Design and Implementation of Virtual Host Based on Machine Learning Algorithm

Xuan Ji*
Huangshan University, Huangshan, Anhui, 245041 China

*Corresponding author: h8866t@163.com

Abstract. The research of face modeling and animation technology has a history of 30 years. With the development of virtual reality, video entertainment, videophone, human-computer interaction and other applications, this research field has attracted more and more attention. Based on machine learning algorithm, this paper studies the personalized 3D face modeling and real-time rendering of virtual hosts. The research lines of this paper are the establishment of neutral face model data structure, face feature extraction, improvement of personalized face model matching and real-time rendering in scene roaming. A realistic 3D facial lip animation system based on MPEG-4 standard and driven by voice files is designed and implemented. The system is easy to operate, versatile and efficient, and meets the needs of real-time voice-driven lip animation.

Keywords: Machine learning algorithm; Virtual host; Face animation; Voice drive

1. Introduction
While foreign technology giants are actively deploying virtual reality, domestic capital is also actively participating. High-quality listed companies with industry barriers in the field of virtual reality have appeared one after another. The interaction between real people and virtual characters is also increasing, which puts forward higher visual requirements for the humanization of human-computer interaction interface. Under this background, virtual human synthesis technology came into being [1]. On the one hand, the modeling of real human face is of great academic significance, and it is one of the most fundamental and difficult problems in the field of computer graphics and computer vision at present. At the same time, it is also an interdisciplinary problem including physiology, psychology, physics and other fields. On the other hand, 3D face model has been widely used, and with the development of various technologies, 3D face model will be applied more deeply and widely in various aspects.

The 3D face animation technology synchronized with Chinese speech can be applied to most fields in real life, such as news broadcast, dialogue system, virtual conference, 3D game entertainment, etc., which can not only improve the realism of pictures, but also enhance the human-computer interaction and comprehensibility [2-3]. At present, the goal in the field of cultural creativity in China is to open up artistic expressions with the technical characteristics of visual immersion, dynamic experience and interactive multi-mode, extend the animation creative industry chain, innovate the profit model of
digital media, and enhance the support and leading power of science and technology to promote the development of cultural industry. This paper introduces the design and implementation of a virtual host based on machine learning algorithm. By using Active X technology, adopting the H-Anim standard of VRML [4] and the definition of standard virtual human in MPEG4, and integrating our Chinese sign language synthesis technology, an online virtual host is realized.

2. Extraction of speech feature parameters

2.1 Digitization of speech signal

In order to store the analog voice signals from human beings in a computer, it is necessary to digitize the voice and change it into digital signals with discrete time and amplitude. This process includes two steps: sampling and quantization [5]. The process of speech signal digitization is shown in Figure 1.

![Fig.1 Digital processing process](image)

In the sampling process, the original analog signal is firstly cut into several equally spaced speech segments in the time domain, and then extracted from these speech segments. Finally, the analog audio sequence is converted into a file composed of binary codes, namely digital audio. First, we divide the amplitude value of the waveform into some finite intervals, and then give the same amplitude to those samples that fall within the same interval. There may be a certain error between the quantized signal value and the original sample value, which is called quantization error or quantization noise.

2.2 Pretreatment of speech signal

Usually, we preprocess the speech signal before extracting the feature parameters of the speech signal, and distinguish the high-frequency part from the low-frequency part of the speech signal, so that the feature parameters can be extracted more easily and the error is smaller. Pre-processing includes pre-emphasis, framing, windowing and endpoint detection. The flow of pretreatment process is shown in Figure 2.

![Fig.2 Pretreatment process](image)

According to the sign language movement of virtual human, a virtual human movement consists of several frames of virtual human gestures, and each gesture is an instantaneous state of virtual human. Natural images and computer-generated images have different imaging channels, and naturally they also have different channel characteristics, which are of great significance to image source identification [6]. This is the purpose of pre-emphasis. After pre-emphasis processing, the speech signal will be beneficial to analyze the channel parameters and frequency spectrum. The first-order digital filter is usually used to realize pre-emphasis of digitized speech signals, and the transfer function of the digital filter is expressed by formula (1):

$$H(z) = 1 - \mu z^{-1} \quad 0.9 \leq \mu \leq 1.0$$ (1)

Where, $\mu$ represents the pre-emphasis coefficient, which is usually 0.94.

The result $Y(n)$ after pre-emphasis processing is expressed by formula (2):
\[ Y(n) = x(n) - \mu x(n-1) \] (2)

Where \( x(n) \) is the speech sampling value at \( n \) time.

Speech signal is non-stationary and changes with time. However, according to the pronunciation mechanism, the muscle movement speed of the pronunciation organ is slower than the sound vibration speed, and scholars have found that the spectrum characteristics and some physical characteristic parameters of the speech signal are relatively stable in a short time range (usually 5ms to 50ms), so scholars think that it can be regarded as a quasi-steady process in a short time range. Therefore, based on the idea of short-term stationarity, we usually use the method of framing to analyze and process speech signals.

Short-time analysis of speech signal is realized by window segmentation. The original speech signal \( x(n) \) is multiplied by a window function \( w(n) \) with finite length to intercept the speech signal to form an analysis frame. The windowed speech signal \( x_w(n) \) is expressed by formula (3):

\[ x_w(n) = x(n) \times w(n) \] (3)

Time-domain analysis of speech signal has the characteristics of simple implementation and small amount of calculation, which is the earliest and most widely used analysis method [7]. However, after being digitized, the speech signal will become a time-varying signal and cannot be directly analyzed. Therefore, we will take the idea of short-time stabilization as the premise and adopt the processing method based on time domain features to process the speech signal.

Short-term average energy represents the energy of each frame of speech after framing and windowing, and the definition of short-term average energy \( E_n \) of speech frame \( x(n) \) is shown in formula (4).

\[ E_n = \sum_{m=-\infty}^{\infty} [x(m)w(n-m)]^2 = \sum_{m=-N+1}^{n} [x(m)w(n-m)]^2 \] (4)

In which \( n \) represents the frame number, \( N \) represents the frame length, and \( w(n) \) represents the window function.

Short-term average energy plays the following roles in speech detection: firstly, because the short-term average energy value of unvoiced segment is much smaller than that of voiced segment, it can be used as a parameter to distinguish unvoiced segment from voiced segment; Secondly, because short-term average energy can determine the starting and ending positions of finals and initials in speech signals, it can be used to obtain the effective time of finals and initials; Finally, as an important index parameter, short-term average energy plays a key role in speech recognition.

2.3 Extraction of MFCC parameters

MFCC is designed based on the auditory mechanism of human ears. Figure 3 shows the extraction process of MFCC [8].

![Fig.3 The extraction process of MFCC](image)

The MFCC calculation steps are as follows:

(1) The speech signal is pre-emphasized and Hamming window is added to output the short-time
speech signal, and then the time-domain signal is transformed into the frequency-domain signal by FFT.

(2) The short-time energy spectrum of speech signal is calculated, and then m Mel band-pass filters are used for filtering. The energy in each Mel band-pass filter band is superimposed.

(3) Take the logarithm of the output result of each filter to obtain the logarithmic power spectrum of the corresponding frequency band. Next, the logarithmic power spectrum is processed by inverse discrete cosine transform to obtain L Mel cepstrum coefficients, which are expressed as follows [9]:

\[ C_n = \sum_{k=1}^{M} \log r(k) \cos \left( \pi (k - 0.5)n/M \right) \quad n = 1, 2, \cdots, L \]  

In this paper, the first 12 dimensions of MFCC speech feature parameters are taken, because the first and last dimensions of MFCC have strong ability to distinguish speech.

2.4 Metaphor recognition based on supervised learning

The principle of supervised learning algorithm in metaphor recognition is to transform metaphor recognition into binary classification of metaphorical meaning and literal meaning, train classifier parameters by labeling a certain scale of metaphor corpus, and then use the trained classifier to automatically recognize metaphor in open corpus.

In this computing model, the context is divided into left context and right context. The position and word order of each word in the context and the relative word frequency information in the corresponding position are considered, and these relative word frequency information are accumulated. This algorithm simulates the process that people synthesize all kinds of knowledge accumulated in the past to resolve ambiguity in context. The specific description is as shown in formula (6):

\[ SUM_m = \sum_{i=1}^{k} f_{m,\text{left}}(W_i) + \sum_{i=k}^{l} f_{m,\text{right}}(W_i) \]  

In which \( i \) represents the window of context, and the value range is \(-1 \leq -k or l \leq k\). \( W_i \) represents the words in \( i \) window position, \( f_{m,\text{left}}(W_i) \) represents the relative word frequency ratio of words in corresponding position in left context, and \( f_{m,\text{right}}(W_i) \) represents the relative word frequency ratio of words in corresponding position in right context. The RFR values of the words in the context at high frequency and low frequency are added separately, and the categories can be distinguished by making decisions. The algorithm is mainly aimed at binary classification.

3. Personalized face modeling

There are many feasible methods to construct the real touching face in the virtual world. The most classic method is to modify a general model with the information of eyes, nose, mouth, ears and hair of a specific face. In this process, the accurate extraction of personalized face information is a very important prerequisite. At the same time, the transformation from neutral face model to personalized face model is the key technology in the whole process, which directly affects the synthesis effect of personalized face.

3.1 Extracting three-dimensional information from two-dimensional images

Face photo is a projection image of a face collected by a camera at a certain angle, which can only reflect the information of the two-dimensional position distribution of various organs of the face on the projection plane, but we need to get the three-dimensional geometric information of the face when we carry out the three-dimensional modeling of the face.

In order to build 3D face model, we need to get 3D geometric information of face. According to the principle of computer vision [10] (as shown in fig. 4), a triangle \( \triangle ABC \) is used for an object in space, and there are projections \( \Delta \hat{A} \hat{B} \hat{C} \) and \( \triangle \hat{A} \hat{B} \hat{C} \) in front and side photos respectively.
In order to simplify the algorithm, it is assumed that the camera makes an orthographic projection on
the face. According to the orthogonality of photos, if the plane coordinates of points $A', B', C'$ and $A'', B'', C''$ on the front and side are known, the real coordinates of space object $ABC$ can be easily obtained. For human faces, if we know the corresponding points of each front projection point in the side photos, we can easily estimate the geometric shape of human faces. Therefore, the face reconstruction method based on orthogonal photos can be attributed to the matching problem of corresponding points in front photos and side photos.

In this paper, the spatial position of feature points is obtained by projection method, that is, the $X,Y$ coordinates of feature points are obtained from front photos and the $Z,Y$ coordinates of feature points are obtained from side photos, so the coordinates in $Y$ direction must be unified to the same scale. Firstly, the $X,Y,Z$ coordinates on the model are calculated, and the ranges are denoted as $x_{\text{min}}, x_{\text{max}}, y_{\text{min}}, y_{\text{max}}, z_{\text{min}}, z_{\text{max}}$, so that the height and width $H_f, W_f$ of the head in front photos and the height and width $H_p, W_p$ of the head in side photos can be obtained. Then the feature points are projected on the front and side photos, and the accurate spatial position of the feature points can be calculated.

Remember that the coordinate of a feature point on the 3D face model is $(X,Y,Z)$, and the corresponding two-dimensional coordinates of feature points on the front and side images are $(x_f, y_f)$ and $(z_p, y_p)$, respectively, and realize the reconstruction from two-dimensional coordinates to three-dimensional coordinates by the following formula.

$$(X,Y,Z) = \left( x_f, \left( y_f + y_p \right)/2, z_p \right)$$ (7)

As $Y$ coordinate will appear in both directions, $y_f$ and $y_p$ may be different, so take the average value.

3.2 Face texture mapping

Two-dimensional texture mapping is essentially a mapping $M$ from two-dimensional color texture plane to three-dimensional scene surface. Generally speaking, the definition of texture map $M$ is related to the representation of scenery, but not unique. Because different definition methods have different mapping effects, the determination of $M$ plays a very important role in texture mapping algorithm, which directly affects the authenticity of the final generated graphics. From a mathematical point of view, the mapping $M$ can be described by the following formula:

$$(u,v) = F(x,y,z)$$ (8)
In which \((u, v)\) and \((x, y, z)\) are points in texture space and object space respectively. Parametric surfaces define a mapping relationship from two-dimensional parametric space to three-dimensional space. Therefore, when the parameter space and texture space are regarded as equal, the mapping \(M\) is equivalent to the inverse mapping defined by the parametric surface itself. For complex high-order parametric surfaces, the texture mapping function can't be expressed analytically, which usually needs to be solved discretely by numerical solution technology.

In this paper, the shape of the head is similar to an ellipsoid, so the cylindrical surface can be selected as the middle surface and the texture coordinates of each point can be obtained by two-step texture mapping method. The two-step texture mapping method firstly maps the texture to a cylindrical middle surface, and then maps the deformed mapping result to the final surface. Assume that the coordinate system of the face model is as shown in Figure 5, the central axis of the human head coincides with the \(Y\) axis, the direct eyes of human eyes are facing the \(Z\) axis direction, and the schematic plan view of the face surface seen from top to bottom against the \(Y\) axis is as shown in Figure 6:

![Fig.5 Three-dimensional coordinate system](image)

![Fig.6 Top view of human face surface](image)

In fig. 5, the inner circle approximates the contour of a human face, the outer circle represents a projected cylindrical surface, and a straight line intersecting the inner circle is projected from the cylindrical axis to point \(P\), where the texture coordinate of \(P\) is \((u, v)\), then \(v\) is the height of the model, \(v = y\), and \(u\) represent the arc length from the positive half axis of the \(x\) axis, and the following formula is obtained from the definition of the arcsine function:
In this way, the range of \( u \) obtained by this formula is \( [-\pi, \pi] \), and the range of \( v \) is the range of model height \( y \). Expand the cylindrical surface into a plane, and in order to meet the requirements of texture coordinate range in texture mapping in the future, classify the range of \( u \) and \( v \) into the range of \([0,1]\) respectively, and assume that the model height range is \([a, b]\), as shown in the following formula:

\[
\begin{align*}
    u &= \frac{(u + \pi)}{2\pi} \\
    v &= \frac{(v - a)}{(b - a)}
\end{align*}
\]  

Through cylindrical mapping, the two-dimensional unfolded plan of three-dimensional face is obtained, and the point coordinates in the two-dimensional unfolded plan corresponding to each three-dimensional point are the texture coordinates of that point.

4. System implementation

The development platform of this experiment is developed by using Visual C++ .NET and Open GL general graphic development library, and analyzed, designed and realized by using object-oriented technology. Open GL is an excellent professional 3D API. In this paper, the frame structure of single document and single view based on MFC and Open GL are adopted to realize the prototype of virtual host face modeling and animation generation system based on machine learning algorithm under Windows operating system platform.

4.1 Three-dimensional display module

Model display includes the display of grid model and the display of real touching face model after texture mapping. Users can observe the model from any angle and size by mouse. The module also provides an interactive interface for editing facial expressions. Users can change the state of each movement unit by selecting their own muscle insertion points, thus synthesizing various expressions.

4.2 Implementation and use of plug-in

According to the above analysis, we have used the ATL3.0 and MFC integrated development environment of VC60, integrated the virtual human synthesis and display development package Virtual Human View virtual human motion control development package Gesture View, and developed a universal browser plug-in containing all the ideas of this paper. Using this plug-in, you can quickly customize the virtual host's online news publishing interface, and truly realize the virtual host who publishes news online based on machine learning algorithm.

4.3 Virtual accessories try-on system

In this project, the technologies of virtual host posture perception, skeleton-driven animation, face image transplantation, skin color enhancement, etc. have been realized, and mastering these technologies will help us to quickly realize a virtual dressing system or virtual clothing try-on system. According to the skin color, shape and other information of the real image.

4.4 Evaluation results of system synthesis effect

The evaluation video used in the experiment is recorded by the speaker reading a piece of material at fast, medium and slow speed, and then using the same voice file to drive the system to synthesize lip-shaped animation with three different speech speeds. After the subjects watch the recorded video, the
answers should be selected from the alternative lip animation, and the five evaluation indexes should be scored respectively. The experimental results are shown in Table 1.

**Table 1 Evaluation results of system synthesis effect**

| speed of speech | Fast and slow truth | Lip transition effect | Lip movement range | Lip fidelity | Corner naturalness | Overall effect | Standard deviation |
|-----------------|--------------------|-----------------------|--------------------|-------------|--------------------|----------------|--------------------|
| fast            | 91.2               | 89.1                  | 86.4               | 86.9        | 88.2               | 89.3           | 1.79               |
| Intermediate speed | 94.0               | 92.2                  | 89.3               | 88.2        | 91.6               | 90.7           | 1.88               |
| low speed       | 88.6               | 91.7                  | 85.4               | 89.0        | 92.7               | 88.3           | 1.26               |

The experimental results show that when the speech speed is at medium speed, the system's fast and slow fidelity, lip transition effect and lip fidelity are all ideal, and the acceptability of the system is the best at this time; when the speech speed is fast, the acceptability of the system is slightly worse. Experiments show that the speech-driven system of realistic 3D face lip animation realized in this paper can simulate lip movements when people speak. The system has low computational complexity, good real-time, simple operation and good interactivity, and meets the needs of real-time speech-driven lip animation.

5. Conclusion

Virtual host is an important subject in human-computer interface. Virtual host must show behaviors similar to real people, including gestures, expressions, lip movements, language and the coordination between these behaviors, so as to meet the needs of human-computer interaction in the network environment. In this paper, the virtual host's face animation and voice drive based on machine learning algorithm are designed and implemented. The design idea of the system is as follows: firstly, select 3D modeling software to establish the extraction of voice feature parameters, and map the face images onto the 3D grid model to obtain a 3D face model with personalized features. Then, under MPEG-4 standard, the grid points in 3D face model are moved by FAP stream, and the lip animation system with synchronized voice and lip is obtained after rendering. The experimental results show that the system realized in this paper can generate lip animation with high reality in real time, with smooth picture and strong operability.

References

[1] Ding Baohua, Wang Danni. On the integration, innovation and development of virtual hosts and traditional hosts. Media Review, no. 11, pp. 92-93, 2019.
[2] Dong Changfu, Wang Yulin, Zhu Congxiao, et al. Design and implementation of virtual training system for three-dimensional placement of audio and video equipment. Chinese Science and Technology Thesis, vol. 13, no. 20, pp. 60-64, 2018.
[3] Yin K, Zhao G, Xu C, et al. Prediction of Toxoplasma gondii virulence factor ROP18 competitive inhibitors by virtual screening. Filaria Journal, 2019, 12.
[4] Zhang Lin. Design and Implementation of Virtual Learning Community. Western Quality Education, vol. 003, no. 007, pp. 282, 2017.
[5] Yu Lu, Jia Shaowen, Huo Liping. Design and implementation of Web 3D virtual training software based on WebGL. Instrumentation Technology, vol. 000, no. 002, pp. 40-42, 2017.
[6] Howe A, Campion P, Searle J, et al. New perspectives-approaches to medical education at four new UK medical schools. BMJ no. Clinical Research ed.), vol. 329, no. 7461, pp. 327-31, 2018.
[7] Lin Min. Design and Implementation of Online Live Platform for Virtual Courses Based on Open Source Architecture. Jiangsu Science and Technology Information, vol. 037, no. 008, pp. 56-59, 2020.
[8] Ji Xuan. Discussion on the Development Dilemma of Virtual Hosts in China. Research on Communication Power, vol. 000, no. 018, pp. 151-151, 2019.
[9] Yang Yongxian, Chang Yufeng. Design and Implementation of Virtual Studio Based on Cloud
Computing. Television Technology, vol. 42, no. 10, pp. 69-71+104, 2018.
[10] Yuan Xiaoyan. Design and Implementation of Video Editing Teaching Game Based on Virtual Reality Technology. Popular Literature and Art, no. 06, pp. 240-241, 2019.