Visual Question Answering

Prof. Li Hongsheng’s group has recently developed an AI program that can answer questions related to an input image. Such a challenging task is called Visual Question Answering (VQA). The AI program is required to understand and reason the contents of the image as well as the question to predict the correct answer. The developed AI program detects objects in the input image and infers the relationships between them. The cross-modal relations between objects in the image and words in the question are also automatically modeled by the AI program. The answers can be correctly predicted based on the cross-modal reasoning. The proposed AI program currently ranks 1st in the public VQA benchmark VQA 2.0.

Q: How many vehicles are in the street? A: 1
Q: Are there any people on the street? A: No
Q: Where will the bus go? A: School

Intelligent Vision Systems

Leading by Prof. LIN Dahua, the current AI research in the Multimedia Lab has three main themes. Firstly, by applying the technology of machine learning and deep learning, we aim to build an intelligent system that is able to extract useful information from our daily videos (bottom figure). For example, our system can automatically recognize the persons, objects as well as the mood and actions in videos. Secondly, our research endeavor also lies at the interaction of images and languages (top right). Human intelligence has a remarkable ability of using languages to describe the visual world; we would like to equip machines with a similar capacity. Lastly, we teach creativity and aesthetics to computers (top left). Our newly developed AI agent is able to create its own oil paintings.

Considering the Motion Speed and Orientation in 3D Human Action Recognition Using Two-stream CNN

Prof. LIU Yunhui’s group recently published paper about the 3D action recognition problem. Based on human skeleton data, the skeletal kinematics features are calculated first, i.e., linear velocity and orientation displacement, to capture the action variation along time. Then a novel image encoding method is introduced to encode the proposed kinematics features into images. Key frame selection scheme is also proposed to guarantee the same image input size. When training the CNN, the popular two-stream CNN architecture is adopted, incorporating spatial and temporal networks. The spatial ConvNet is trained on still RGB images, while the temporal ConvNet is trained on the proposed encoded kinematics features. This method is evaluated on a challenging multi-view dataset and the experiment results show that the proposed method is fast to train and outperforms many handcrafted features. (Published as conference paper “Kinematics Features for 3D Action Recognition Using Two-Stream CNN” on WCICA 2018)

A Robot for Baby-caring: Detecting the Dangerous Behaviors Conducted by Young Children in Daily Life

Prof. LIU Yunhui’s group recently developed a child caring robot, for detecting some dangerous behavior performed by child in the domestic environment based on the human action recognition and object recognition technologies. A human behavior is an interactive process between human and objects. Therefore, three factors need to be considered: the engaged objects, human actions and the relationship between human and the engaged objects. For the human action recognition, a new motion encoding is introduced and a convolutional neural network is utilized. Evaluation on the Northwestern-UCLA dataset verified the effectiveness of this method when action categories are small. The proposed action recognition method is simple and efficient, which is crucial for online behavior detection. Extensive experiments in the real physical world for detecting the behavior of eating allergic fruit and touching/playing electrical socket have achieved good performance. (Published as Conference paper “A child caring robot for the dangerous behavior detection based on the object recognition and human action recognition” on IEEE INTERNATIONAL CONFERENCE ON ROBOTICS AND BIOMIMETICS (ROBIO) 2018)