Recognizing Scenes by Simulating Implied Social Interaction Networks

MaryAnne Fields and Craig Lennon
Army Research Laboratory, Aberdeen, MD, USA

Christian Lebiere and Michael Martin
Carnegie Mellon University, Pittsburgh, PA, USA

To appear in the Proceedings of the 8th International Conference on Intelligent Robots and Applications (ICIRA 2015)

Exploiting Cognitive Context

OBJECTIVE & BENEFITS

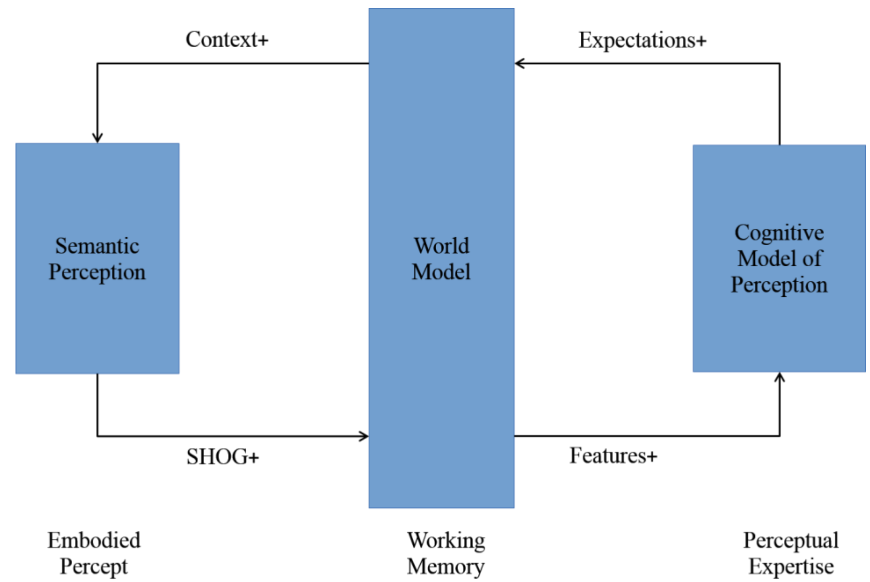
- Exploit cognitive context to augment bottom-up perceptual approaches
- Leverage activation mechanisms in ACT-R to provide contextual expectations
- Develop techniques to exchange information between cognitive and perceptual systems
- Benefits include improved object and scene recognition, and support for active perception

TECHNICAL APPROACH TO OVERCOME BARRIERS

- Establish a feedback loop between perceptual and cognitive systems via the World Model
- Encode Spatially-organized Hierarchical Object Graphs (SHOGs) from perceptual system
- Augment context via semantic priming in ACT-R
- Share contextual information from ACT-R with perceptual system

STATE OF THE ART & BARRIERS

- Perceptual systems tend to feed forward to cognitive systems that provide little feedback.
- General world knowledge, ontologies, goals and preceding cues create expectancies in ACT-R that have been used in perception as context for anticipating and resolving ambiguities about objects or scenes.
- Our main challenge is to provide the cognitive system with usable information based on the semantic label distributions for objects and regions generated by our perceptual approach.



Robot Readable World

by [Timo Arnall]

http://berglondon.com/blog/2012/02/06/robot-readable-world-the-film/

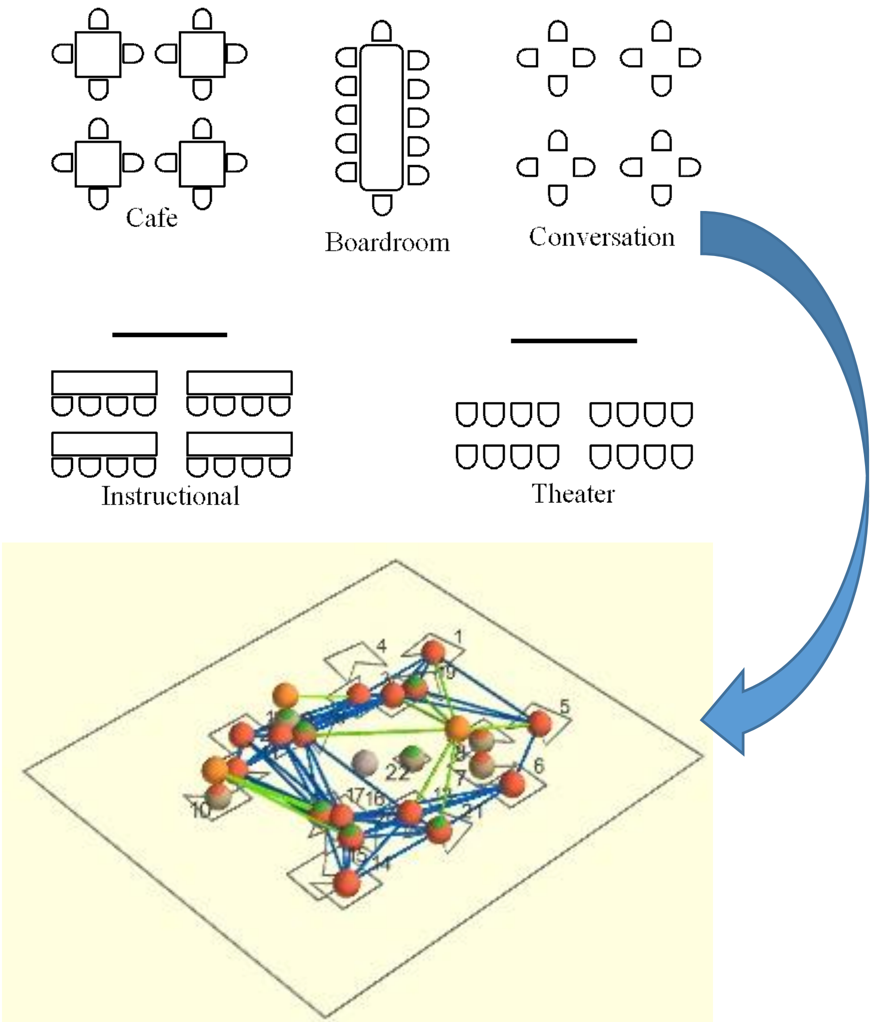
-- Embedded Video Removed (see URL above) --

ARL

Carnegie Mellon University

Public Spaces

Object recognition not usually sufficient for scene recognition. Configurations required to disambiguate.



- Developed room simulator to create notional SHOGs containing tables and chairs
  - These SHOGs code social affordances
  - Social immediacy operationalized in terms of object proximity and orientation
- Developed approach to encode semantic perception knowledge structures (SHOGs) to cognitive models
  - Instance-based learning in ACT-R
  - Global graph properties = scene gist
  - Local graph properties = exemplars of object in context (scene content, inter-object configuration, affordances, etc.)
  - Centrality guides order of object encoding (attention)
- Demonstrated utility of relational features in discriminating spaces with similar objects & similarity of KNN to ACT-R partial-matching and blending mechanisms (Fields, Lennon, Lebiere, & Martin, in press)

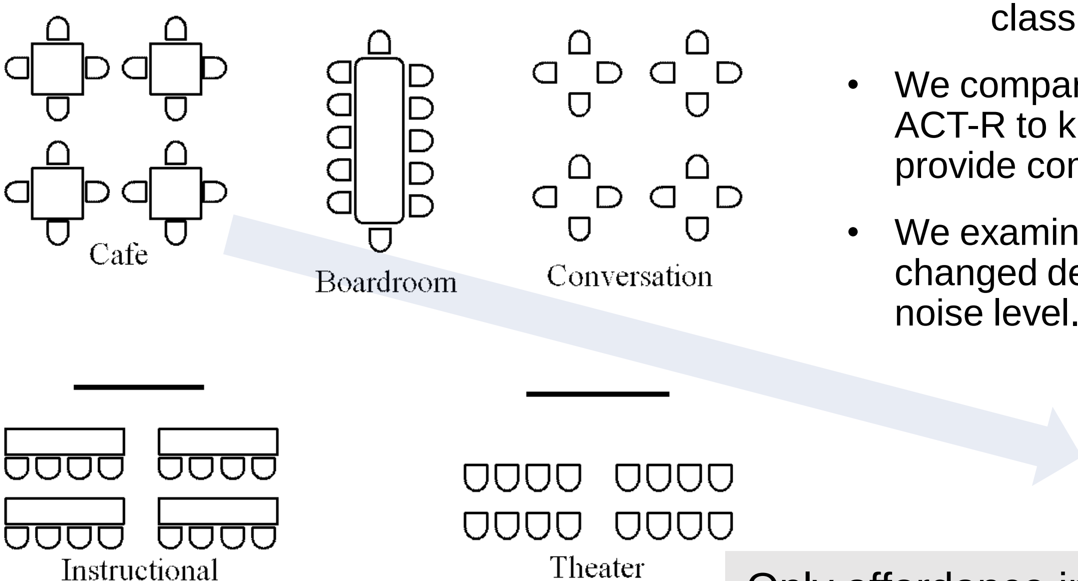
<u>**METRICS**</u>

- Confusions and error rates

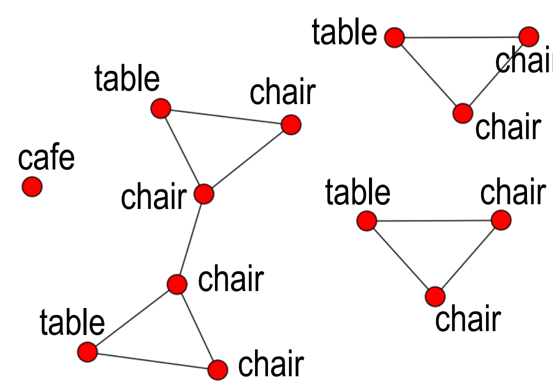
Scene Classification

- Indoor scene recognition remains challenging
- Current methods use object or parts recognition, along with the co-occurrence of salient features, to recognize interior scenes
- Rooms that contain collections of commonplace objects (e.g., tables & chairs) are vexing

- We tested a method to classify scenes based on how arrangements of constituent objects might impact social interactions
  - Chairs acted as surrogates for imagined humans so we could define social affordances based on spatial layout.
  - We compared the impact of affordance-based vs object-based features on room classification performance.
- We compared pattern-matching mechanisms in ACT-R to k-Nearest Neighbor classification to provide common ground.
- We examined how classifier performance changed depending on training set size and noise level.



Only affordance in this instance of a café is proximity

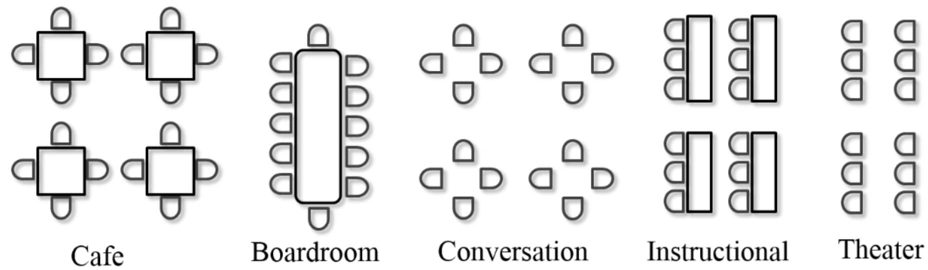


Experiment

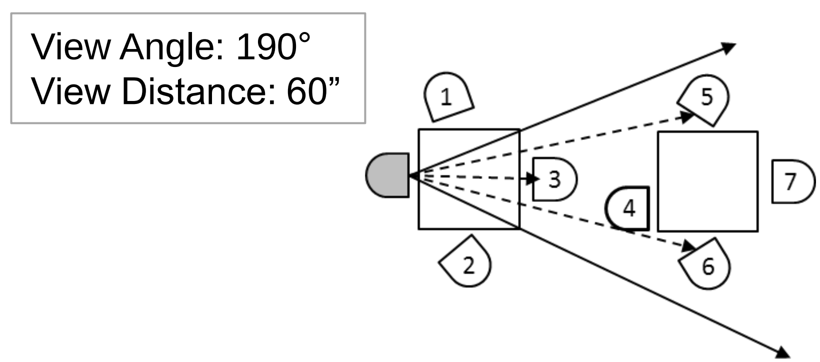
- We simulated <u>5 highly confusable room-types</u> (café, boardroom, conversation areas, instructional rooms, theaters)
- Canonical room-types (except boardroom) were populated with a <u>variable number of chairs and tables</u>, ranging from:
  - 2-4 rows
  - 2-4 sections within each row
  - 2-6 chairs grouped with 1 table (or focus point) within each section
- Each room was generated in a fashion that allowed testing of the robustness of classification to <u>2 levels of noise</u> (low, high) in:
  - social dynamics (chairs shifted and rotated from their canonical positions)
  - object identification (chairs mislabeled as tables or tables as chairs).

Noise Level	x (left/right)	y (front/back)	s	Labeling error
Low	[-6, 6] in.	[0, 6] in.	15°	0.05
High	[-12,12] in.	[0, 12] in.	45°	0.20

- We created 100 simulated rooms of each room-type x room-size combo for a total of 18,500 instances at each level of noise.



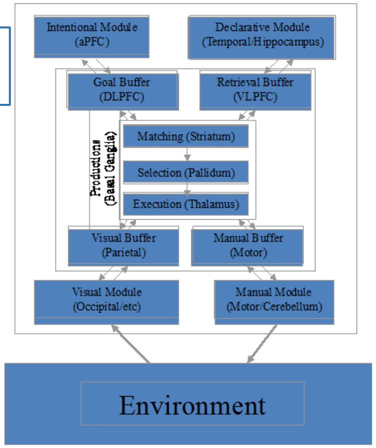
- We created <u>2 feature sets</u>
  - object-based – node counts
    - chairs
    - tables
  - affordance-based – binary link counts
    - proximity edges (60”)
    - mutual visibility edges (potential “eye-contact” based on orientation)



- Classifier robustness was further tested across <u>3 training set sizes</u> (1%, 10%, 100%)

Classifiers

Subtle interplay of environment, an agent's relevant knowledge and the agent's goals → Context



Some mechanisms underlying this interplay are inherent part of ACT-R
Similar to ML techniques, but integrated in a unified cognitive architecture

KNN

- Requires training set with quantitative features, associated labels, and a similarity metric (Euclidean distance in this case).
- Assumes feature space is continuous enough that a point w/in it is likely to have same label as points near it.
- Classifies new observations according to modal label of <u>K closest training set points</u>.
- We set neighborhood size of k = 1, 5, & 10 (for the 1, 10 and 100% training sets, respectively)

ACT-R

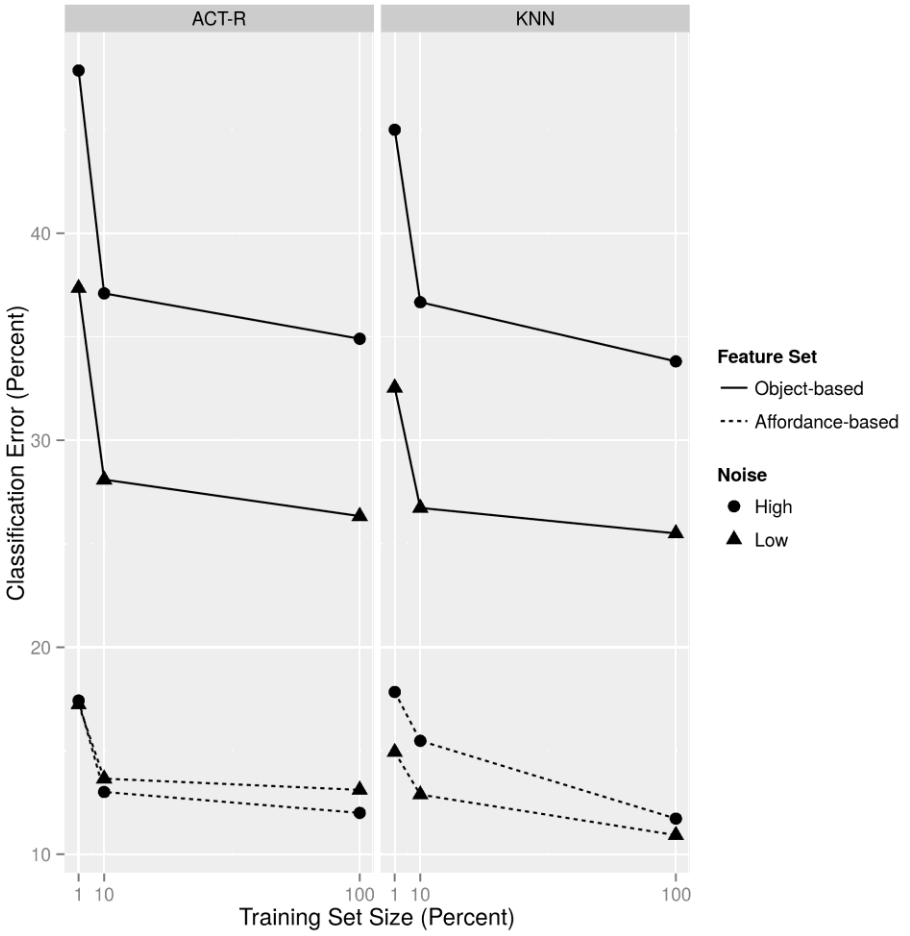
- Classification based on retrieval of knowledge patterns (chunks) from declarative memory
  - Chunks are data structures associating small sets of data items
  - Retrieval governed by statistical quantities reflecting history, associations, similarities.
  - Classification reflects <u>entire training set</u>

ACT-R Mechanisms

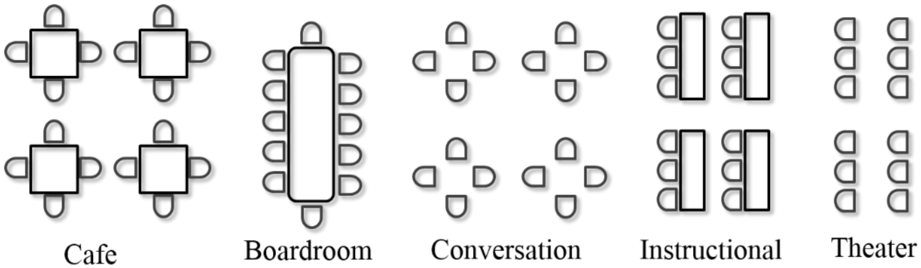
- Activation: sum of Bayesian factors reflecting chunk's content & history of use:
  - $A_i = \log \sum_k t_k^{-d} + \sum_j W_j S_{ji} + N(0,\sigma)$
  - Base-level = prior history
  - Spreading activation = current context
  - Noise = stochastic retrieval process
- Retrieval process
  - Specify situation as pattern in retrieval buffer
  - Compute match score for all chunks in DM
    - $M_i = A_i + MP \sum_d sim(v,d)$
    - Similarity: $sim(x,y) = \frac{min(x,y)}{max(x,y)} - 1$
  - Return consensus value by blending
    - $V = argmin \sum_i P_i \left(sim(V,V_i)\right)^2$
    - $P_i = \frac{e^{M_i/t}}{\sum_j e^{M_j/t}}$

Classification Error

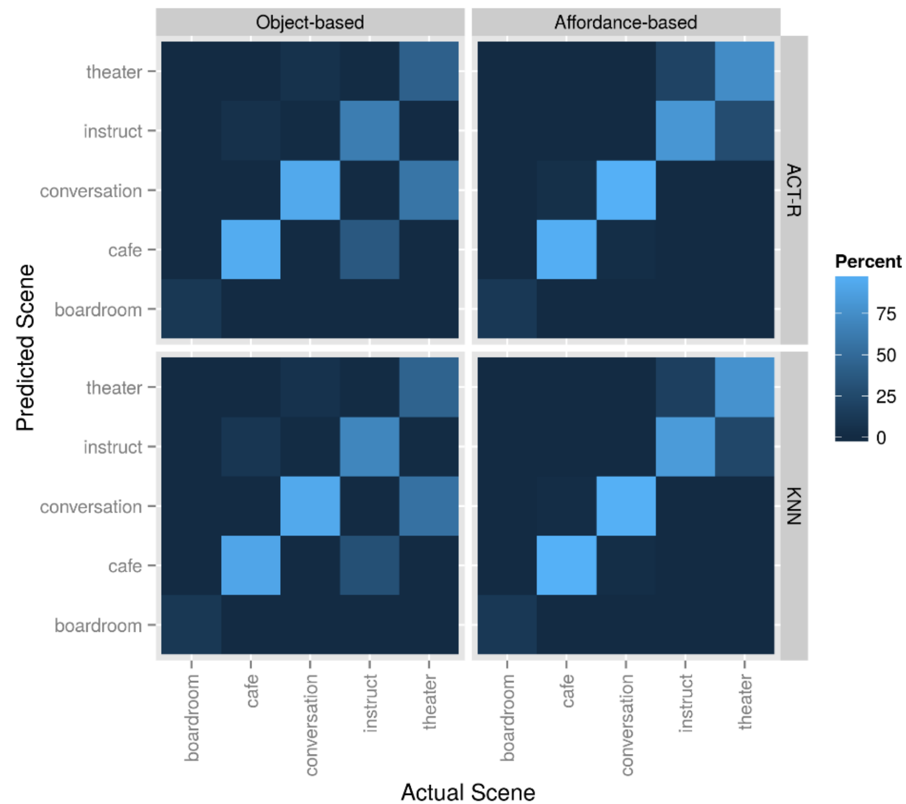
- Both classifiers (KNN, ACT-R) recognized rooms more accurately by using affordance-based features rather than object-based features
- Both classifiers responded similarly to the degree of noise present in the stimuli (high, low) especially for the object-based features
  - Low noise stimuli tended to reduce classification errors relative to high noise stimuli.
  - However, for affordance-based features <u>high noise improves performance marginally in the ACT-R classifier while still decreasing it slightly for KNN</u>.
- Both classifiers were robust to decreases in training-set size (1%, 10%, 100%).
  - They performed best with full sampling (i.e., 100%).
  - Performance at 10% sampling was nearly as good.



Confusions



- A close similarity in classifier performance can be seen in confusion patterns, too.
- Social affordances were more effective than pure object-based feature sets
- Confusion pairs for object-based features
  - Theater/conversation → no tables
  - Instructional/café → tables
- Confusion pairs for affordance-based features
  - Theater/instructional → same social structure except for tables
  - Café/conversation → same social structure except for tables
- Boardroom is not very confusable in either feature set because of its unique structure



Room-type confusions for each classifier for full sampling with low noise.

Classifier Comparison

- Similarities between ACT-R memory retrieval and KNN.
  - Each chunk in declarative memory corresponds to a training instance.
  - The partial matching mechanism is akin to the distance computation in KNN
  - Blending and KNN classify by summing over instances
- Differences between ACT-R Model & KNN Algorithm
  - Ratio similarity vs linear distance
  - Manhattan distance vs Euclidean distance
- ACT-R memory retrieval is more general than the KNN voting process
  - ACT-R activation equation captures recency, frequency, and semantic priming effects
  - Blending operates over all instances in memory rather than the most similar K of them
    - Broadens the experience base upon which the decision is made
    - Removes the need for modelers to specify a proper value for the K parameter
  - More similar examples have a higher impact than more distant ones because of weighting term
  - Process of aggregating answers in blending is more general than KNN voting process
    - Can also average over values for which similarity functions are defined (e.g., numbers)
    - Can find consensus values among symbolic chunks for which similarities are defined.
- Embedding generalizations of machine learning algorithms such as KNN, RL, and Bayesian Learning in cognitive architectures enables them to be integrated with other cognitive mechanisms.
  - Flexible ways of reflecting cognitive context in perception and decision making
  - Leverage knowledge about the semantics of the domain

Next Steps

- Revise room simulator to include perceptual errors and metric info in notional SHOGs
  - Mislabeled, missing, hallucinated
  - Metric distances, sizes, orientations
- Incorporate incremental perception
- Incorporate recency, frequency, and semantic priming effects
- Map SHOG network properties to Gestalt principles where possible
- Explore feature sets co-developed with the perceptual system
- Integrate with semantic perception algorithms