Vision and World Models

Alan Yuille

Bloonberg Distinguished Professor

Depts. Computer Science and Cognitive Science

What is Vision?

- Humans can extract an enormous amount of information from this image. *We can perform many visual tasks* -- recognize all these objects, estimate their 3D pose and other properties, their positions in the 3D world. **We can construct a 3D representation of the physical scene.**
- *We can combine this with common sense knowledge about* the physical/functional/social properties of objects, agents, and their interactions.
- We can answer questions like *“what accidents are waiting to happen in this picture?”*

Vision and the full AI problem

- Understand objects, scenes, and events. Describe them in language.
- Reasoning about functions and roles of objects, goals and intentions of agents, predicting the outcomes of events.
- World Models.

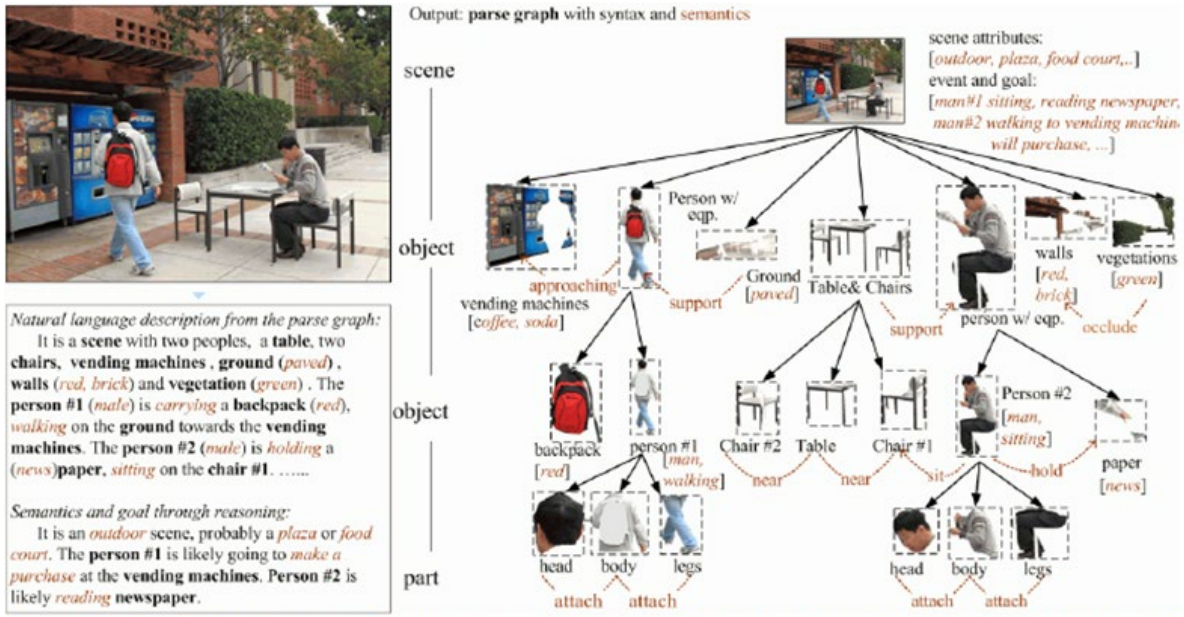


Figure 1. Example of image understanding. Analysis of the image (top-left) produces a parse graph (right) representing hierarchically objects, contextual relations, and semantic associations (in italic orange font) for attributes, functions, roles, and intents. The parse graph maybe converted to a description in natural language (bottom-left).

Vision Systems and knowledge of 3D World

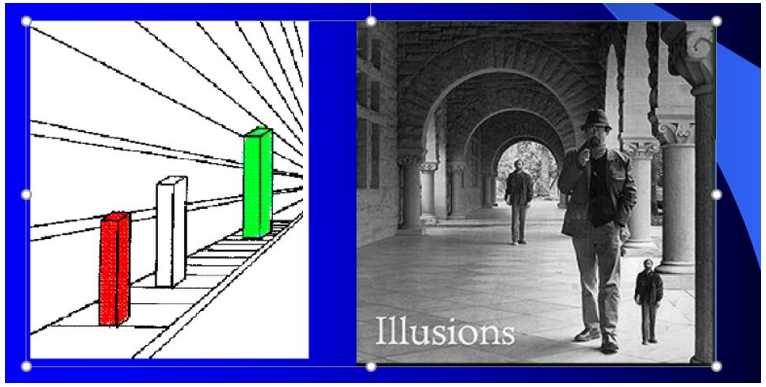
- Knowledge of the objects and scene structures in the world. Gibson's ecological constraints, Marr's natural constraints. Properties of typical visual environment and the 3D physical world.
- Intuitive Physics. How objects interact which each the environment and with each other.
- Human Activities. How humans interact with objects.
- Social Interactions. How humans interact with each other.
- World models consist of representations of the world together with prior knowledge of how these representations change with time including intuitive physics, human activities, human social interactions.

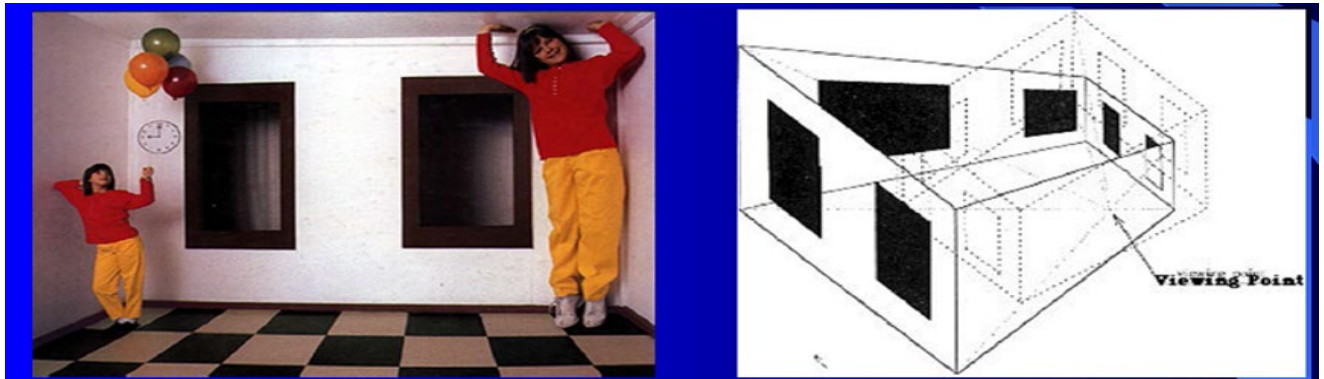
How do Humans Learn World Models?

- A baby is faced with a “buzzing, blooming confusion” of stimuli (visual, auditory, tactile, taste, smell) and has to make sense of it. The baby can perform limited actions, including crying, touching, tasting, etc (all with very limited control).
- Over time the infant learns that there is an external 3D world that generates all these stimuli. This 3D world consists of objects and agents (humans). These objects and agents obey physical laws (intuitive physics) and social constraints (video of social).
- Understanding this world enables infants to make predictions about what is likely to happen – if they drop an object, if they taste food that they know, if they cry, if they ask adults for something.
- This ability to take actions and predict their consequences.

Humans have knowledge of the 3D World

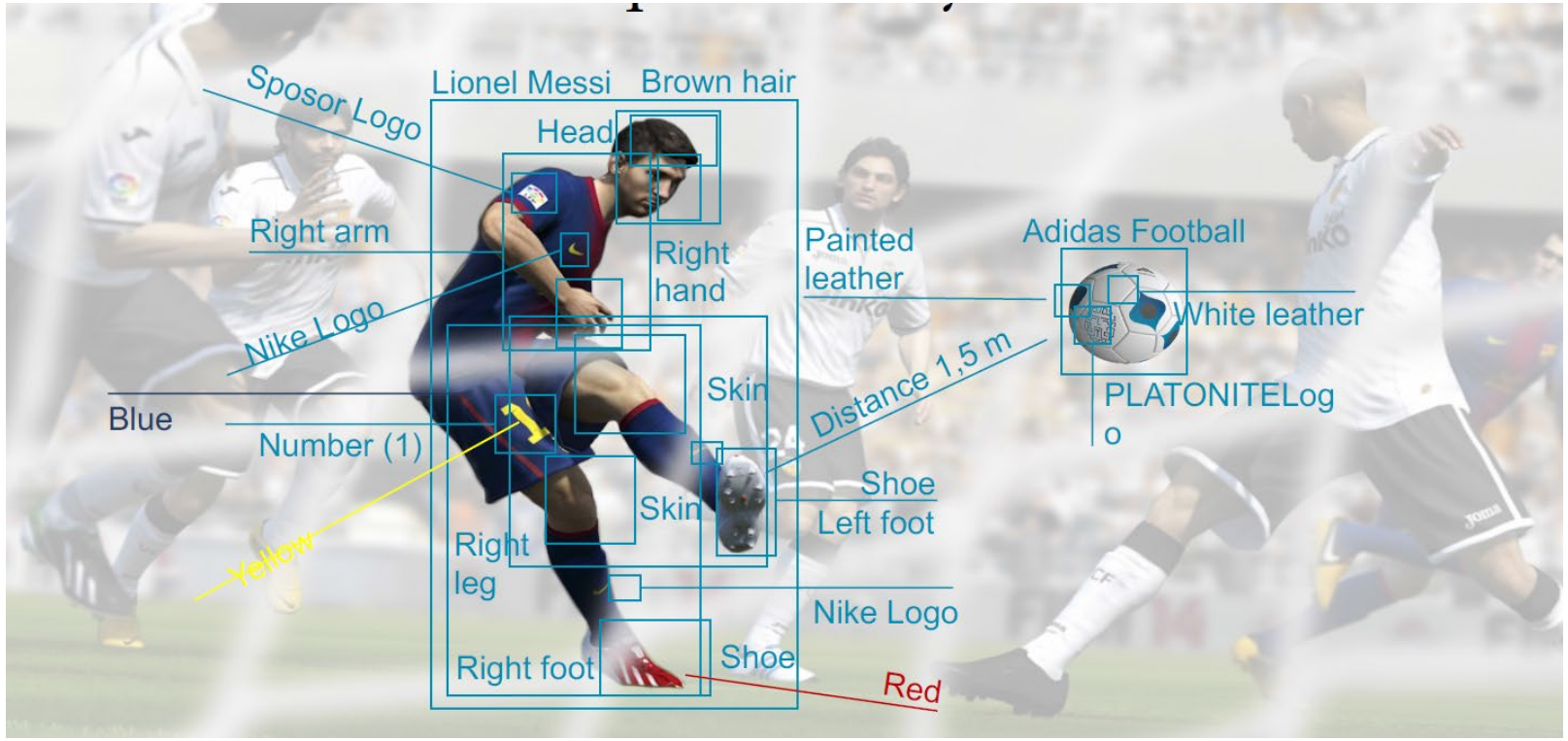
- The world is 3D. We interpret images assuming normal 3D structure (ground planes, shadows, and 3D structure), but can be fooled. *Easy to demonstrate but hard to prove.*





Humans have knowledge of parts.

- Part Examples (from von der Malsburg). Detecting and describing the footballer in terms of his parts is necessary for understanding his actions.



World Models unify Vision, Robotics, Language

- World Models are a common theme that includes all senses – vision, speech/audition, touch, taste, smell – and enable interactions with the world.
- This parallels how infants learn by combining information from all their senses and by interacting with the environment.
- From this perspective, the goal of computer vision is to construct rich 3D representations of the environment from images and image sequences.
- But this is problematic for vision because of the lack of suitable datasets. It requires realistic-synthetic datasets with 3D annotations, and mirrored real image datasets of virtual worlds. These have the capability of evaluating any visual tasks and can be extended to include language and robotics.