International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 05 | May 2022 www.irjet.net p-ISSN: 2395-0072

Real Time Head Generation for Video Conferencing

Prof. Sumedha Ayachit[1], Rohan Sabale[2], Avinash Parit[3], Shreya Shere[4] , Varun Raikar[5]

[1]Prof. Dept. of Information Technology, JSCOE, Pune, Maharashtra, India
[2,3,4,5] Dept. of Information Technology, JSCOE, Pune, Maharashtra, India
Department of Information Technology, Jayawantrao Sawant College of Engineering, Pune

---

Abstract - *Video conferences area unit receiving genuine interest as **a** technique of interaction as web conferencing has improved dramatically in the last few years. WebRTC technology uses a high-speed data transmission channel to establish communication between users. We propose a real-time talking-head video synthesis model for video conferencing. Our model reconstructs a source video at the receiver side to maintain a steady and lag-free experience for face-to-face video conferencing. It uses less bandwidth compared to the commercial H.264 standard. It extracts and retargets motion from the sender video frame by frame and synthesis video on the receiver end using the first image and motion keypoints. It successfully takes out facial expressions, head poses, and eye movements from the faces. It synthesizes the talking-head from the original viewpoint of the source image.*

Key Words: **Video conferencing, Machine Learning, WebRTC**

1. INTRODUCTION

Digital property composed of wireless, satellite technologies, and wired technologies these are the utility of the twenty first century. Digital life helps us improve many acreages of our lives like virtual education, monitoring health, and keeping physicians informed using internet technology and other aspects of life.

Due to COVID-19 pandemic there is an unprecedented need of cyber technologies and its related online services. One of the main key eventful changes concerned was how individuals and communities socialize and communicate with each another. Virtual proceedings have become the most sort of top of the list way of communicating and holding meeting or teaching all over the world from businesses to schools and colleges as they can communicate from the comfort of homes.
Like current one of the biggest video conferencing service providers, Zoom had usage of 10 million daily meeting participant's pre covid which are far lesser than now. Post Covid they have now hundreds of million daily participants. But for many people, the increased usage of this growing tech has been troublesome. It has many challenges, like digital infrastructure, affordability of data connection, and quality internet access.

The driving force of our new modern era has become the new and advanced technology. Hence, using technology like WebRTC and FOMM, we developed an advanced Video Conferencing System which can even be used in lower bandwidth and poor network connections which will help people facing these issues. This paper mainly focuses on outlining and executing a web-based conferencing system by initiating peer to peer connection using Web-RTC and face reconstruction and reformation by the facial data point already extracted using First Order Motion Model. This video conferencing system will help connect people who are having poor digital Infrastructure and cannot afford high speed data connection.

1.1 Background and Related Work

The overall design of the WebRTC system is arranged out in [1] [2], and therefore the data is pictured to flow in a very peer-to-peer fashion directly between the two net browsers. Therefore the protocols used to transport, communicate and secure the encrypted media are specified. The main points relating to the WebRTC information channels square measure arranged clearly. The problems associated with the transport layers unit of measurement revolve around NAT (Network Address Traversal) and firewall traversal, multiplexing of information and media over one transport flow is additionally addressed.

Earlier works on deep video generation are mentioned. However, Spatio-temporal neural networks get render video frames from noise vectors [6]. A lot of recently, many approaches tackled the matter of conditional video generation. For example, Wang et al. [7] combine a recurrent neural network with a VAE to get face videos. Considering a more comprehensive variety of applications, Tulyakov et al. [8] introduced MoCoGAN, a continual design adversarially trained to synthesize videos from noise, categorical labels, or static pictures.

X2Face [5] uses a dense motion field which is extracted by the image to get the output video via image distortion. Equally to us, they use a reference create that's wont to get a canonical illustration of the article. In our formulation, we don't need an exact connection to make, resulting in considerably less complicated optimization and improved image quality.

2. Method

2.1 WebRTC

WebRTC has the semantic client-server organizer with the concept of peer-to-peer communication among various browsers. These connections manages each and every media paths which is responsible for a direct flow between browsers. Network signaling is done in the Web Servers that help in changing, managing or interpreting the signals, as it is required by Web Sockets or HTTP. It was recognized that the signals between the browsers and servers are not constant in WebRTC, where they are the members of the application. Web servers usually and mainly performs interactions by the signaling protocol such as SIP. Or else, in order to achieve the goal property signaling protocol can be used.

The WebRTC web application interacts with browsers which uses both standard applications of API and WebRTC application, enterprising (for e.g. IBC inquiry browser competency) and an interactive method (for e.g. receiving a cross browser notification) .The real time imaging communication such as video and audio call among various browsers consist of media streaming between two browsers, within the media path and creates an instance of multifaceted interaction.

The World Wide Web Consortium WebRTC API permits JavaScript applications to take the benefits of the real time features of an uncommon browsers. Real time browser functions which were performed on the browser and gives the required functions for setting the video, audio and data channel which is required and permitted. Three concepts that has been relied upon in the designing of the APIs which stands for Application Programming Interface are Media Stream, Peer Connection and Data Channel.

1. Media Stream : It represents media streaming from local media devices such as webcam and microphone .The web application needs to request user access to make use of a local stream through the function called as GetUserMedia [1].

2. Peer Connection: It mainly reads the output data from Media Streams and make the connections between two users. In order to establish peer to peer connections, TURN and STUN protocols are mainly used by an Interactive connectivity establishment structure as it is a core part of Network address translator , where protocols are mainly given by the Google.

3. Data Channel: It is a bi-directional data channel between two peers which provides the possibility of exchange of random data among them. Each RTC Data Channel provides which is given as follows:

- Reliable or unreliable transportation of messages.
- In order or out order transportation of messages

2.2 FOMM

For the Creation of animated videos from the still image, we make use of the First-order motion model.

Here the source image is taken first from the driving video and its video sequence is created concurrently with the motion of the driving video.

In this framework, the model doesn't require annotation or key points or prior information about the targeted object to animate. Once it is trained on a set of videos having similar categories, like faces usually have the same data points, then it can be easily applied to any object of the concerned class.

For achieving perfect output, it takes the help of self-supervised formulation in which it decouples the appearance of the current face frame and motion information like facial key points already extracted. For faces, it successfully extracts and retargets from driving video head poses, eye movements, and facial expressions.

This model basically works on two different modules, the motion estimation module and the image generation module. The capability of the first module is to predict and gather a dense motion field from the input image. Here we assume the existence of an abstract reference frame. We basically estimate two transformations first one being a reference to source and the other one being a reference to driving. By this model, we can process source and driving frames independently. This is carried out because at the time of testing or training model can receive pairs of the source images and driving frame sampled from a different video which can be very different visually.

The second module, which is the generation module, is used to render an image of the source object moving as given in the base driving video. Here the model makes use of the generator network which is used to warps the source image according to dense motion and paints the image parts that occluded in the source image.

3. SYSTEM ARCHITECTURE

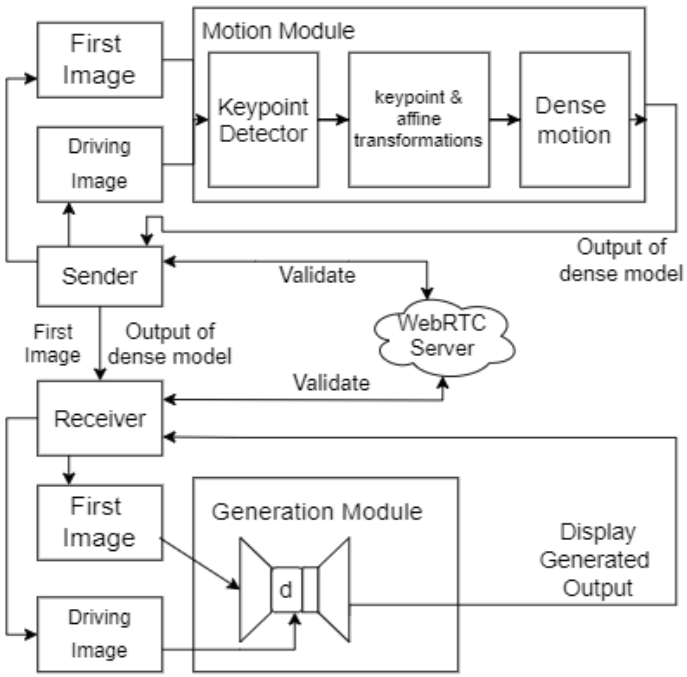

Figure 3.1: System Architecture

4. CONCLUSIONS

Real-Time head generation for Video Conferencing systems helps people socialize over the internet. This system works even in lower bandwidth and remote areas where infrastructure is not developed yet.

Many different methods have been used in the making of this system. Among them, Web-RTC and machine learning play a vital role. Web RTC has the semantic client-server organizer with the concept of peer-to-peer communication among various browsers. Machine Learning makes a significant usage of the First Order Motion Model.

REFERENCES

[1] Zinah Tareq Nayyef, Sarah Faris Amer, Zena Hussain: "Peer to Peer Multimedia Real-Time Communication System based on WebRTC Technology" In IJHET, 2019

[2] Felix Weinrank, Martin Becke, et al, "WebRTC Data Channels", IEEE Communications Standard Magazine, vol. 1, no. 2, July 2017, pp. 28-35, DOI: 10.1109/MCOMSTD.2017.1700007

[3] Aliaksandr Siarohin, Sergey Tulyakov,Elisa Ricci, St´ephane Lathuili`ere, and Nicu Sebe. "First order motion model for image animation". In NeurIPS, 2019.

[4] Qiming Hou, Chen Cao, and Kun Zhou. "Displaced dynamic expression regression for real-time facial tracking and animation". TOG, 2014.

[5] OliviaWiles, A Sophia Koepke, and Andrew Zisserman. "X2face: A network for controlling face generation using images, audio, and pose codes." In ECCV, 2018.

[6] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. "Generating videos with scene dynamics." In NIPS, 2016.

[7] Wei Wang, Xavier Alameda-Pineda, Dan Xu, Pascal Fua, Elisa Ricci, and Nicu Sebe. "Every smile is unique: Landmark-guided diverse smile generation". In CVPR, 2018.

[8] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. Mocogan: "Decomposing motion and content for video generation". In CVPR, 2018.