An Online Intelligent Visual Interaction System

Anxiang Zeng[1] , **Han Yu**[1] , **Xin Gao**[2] , **Kairi Ou**[2] , **Zhenchuan Huang**[2] , **Peng Hou**[2] , **Mingli Song**[3] , **Jingshu Zhang**[4] and **Chunyan Miao**[1]
[1]Nanyang Technological University,
[2]Alibaba-group,
[3]Zhejiang University,
[4]Sichuan University
{zeng0118, han.yu, ASCYMiao}@ntu.edu.sg, {suzhe.okr}@taobao.com, {zimu.gx, zhenchun.hzc, houpeng.hp}@alibaba-inc.com, brooksong@zju.edu.cn, zhangjingshu@scujcc.cn

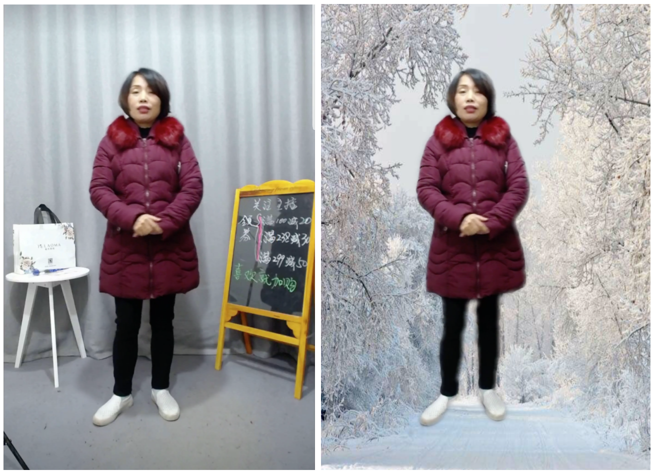
Abstract

This paper proposes an Online Intelligent Visual Interactive System (OIVIS), which can be applied to various short video and live streaming apps to provide an interactive user experience. During the live streaming scene, the anchor can issue various commands by using pre-defined gestures, and can trigger real-time background replacement to create an immersive atmosphere.To support such dynamic interactivity, we implemented algorithms including real-time gesture recognition and real-time video portrait segmentation, developed a lightweight deep neural network inference engine called MNN, and a real-time rendering framework AI Gender at the front-end to create a complete set of visual interaction solutions for using in resource constrained mobile.

1 Introduction

In recent years, short video apps like Instagram, Snapchat and Like apps are very popular around the world. In China, another similar kind of app is also getting popular: live streaming apps. Take taobao live for an example, the platform has empowered farmers, business owners and self-employed entrepreneurs to promote their specialties virtually face-to-face with interested buyers. Different from the traditional online shopping mode, the huge amount of traffic brought by the anchor can attract lots of viewers, thus completing the amazing volume of goods. Taking Li Jiaqi, a lipstick seller, as an example, as a taobao anchor, with huge traffic on the live platform, he can sell 15 thousand lipsticks in just 1 minute. He sold 320 thousand items with sales of 67 million during taobao's double eleven festival.

In ordinary online streaming apps, the anchor often needs to click the corresponding button on the screen and has to put down the items being displayed in hand when they are selling goods. The experience is not good for the anchor and the viewers. With the application of new technology including virtual reality and augmented reality, live-streaming sales would bring a more diversified interaction between online sellers and shoppers and generate more profits. In this paper, we demonstrate an Online Intelligent Visual Interaction System (OIVIS) to bring a novel interactive experience in live streaming apps. It uses the camera of a mobile terminal to develop a complete set of intelligent visual interactions. By collecting and processing the image/video information uploaded by a user, the virtual person controlled by an intelligent agent can provide immersive interactions and bring about a more realistic experience. When the anchor broadcasts the video, the smart background can be switched in real time by gestures. For example, the anchor can instantly switch the visual interaction scene into white snow scenes to create a realistic experience (Figure 1).



(a) Original frame. (b) Background switching.

Figure 1: Real-time portrait segmentation.

2 System Architecture

The goal of OIVIS is to provide interactive capacities for anchors and viewers in short-video and online-streaming apps, providing various functions such as gesture recognition, background replacement, etc. Different from Kinect, RealSence and other devices that use depth sensors, our system only requires the camera in mobile devices to achieve background switching in real-time. In order to ensure better user experience, we optimized the entire framework from the back-end to the front-end. We develop AI Engine which con-

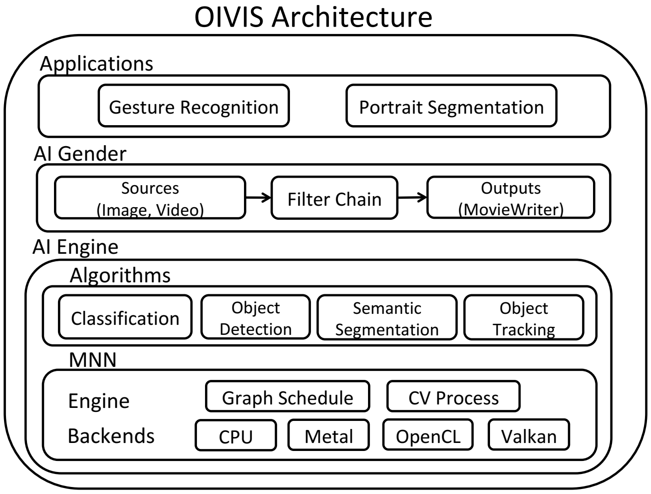


Figure 2: OIVIS architecture.

sists of deep model algorithms and a lightweight deep neural network inference engine called MNN which is specially optimized for embedded platform. We also built a multimedia cross-platform rendering framework called AI Gender. By leveraging the filter chain design, various filter effects can be combined efficiently. The whole system can run fluently even on the mid-range Android devices. Figure 2 shows the architecture of the OIVIS.

3 The AI Engine

The AI Engine is the most important part in OIVIS, which consists of deep model algorithms and a lightweight deep neural network inference engine called MNN. MNN is now open sourced and can be used for everyone, the performance is comparable and even better to tflite. The following mainly introduces gesture recognition and portrait semantic segmentation algorithms of the AI Engine.

3.1 Gesture Recognition

The gesture recognition problem is defined as a detection plus classification problem . Compared with traditional detection and classification tasks, the gesture recognition problem in the live broadcast scene has the following difficulties: 1) accurate identification of the gesture under motion blurring in video scenes; and 2) miniaturization of the model with fast response time.

For the object detection task, the one-stage model Single-Shot Multibox Detector (SSD) algorithm [Liu *et al.*, 2016] is fast, effective, and easy to expand, with performance well-balanced in all aspects. For real-time scenarios on mobile phones, we have optimized the backbone network based on SSD so that it can meet the system operation requirements on resource constrained devices. For gesture detection, the hand belongs to small object, so we use the Feature Pyramid Network(FPN) structure [Lin *et al.*, 2017] to optimize the small object detection. In the definition of the loss function, we choose sigmoid loss as the target loss, which is a multiple dichotomy problem. Our model is only 1.9MB, reaching 0.984 AP(0.5) on our test data, with an average inference time of only 17ms on IOS at input size 224*224.

3.2 Human Semantic Segmentation

For each frame in a given video, the portrait semantic segmentation task requires pixel-level classification based on different categories of objects. Compared with gesture recognition, this task is more complex. Both the data labeling cost and the real-time requirement are higher. To achieve this goal, we optimize both the data level and the model level.

To reduce data labeling cost, we create massive samples through image synthesis. In order to simulate the real data distribution, the color migration algorithm from [Reinhard *et al.*, 2001] is adopted to adjust the illumination. The person's position distribution statistics are used to him/her in a reasonable position. By artificially synthesizing high-quality data, we are able to obtain more segmented samples by an order of magnitude.

At the model level, we choose Mobilenet-v2 [Sandler *et al.*, 2018] as the backbone , use UNet's Encoder-Decoder structure [Ronneberger *et al.*, 2015] and DeepLab's ASPP structure [Chen *et al.*, 2018] to improve the accuracy. We analyze the timeline bottlenecks of the backbone, and perform model pruning to reduce the number of channels. Fast downsampling is used to reduce the size of feature maps as early as possible. Through the above modification, our model size is only 1.7MB, and the speed is 19ms/frame under Snapdragon 835CPU at input size 320*240.

In order to further improve the accuracy of our model, we adopt the knowledge distillation method [Hinton *et al.*, 2015] to migrate knowledge from the high-precision model to the real-time model by collecting a large amount of unlabel data. We used Xception-65 [Chollet, 2017] as the teacher model to increase the accuracy of the Mobilenet-v2 real-time model by 3 percentage points.

4 The Demonstration System

OIVIS provides interactive capacities such as gesture recognition, background replacement for anchors and viewers in short-video and online-streaming apps. By pre-defined gestures such as clenching, opening and sliding, the anchor can perform basic functions such as confirmation, return, swipe forward, and swipe backward through gestures in real time without having to click any button on the screen. In addition, we have developed a series of gesture-based interactions for fun use. For example, the anchor can play the game of scissors, paper and stone with fans online. Such fun use not only helps the anchor to bring more traffic, but also supports highly interactive user experience. In addition, through the opening of platform capabilities, we allow third parties to develop interesting and practical applications based on OIVIS platform, which are provided to the anchor or users in the form of plug-ins.

Acknowledgments

This research is supported, in part, by Nanyang Technological University, Nanyang Assistant Professorship (NAP).

References

[Chen *et al.*, 2018] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, 2018.

[Chollet, 2017] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1251–1258, 2017.

[Hinton *et al.*, 2015] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[Lin *et al.*, 2017] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2117–2125, 2017.

[Liu *et al.*, 2016] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European Conference on Computer Vision*, pages 21–37, 2016.

[Reinhard *et al.*, 2001] Erik Reinhard, Michael Adhikhmin, Bruce Gooch, and Peter Shirley. Color transfer between images. *IEEE Computer Graphics and Applications*, 21(5):34–41, 2001.

[Ronneberger *et al.*, 2015] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241, 2015.

[Sandler *et al.*, 2018] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, 2018.