TEACHING COMPUTERS TO SEE

Statistics professor Song-Chu Zhu studies artificial intelligence to help computers understand one of the most challenging skills of all.

By Kim DeRose

As a university student, UCLA professor Song-Chun Zhu was inspired to study artificial intelligence when a computer became an unlikely competitor in a regional chess championship.

"For computers to play chess with humans," said Zhu, "sounded like fairy tales to me and inspiring."

Yet while humans and machines were on equal footing when it came to chess, Zhu realized a far greater challenge lay in teaching computers a skill that comes naturally to most humans: understanding how to see.

"Tasks involving symbolic reasoning such as chess are completely disconnected from reality," said Zhu, a professor with a joint appointment in the College's Department of Statistics and computer science in the school of engineering.

"Humans understand what they see without thinking," said Zhu. "To process images on a basic level requires a tremendous amount of common sense," Zhu said. "Robots can beat a chess champion, but they cannot think like a five-year-old child."

Processing images by humans and the challenges for machines to do the same thing are apparent in a casual glance at UCLA students during the lunch hour. An observer in Ackerman Union sees busy students sitting at

Song-Chung Zhu: "Dreams play a major role in human learning, and by creating software models based on real footage, we enable the computer to dream as well."

tables, talking animatedly while lugging textbook-laden backpacks, cell phones, and athletic gear. The human eye takes in the structured chaos without a hiccup, as neurons in the brain identify objects and places, easily sorting the students from the tables piled with notebooks and sandwiches.

"The human brain is a massive parallel machine that is more powerful than any computer in the world when it comes to understanding images," said Zhu, who is principal investigator on major grants from the Office of Naval Research, the National Science Foundation, and the Defense Advanced Research Projects Agency.

"More than 30 percent of the 100 billion neurons in the human brain are involved in visual processing, while a much smaller fraction is required when playing chess or studying for exams."

An electronic eye surveying the same scene would have much more difficulty; a table leg looks much like a human leg and a pencil could just as easily be a plastic straw or a chopstick. Zhu recognized that fledgling computer programs, not unlike young children, need a lot of guidance when it comes to identifying even simple scenes.

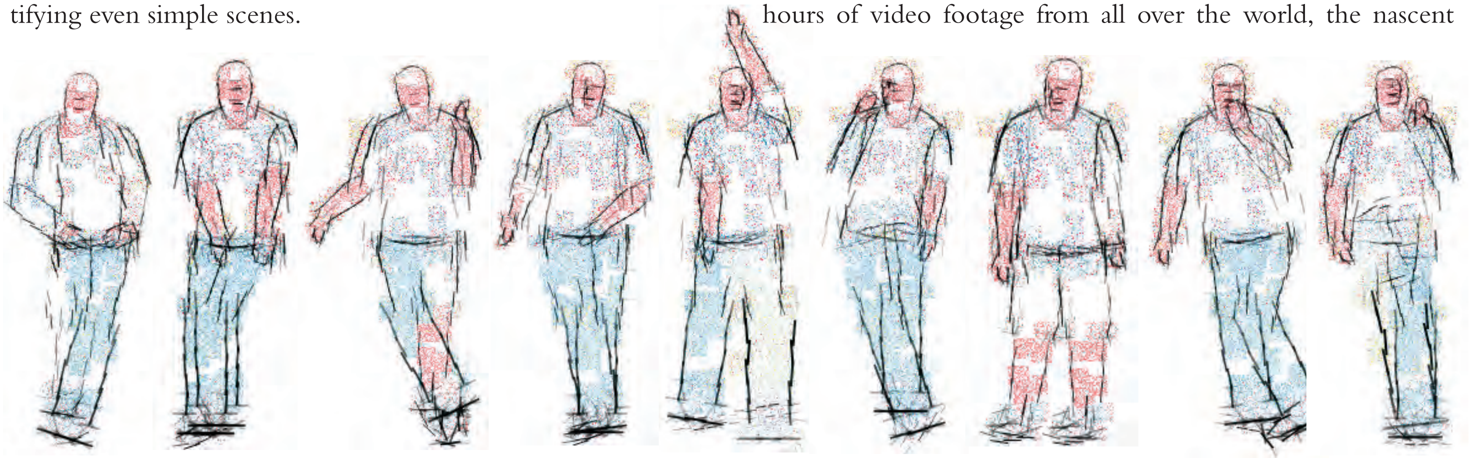
In 2005, Zhu founded the Lotus Hill Research Institute for Computer Vision and Information Science in Ezhou, China—its primary mission to collect and annotate images that help computers identify people, locations, and items from all angles. Breaking down each image into component parts that the computer can more easily digest, technicians have documented and categorized millions of photos: city streets, people, landmarks—even types of cupcakes.

While instructing a machine in the ways of human beings may seem a gargantuan task, the rewards could be just as monumental. A project in Zhu's laboratory titled "Image to Text (I2T)" builds a program to describe in words what it sees in an image or video. While still in its early phases, I2T has the potential to greatly simplify the monitoring of video observation by reducing the process down to a basic text search. With the prospects for analyzing footage in fields ranging from science to business to the military, there are simply not enough human eyes to watch it all, and I2T may one day fill that role.

Enlisting the vast database at its disposal in conjunction with advanced models of places and people, I2T is capable of identifying whether a location is a shopping mall or an office building through a process called 'scene categorization.' Zhu's most recent algorithms are even learning to recognize the gender, age, and social roles of nearby humans—an ability that may be critical for describing an emergency situation where, for example, a firefighter would be expected to behave differently from a bank teller.

Among the most difficult subjects for I2T to recognize are human beings and their behaviors. To the average electronic onlooker, UCLA students would appear to be veritable masters of disguise, obscuring their faces with hoods and sunglasses while sporting a variety of hairstyles and hats.

Much like a human toddler learns that a wall switch controls the illumination of a room without understanding the complexities of power grids or incandescent bulbs, Zhu hopes that his software will someday be able to infer similar connections between people and objects. By analyzing millions of hours of video footage from all over the world, the nascent computer program may even learn to statistically predict future actions in an unfolding event.



"Humans understand what they see without thinking. To process images on a basic level requires a tremendous amount of common sense. Robots can beat a chess champion, but they cannot think like a five-year-old child."

To make the detection of people more reliable, Zhu and colleagues recorded thousands of photos on campus to develop a virtual model of the UCLA student. The scientists confirmed they were on the right track by evaluating what Zhu calls a 'computer dream,' a digital representation of the computer's idea of what a human being should look like, formed by statistically sampling features in passersby. The imaginings are surprisingly lifelike *(see the images above)*, incorporating different clothing, expressions, and poses that signify a nuanced understanding of the human form that cannot easily be replicated.

"As humans, we can close our eyes and imagine many things because all that knowledge is stored in our brains," said Zhu. "Dreams play a major role in human learning, and by creating software models based on real footage, we enable the computer to dream as well." CR

www.stat.ucla.edu