Understanding Scenes and Events through Joint Parsing, Cognitive Reasoning and Lifelong Learning

PI: Dr. Song-Chun Zhu, University of California, Los Angeles
Other Universities: US: CMU, MIT, Stanford, UIUC, and Yale.
UK: Oxford, Birmingham, Glasgow, and Reading

Project Summary

Problem Statement The goal of this MURI team is to develop machines that have the following capabilities: i) Achieve deep understanding of scenes and events through joint parsing and cognitive reasoning about *appearance*, *geometry*, *functions*, *physics*, *causality*, *intents* and *belief* of agents, and use joint and long-ranged reasoning to fill the performance gap with human vision; ii) Represent visual knowledge in probabilistic compositional models across the spatial, temporal, and causal hierarchies augmented with rich relations, which are *task-oriented*, support efficient *task-dependent* inference from an agent's perspective, and preserve uncertainties; iii) Acquire massive visual commonsense through web scale continuous lifelong learning from heterogeneous sources through weakly supervised HCI and dialogue with humans; and iv) Understand human needs and values to interact with humans effectively and answer human queries about what, who, where, when, why and how in storylines through Turing tests.

Technical Approach We take a multi-disciplinary approach that integrates *four areas* illustrated by the four layers: (a) Psychology and cognitive experiments; (b) knowledge representation; (c) lifelong learning; and (d) computer vision tasks in an inference engine. The four areas will be studied in tight loops shown by the arrows in 3 colors. Human experiments bring new paradigms to transform computer vision and machine learning; and the latter drive human experiments to probe the brain mechanisms for representation, inference and learning. Each area consists of a spectrum of tasks which we organize in *three levels* of increasing depth and complexity: (i) Categorical recognition; (ii) joint parsing; (iii) Cognitive reasoning. We have assembled a multi-disciplinary team from both the US and UK to achieve our goals, including: Experimental psychology and cognition, computer vision and learning, Cognitive and mathematical modeling.

Outcome and Impact Our proposed research tackles challenges of pressing importance in the following DoD missions. 1) Persistent surveillance with ground and aerial sensors. Our approach will achieve deep understanding, generate narrative text descriptions, and answer human queries. 2) Web scale commonsense knowledge acquisition and information gathering. We will expand our lifelong learning engine to discover visual and common sense knowledge from heterogeneous sources on the web, organize them into deep structured knowledge on a unified graphical representation. This will fill the semantic gaps, overcome the shortcomings of machine learning approaches. 3) Robot autonomy and situation awareness. We will develop task-oriented representations and learning, so that machines understand the underlying physics, causal-effects, functionality of objects, the use of tools, and human intents, needs and values. This will improve robot autonomy in civil and military tasks in collaboration with humans.

Submitted to Dr. Behzad Kamgar-Parsi, Office of Navy Research
MURI Topic #14: Visual Commonsense for Scene Understanding
Requested fund US DoD $4.5M (3 yr base period) + $3.0M (2 yr option period) = $7.5M (total)
UK DoM $3.0M (3 yr base period) + $2.0M (2 yr option period) = $5.0M (total)