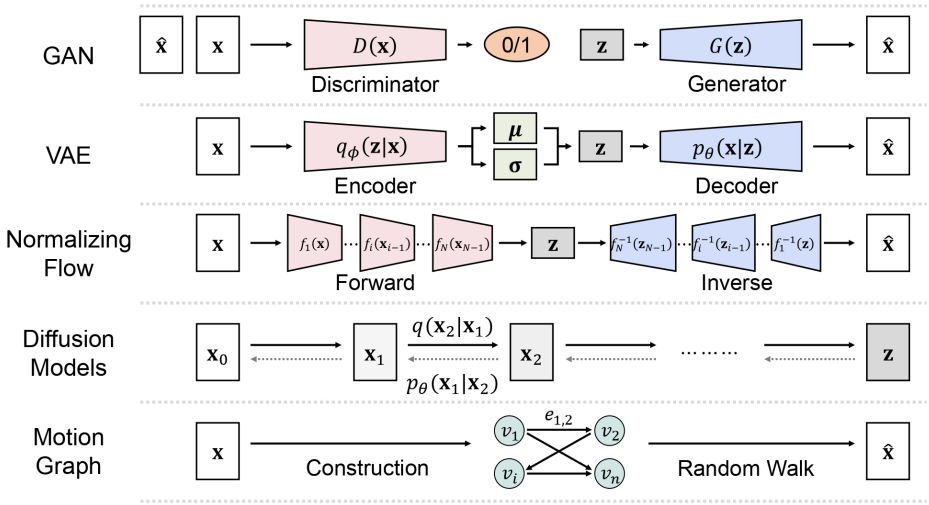
2.1 Phương pháp dựa trên luật

Các phương pháp dựa trên luật thường ánh xạ (mappings) từng âm thanh với từng đơn vị cử chỉ [9]. Và luật được tạo thủ công. Phương pháp dựa trên luật thì chúng ta có thể dễ dàng điều khiển kết quả của mô hình và có khả năng giải thích tốt kết quả dự đoán của mô hình. Tuy nhiên chi phí để tạo thủ công là không khả thi để xây dựng cho các ứng dụng phức tạp đòi hỏi phải xử lý một lượng dữ liệu rất lớn.

2.2 Phương pháp dựa trên thống kê

Tương tự như phương pháp dựa trên luật, phương pháp dựa trên dữ liệu cũng ánh xạ các đặc trưng của âm thanh tương ứng với cử chỉ nhưng thay vì làm thủ công thì được sử dụng học một cách tự động dựa trên dữ liệu. Trong đó có hai phương pháp chính là phương pháp thống kê và phương pháp dựa trên dữ liệu.

Phương pháp thống kê sử dụng phân phối xác xuất để tìm sự tương đồng giữa các đặc trưng âm thanh và cử chỉ [13]. Tác giả [17] xây dựng mô hình để học từng phong cách của từng người nói.



Hình 4: Tổng quan về các mô hình tạo sinh khác nhau.

2.3 Phương pháp học sâu

Phương pháp sinh cử chỉ được chia thành hai nhóm chính. Bao gồm các mô hình ước lượng log likelihood (likelihood-base model) và phương pháp dựa vào các mô hình sinh ngầm định (implicit generative models).

2.3.1 Likelihood-base Model

Phương pháp học ước lượng log likelihood là phương pháp học trực tiếp từ hàm mật độ xác xuất (probability density) thông qua maximum likelihood. Các phương pháp điển hình là autoregressive models, normalizing flow models, energy-based models (EBMs) , và variational auto-encoders (VAEs).

Mô hình được kết hợp với văn bản đầu vào được gắn thẻ với chủ đề, trọng tâm câu và thành ngữ để tạo ra các kịch bản cử chỉ, sau đó được ánh xạ sang một chuỗi các cử chỉ được chọn từ một từ điển hoạt họa. [5] huấn luyện một model classifier neuron network để chọn một đơn vị cử chỉ phù hợp dựa trên đầu vào là lời nói.

Cử chỉ có thể được tổng hợp bằng các mô hình xác định như perceptron đa tầng (MLP) [11], recurrent neural networks [3], [14], [8], [21], convolutional networks [7] và transformer [4]

2.3.2 Implicit Generative Models

Trong các phương pháp dựa vào mô hình sinh ngầm định, phân phối của dữ liệu được học một cách ngầm định thông qua việc quá sinh lấy mẫu (sampling). Ví dụ tiêu biểu nhất là mô hình generative adversarial networks (GANs). Khi dữ liệu được tổng hợp bằng cách chuyển phân phối dữ liệu ban đầu ở dạng phân phối chuẩn về phân phối của dữ liệu.

Trong mục tiêu theo chúng tôi sẽ trình bày các phương pháp sử dụng mô hình Diffusion.

TÀI LIỆU THAM KHẢO

[1] Simon Alexanderson, Gustav Eje Henter, Taras Kucherenko, and Jonas Beskow. Style-controllable speech-driven gesture synthesis using normalising flows. *Computer Graphics Forum*, 39(2):487–496, 2020.

[2] Kirsten Bergmann, Volkan Aksu, and Stefan Kopp. The relation of speech and gestures: Temporal synchrony follows semantic synchrony. In *Proceedings of the 2nd Workshop on Gesture and Speech in Interaction (GeSpIn 2011)*, 2011.

[3] Uttaran Bhattacharya, Elizabeth Childs, Nicholas Rewkowski, and Dinesh Manocha. Speech2affectivegestures: Synthesizing co-speech gestures with generative adversarial affective expression learning. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2027–2036, 2021.

[4] Uttaran Bhattacharya, Nicholas Rewkowski, Abhishek Banerjee, Pooja Guhan, Aniket Bera, and Dinesh Manocha. Text2gestures: A transformer-based network for generating emotive body gestures for virtual agents. In *2021 IEEE virtual reality and 3D user interfaces (VR)*, pages 1–10. IEEE, 2021.

[5] Chung-Cheng Chiu, Louis-Philippe Morency, and Stacy Marsella. Predicting co-verbal gestures: A deep and temporal modeling approach. In *Intelligent Virtual Agents: 15th International Conference, IVA 2015, Delft, The Netherlands, August 26-28, 2015, Proceedings 15*, pages 152–166. Springer, 2015.

[6] Paul Ekman and Wallace V Friesen. The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *semiotica*, 1(1):49–98, 1969.

[7] Ikhsanul Habibie, Weipeng Xu, Dushyant Mehta, Lingjie Liu, Hans-Peter Seidel, Gerard Pons-Moll, Mohamed Elgharib, and Christian Theobalt. Learning

speech-driven 3d conversational gestures from video. In *Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents*, pages 101–108, 2021.

[8] Dai Hasegawa, Naoshi Kaneko, Shinichi Shirakawa, Hiroshi Sakuta, and Kazuhiko Sumi. Evaluation of speech-to-gesture generation using bi-directional lstm network. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, pages 79–86, 2018.

[9] Chien-Ming Huang and Bilge Mutlu. Robot behavior toolkit: generating effective social behaviors for robots. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*, pages 25–32, 2012.

[10] Michael Kipp. *Gesture generation by imitation: From human behavior to computer character animation*. Universal-Publishers, 2005.

[11] Taras Kucherenko, Patrik Jonell, Sanne Van Waveren, Gustav Eje Henter, Simon Alexandersson, Iolanda Leite, and Hedvig Kjellström. Gesticulator: A framework for semantically-aware speech-driven gesture generation. In *Proceedings of the 2020 international conference on multimodal interaction*, pages 242–250, 2020.

[12] Taras Kucherenko, Patrik Jonell, Youngwoo Yoon, Pieter Wolfert, and Gustav Eje Henter. A large, crowdsourced evaluation of gesture generation systems on common data: The genea challenge 2020. In *26th international conference on intelligent user interfaces*, pages 11–21, 2021.

[13] Sergey Levine, Philipp Krähenbühl, Sebastian Thrun, and Vladlen Koltun. Gesture controllers. In *Acm siggraph 2010 papers*, pages 1–11. 2010.

[14] Xian Liu, Qianyi Wu, Hang Zhou, Yinghao Xu, Rui Qian, Xinyi Lin, Xiaowei Zhou, Wayne Wu, Bo Dai, and Bolei Zhou. Learning hierarchical cross-modal association for co-speech gesture generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10462–10472, 2022.

[15] Evelyn McClave. Gestural beats: The rhythm hypothesis. *Journal of psycholinguistic research*, 23(1):45–66, 1994.

[16] Adam Metallo, Vincent Rossi, Jonathan Blundell, Günter Waibel, Paul Graham, Graham Fyffe, Xueming Yu, and Paul Debevec. Scanning and printing a 3d portrait of president barack obama. In *SIGGRAPH 2015: Studio*, pages 1–1. 2015.

[17] Michael Neff, Michael Kipp, Irene Albrecht, and Hans-Peter Seidel. Gesture modeling and animation based on a probabilistic re-creation of speaker style. *ACM Transactions On Graphics (TOG)*, 27(1):1–24, 2008.

[18] Shenhan Qian, Zhi Tu, Yihao Zhi, Wen Liu, and Shenghua Gao. Speech drives templates: Co-speech gesture synthesis with learned templates. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11077–11086, 2021.

[19] Thomas A Sebeok and Jean Umiker-Sebeok. *Advances in visual semiotics: The semiotic web 1992-93*, volume 118. Walter de Gruyter, 2011.

[20] Rebecca A Webb. *Linguistic features of metaphoric gestures*. University of Rochester, 1997.

[21] Youngwoo Yoon, Bok Cha, Joo-Haeng Lee, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. Speech gesture generation from the trimodal context of text, audio, and speaker identity. *ACM Transactions on Graphics (TOG)*, 39(6):1–16, 2020.

[22] Youngwoo Yoon, Pieter Wolfert, Taras Kucherenko, Carla Viegas, Teodor Nikolov, Mihail Tsakov, and Gustav Eje Henter. The genea challenge 2022: A large evaluation of data-driven co-speech gesture generation. In *Proceedings of the 2022 International Conference on Multimodal Interaction*, pages 736–747, 2022.