印第安納大學 David Crandall 教授演講
Egocentric Computer Vision, for Fun and for Science

文／孔啟熙 研究助理

David Crandall 教授在康乃爾大學 (Cornell University) 獲得博士學位後於印第安納大學 (Indiana University Bloomington) 任教並主導 Luddy Center for AI。David 研究以人為本且第一人稱的電腦視覺也包含導入觀察人類發展學習的機制進電腦視覺，其應用包含 AR/VR、自駕車、智能助理、。David 的研究也探討了人類在早年時，如嬰兒，和成人的第一人稱視覺上的差異，並探討為何嬰兒能快速學習而能否用其原理來改善電腦視覺。

David 介紹了第一人稱視覺的獨特性：與靜態資料集相比，第一人稱視覺可以捕捉到與環境互動的資料，例如透過移動來捕捉到感興趣的畫面。David 也介紹了嬰兒的視覺：透過人體實驗觀察發現了嬰兒配戴的相機可以捕捉到更動態的場景，例如嬰兒會更靠近物體而捕捉到更清晰且更大的物體畫面，更重要的是，嬰兒會透過移動身體和轉動物體以得到更多元的視角來觀察物體。David 的團隊將這樣的觀察導入物體偵測，重新設計了具有嬰兒視覺特性的資料集並發現其表現均較以成人視覺特性的資料集還好。

第一人稱視覺也面臨了許多獨特的挑戰，如資料收集的困難、相機劇烈的晃動、隱私等等。其中 David 的團隊更特別為了解決自動偵測第一人稱視覺隱私，而展開了 40 人的實驗觀察日常生活中哪些畫面是受測者最不想要被公開的畫面，並可藉此來訓練自動偵測隱私畫面的模型。

David 又與 Meta 合作參與了大型的第一人稱視覺資料集收集，Ego4D，由超過三千小時的日常活動影片資料集組成，其中包含了各種模態的資料，如聲音、慣性測儀、眼球偵測，以及標記了大量的物體以及文字描述。這樣的標準化大型第一人稱視覺資料集可以幫助推動後續相關的研究發展，就像物體檢測中的 MS COCO 資料集一樣。然而這樣大型的資料集難以保有完好的結構化，如影片活動分類，因此團隊接著展開另一個大型資料搜集 Ego-exo4D，專注於八種日常活動，如烹飪、彈奏樂器、修腳踏車等等，蒐集了人類在第一人稱視角如何做出需要技術的活動。

現今的機器學習距離人類學習的能力還有很大的差距，如人類可以透過少量的經驗有效率的學習一項技術、人類可以解釋做出決定的理由等等，以及第一人稱電腦視覺仍有許多尚未被探討的地方，然而 David 相信這是未來電腦視覺的一個重大領域也是 AI 通往發展更高智能的路徑。

Speech by Dr. David Crandall:
Egocentric Computer Vision, for Fun and for Science

Professor David Crandall completed his doctoral degree at Cornell University and then taught at Indiana University Bloomington where he directed the Luddy Center for AI. His research focuses on human-centered and first-person computational vision, which involves integrating mechanisms observed in human developmental learning into computer vision applications in areas such as AR/VR, autonomous vehicles, and intelligent assistants. Dr. Crandall's work also explores the differences in first-person vision between humans at different stages of life, including infants and adults. He investigates whether the principles underlying infants' rapid learning can be utilized to enhance computer vision.

During the presentation, Dr. Crandall highlighted the uniqueness of first-person vision. Unlike static datasets, first-person vision captures data of human interaction with the environment, such as capturing interesting scenes through movement. Additionally, he introduced the concept of infant vision. Through human experiments, they discovered that cameras worn by infants can capture more dynamic scenes. Infants tend to get closer to objects to capture more accurate and larger images. They also move their bodies and rotate objects to obtain diverse perspectives when observing objects. Dr. Crandall's team incorporated these observations into object detection, redesigning datasets with infant vision characteristics. They found that the model performance improved compared to those trained on datasets based on adult vision characteristics.

The first-person perspective presents several challenges that are not present in other perspectives, such as difficulties in data collection, intense camera shaking, and privacy concerns. To find the automatic detection solution to privacy concerns of first-person vision, Dr. Crandall's team conducted a particular experiment with 40 individuals to identify which scenes in their daily lives they would least want to be publicly disclosed. Meanwhile, the data collected in the experiment can also be used to train a model that could detect private scenes in first-person visual data.

Dr. Crandall has teamed up with Meta to participate in gathering a large-scale first-person visual dataset called Ego4D, comprising over three thousand hours of daily activity video data. This dataset includes various types of data such as sound, inertial sensors, eye tracking, and annotations of numerous objects along with textual descriptions. These standardized large-scale first-person visual datasets can drive further research developments, similar to the MS COCO dataset in object detection. However, maintaining such large datasets in a structured manner, especially when it comes to video activity categorization, can be challenging. Therefore, the team embarked on another large-scale data collection effort called Ego-exo4D, which focuses on eight types of daily activities like cooking, playing musical instruments, and fixing bicycles, aiming to gather data on how humans perform technically intricate activities from a first-person perspective.

Machine learning technology still has a long way to go before it can match human learning capabilities. Humans can efficiently learn a new skill with minimal experience and provide logical reasons for their decisions. Moreover, numerous aspects of first-person computer vision remain unexplored. Nevertheless, Dr. Crandall believes this field holds great potential for the future of computer vision and paves a promising avenue for AI to advance toward higher levels of intelligence.