Building computer vision in the kitchen

May 31 2024, by Alvin Lee

Credit: Pixabay/CC0 Public Domain

Imagine watching a pizza chef going about his work in a kitchen. You see him: weigh flour before adding water and yeast to it; knead the mixture into a dough; leave it to rise while he slices pepperoni and other toppings; stretch out the dough before assembling the pizza and sliding it into an oven.

While most people are unable to fluently execute the steps of pizza-making like an experienced chef, they can see and identify what was done. One could see the chef opening the flour sack and digging into it with a flour scoop, taking the pepperoni out of the fridge and putting it over the slicer repeatedly, or grating cheese with a box grater. At the end of it all, people understand that flour becomes dough, which in turn becomes pizza.

Can a computer vision software make the same connection?

Annotating for success

For SMU Assistant Professor of Computer Science Zhu Bin, the answer lies in the VISOR (VIdeo Segmentations and Object Relations), a dataset Professor Zhu and his collaborators have been working on.

By outlining certain objects such as hands, knives, flour scoops, graters, etc. and assigning identifying labels to them on first-person videos—also called egocentric videos—VISOR aims to: better identify separate objects; understand how hands and objects interact; achieve better reasoning and understanding of object transformation, such as flour becoming dough or a potato turning into fries.

This process of outlining and labeling objects is known as "annotation," and it can be achieved either via a "sparse mask" or a "dense mask."

"Sparse masks are annotations applied to select key frames within a video rather than every frame," explains Professor Zhu.

"These masks are curated to outline objects at significant moments or intervals in the video sequence. Dense masks are detailed, continuous pixel-level annotations that cover every frame in a segment of a video. In VISOR, these are often generated through interpolation between sparse

masks, using computer vision algorithms to fill in the gaps.

"Sparse masks are very useful for fine-grained egocentric video understanding, such as action recognition, e.g., 'chop potato,' and object state change. In contrast, dense annotations enable analysis of how objects are manipulated over time, providing insights into human-object interactions that sparse annotations alone could miss."

VISOR features over 10 million dense marks in 2.8 million images, and each annotated item has a mask that is assigned an entity class ("knife," 'fork," "plate," "cupboard," "onion," 'egg," etc.) and a macro-category ("cutlery," "appliance," "container," "vegetable," etc.). For instance, the entity classes "knife" and "fork" are classified into the macro-category "cutlery." All in all, VISOR features 1,477 labeled entities that identify and annotate many kitchen objects.

Other than identifying objects and annotating how items and human hands interact, VISOR also proposes a task called "Where did this come from?". In the case of the pizza chef, flour would be identified as coming from the flour sack. VISOR annotations cover videos with an average duration of 12 minutes, which is significantly longer than most existing datasets. This allows for an in-depth analysis and reasoning about object states over extended periods, facilitating studies on sustained interactions and changes.

Obstacles and future uses

Unlike many other datasets, such as UVO (Unidentified Video Objects) that focus on third-person perspectives, VISOR's use of egocentric videos from the EPIC-KITCHENS dataset presents extra challenges. Egocentric videos are dynamic by nature: objects often get blocked when hands move over items, and items transform as seen with the flour-to-dough-pizza example.

VISOR aims to overcome the obstacles in the following ways:

- Fine-grained egocentric video understanding: The object masks provided by VISOR clarify the boundaries of objects even through significant transformations. This precision enables the development of advanced deep models for analyzing fine-grained interactions and transformations within videos, such as egocentric action recognition and object state analysis.
- Enhancing interaction understanding: The detailed annotations of how hands interact with various objects help in studying and modeling human behavior, particularly in naturalistic settings like kitchens.
- Long-term video understanding: With continuous annotations across actions and transformations of objects (like an onion being peeled and cooked), VISOR supports research into long-term reasoning in videos, such as long-term object tracking.

"As the technology matures and technical challenges such as real-time processing are addressed, technology such as VISOR can be used to develop assistive technologies that help individuals with disabilities, or the elderly navigate and manage real-world tasks more independently," Professor Zhu tells the Office of Research.

"Robots equipped with the capability to understand complex object interactions and predict future actions can be employed in various activities, such as cooking, cleaning and manufacturing."

He adds, "Egocentric video understanding can also be used to develop virtual reality (VR)- or augmented reality (AR)-based training and educational tools, providing step-by-step guidance from the first-person view."