Grammars and Learning for Scene Classification: A Survey

Hemerson Pistori, Andrew Calway
and Peter Flach

SUN Database: Large-scale Scene Recognition from Abbey to Zoo

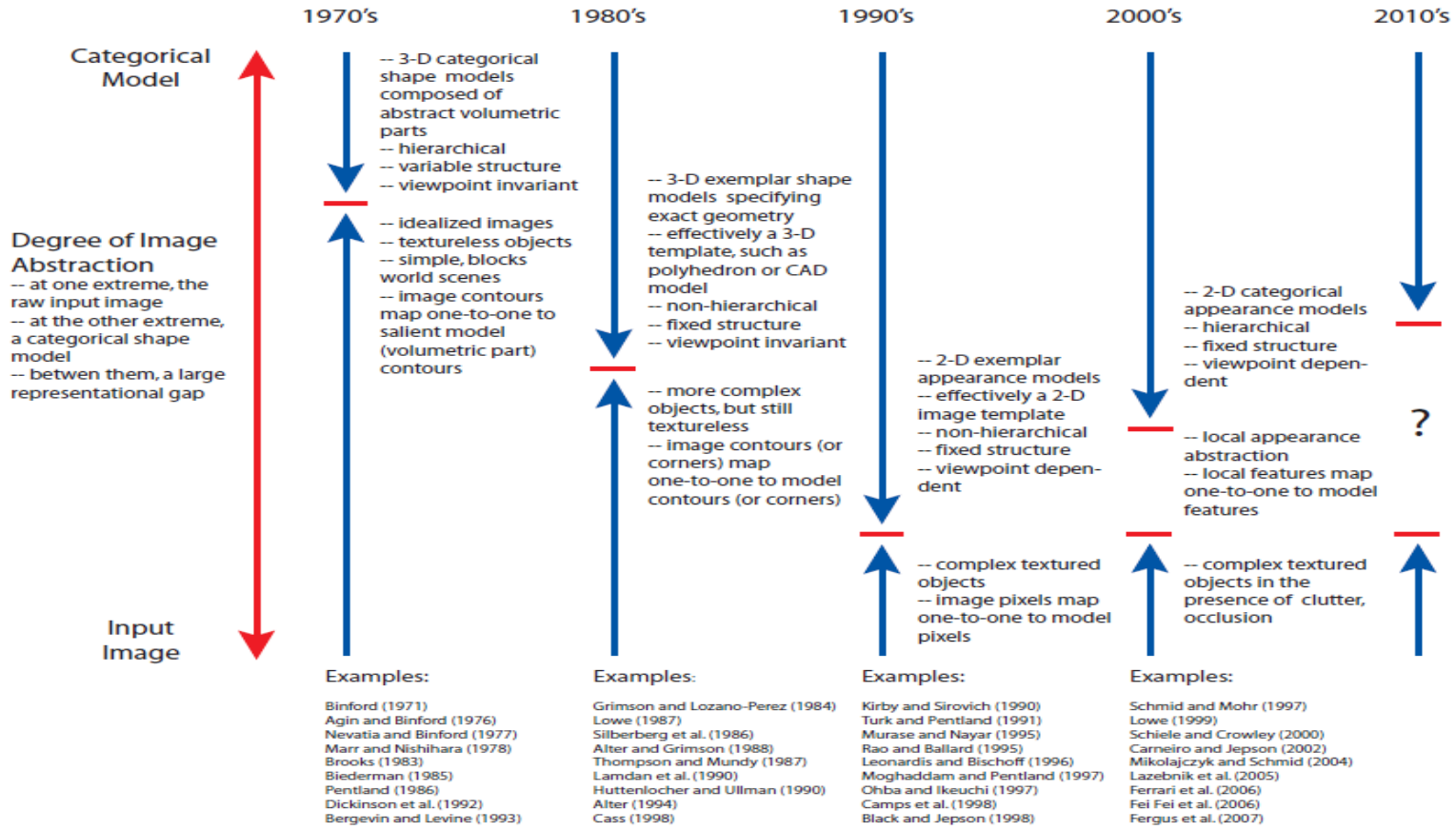
Jianxiong Xiao – MIT and Brown University

- First large scale image database for scene recognition
- 130,519 images and 899 categories
- Evaluation of state-of-the-art algorithms and human performance
- Introduces the scene detection problem (in multiple scenes images)

The Evolution of Object Categorization and the Challenge of Image Abstraction

Sven Dickinson – University of Toronto

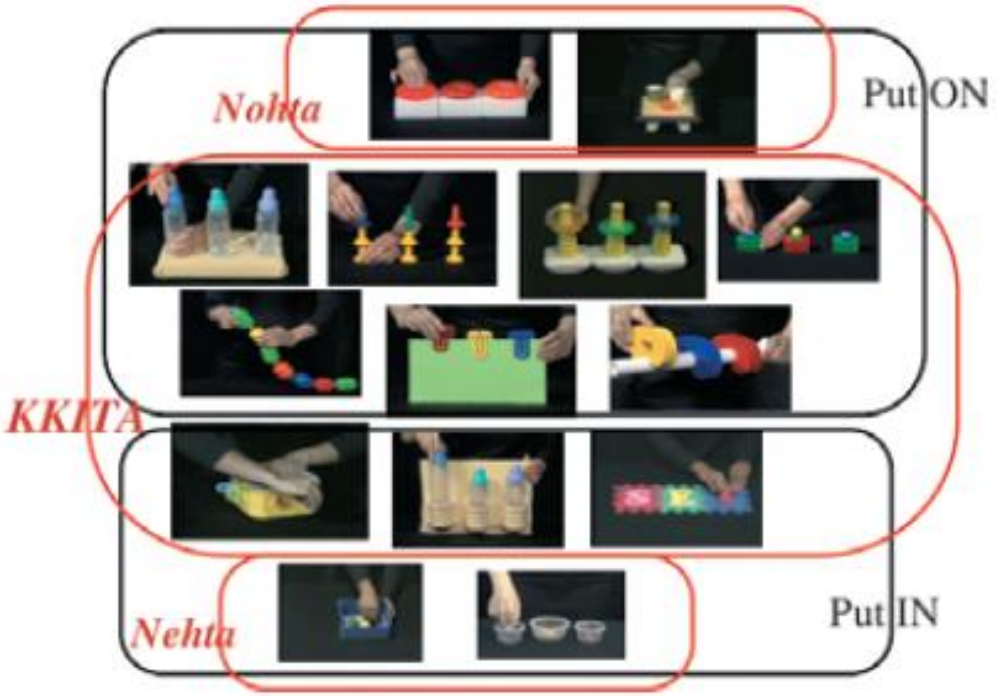
- History of object categorization and degree of image abstraction
- Need for new image abstraction mechanisms
- Large survey (275 refs)



Relative Contribution of Perception/Cognition and Language on Spatial Categorization

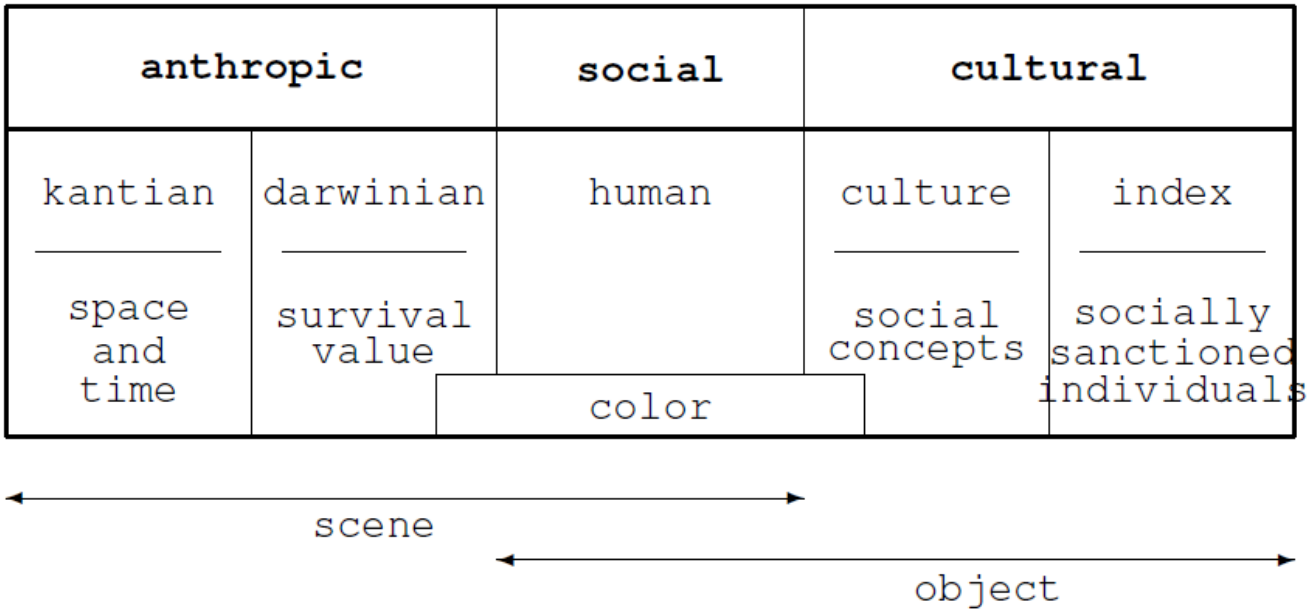
Soonja Choi and Kate Hattrup – San Diego State University

- Both universal cognition/perception and language-specific semantics guide spatial categorization
- Use of Universal or Language specific resources depends on type and context
- Experiments with Koreans and English speakers



On the Definition of Categories for Image Classification Evaluation
Simone Santini

- Semiological level of signification (portion of meaning that does not depend on context)
- Coherent set of categories for image analysis evaluation
- Semantic axis:
  - **Anthropic**: openness, expansion, transience, concealment, navigability, fractality;
  - **Social**: presence of people, color;
  - **Cultural**: "there are so many social forces that influence the connotations of a picture that isolating a semiological level might be impossible"



A Survey of Grammatical Inference Methods for Natural Language Learning
Arianna D'Ulizia et alii – CNR, Italy

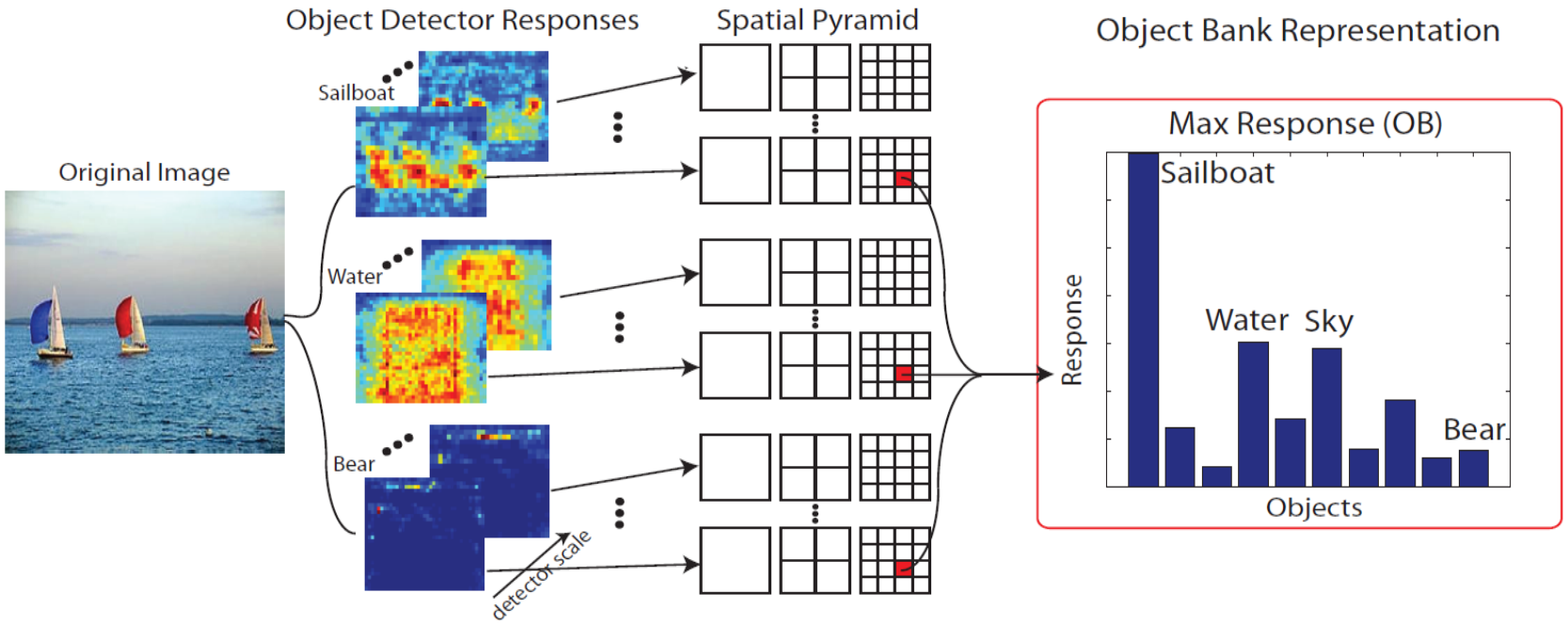
- Context Free Grammars Only
- Literature review only

	Presentation set		Type of information		
	Text	Informant	Supervised	Unsupervised	Semi-supervised
ADIOS	X			X	
EMILE		X	X		
e-GRIDS	X			X	
CLL	X			X	
CDC	X			X	
INDUCTIVE CYK		X		X	
LAgtS	X			X	
GA-based		X	X		
ALLis	X		X		
ABL	X			X	
UnsuParse	X			X	
Incremental parsing	X			X	
Self-training	X				X
Co-training	X				X

Object Bank: A High-Level Image Representation for Scene Classification & Semantic Feature Sparsification

Li-Jia Li et alii – Stanford and CMU

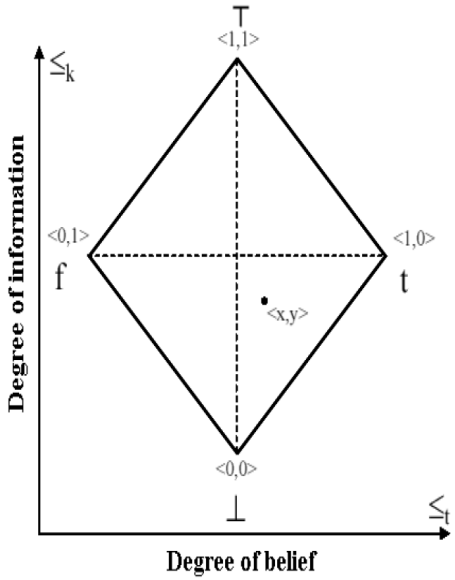
- Visual knowledge readily available on the Internet (LabelMe and ImageNet)
- Features: Responses from object sensing filters built on a generic collection of labeled objects
- Object detector of Felzenszwalb et al.
- 200 most frequenty objects in Labelme, Imagenet, ESP and Flickr



Predicate Logic based Image Grammars for Complex Pattern Recognition

Vinay Shet et alii – Siemens and University of Maryland

- Bilattice based Logical Reasoning Framework
- Pedestrians and Surface to Air Missile Site detection
- Contextual Cues, Scene Geometry and Object Pattern Constraints as Rules in a Logic Programming Language
- Pattern Grammar as logical rules over predefined atoms
- Rules are manually encoded and designed to facilitate scalability
- Parts Detectors (E.g.: Human Parts) get the facts or evidences
- Rule weight optimization using KB-NN



Assume the following set of rules and facts:

Rules	Facts
$\phi(human(X,Y,S) \leftarrow head(X,Y,S)) = \langle 0.40, 0.60 \rangle$	$\phi(head(25,95,0.9)) = \langle 0.90, 0.10 \rangle$
$\phi(human(X,Y,S) \leftarrow torso(X,Y,S)) = \langle 0.30, 0.70 \rangle$	$\phi(torso(25,95,0.9)) = \langle 0.70, 0.30 \rangle$
$\phi(\neg human(X,Y,S) \leftarrow \neg scene\_consistent(X,Y,S)) = \langle 0.90, 0.10 \rangle$	$\phi(\neg scene\_consistent(25,95,0.9)) = \langle 0.80, 0.20 \rangle$

Inference is performed as follows:

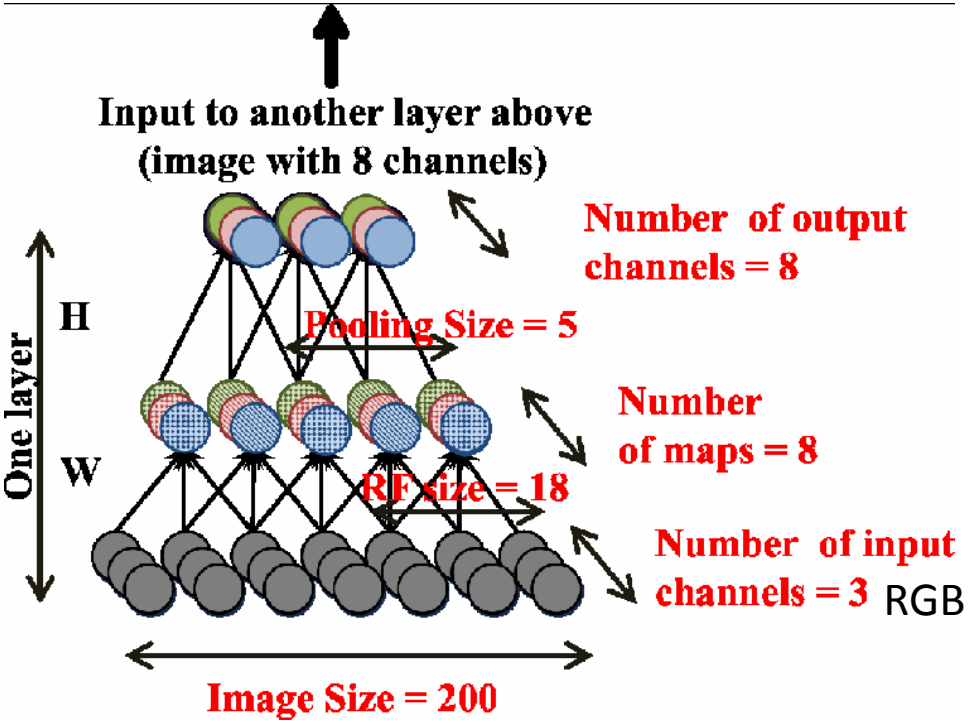
$$cl(\phi)(human(25,95,0.9)) = \langle 0,0 \rangle \vee [\langle 0.4,0.6 \rangle \wedge \langle 0.9,0.1 \rangle] \oplus \langle 0,0 \rangle \vee [\langle 0.3,0.7 \rangle \wedge \langle 0.7,0.3 \rangle] \oplus \neg(\langle 0,0 \rangle \vee [\langle 0.9,0.1 \rangle \wedge \langle 0.8,0.2 \rangle])$$
$$= \langle 0.36,0 \rangle \oplus \langle 0.21,0 \rangle \oplus \neg \langle 0.72,0 \rangle = \langle 0.4944,0 \rangle \oplus \langle 0,0.72 \rangle = \langle 0.4944,0.72 \rangle$$

Building high-level features using large scale unsupervised learning
Quoc V. Le et alii – Stanford

- 10 million 200x200 image frames from youtube videos
- 17% - percentage of faces in dataset (frames containing faces or area covered by faces ?)
- 1000 machine during one week
- Grandmother cell
- Deep autoencoder with sparsity
- Local receptive fields (RF)
- Learning: Topographic ICA
- 6 Sublayers (3 layers)
- Experiment: selection of best

neuron and thresholds uses

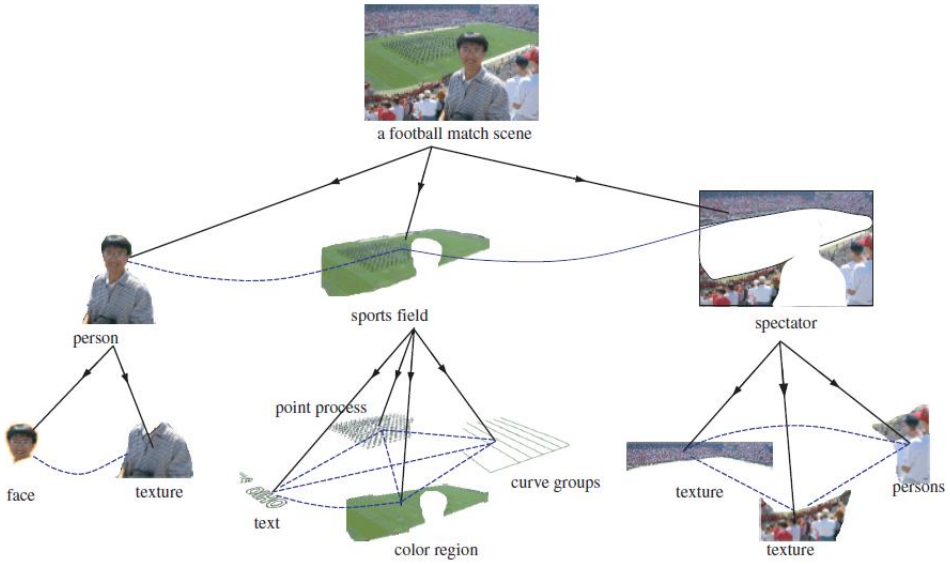
classif. Accuracy (supervised)

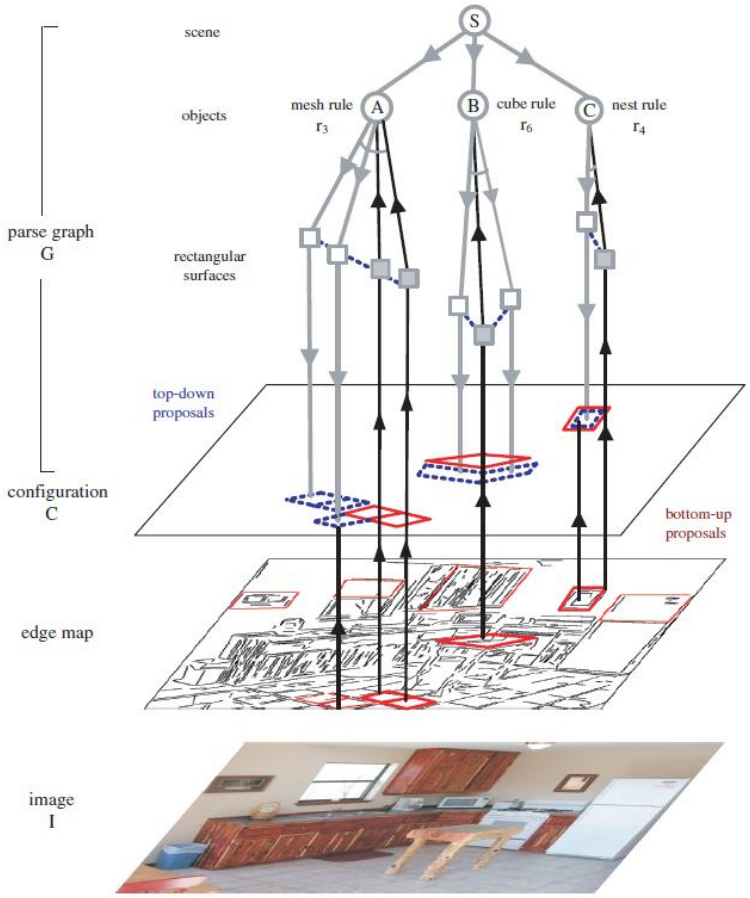


A Stochastic Grammar of Images

Song-Chun Zhu et alii - University of California

- Visual dictionaries and and-or graph composition to fill semantic gap (symbols and raw signals)
- Stochastic grammars
- Combine bottom-up and top-down procedures
- Graphical models (MRF)
- Semi-automatic parsed images: http://www.imageparsing.com/

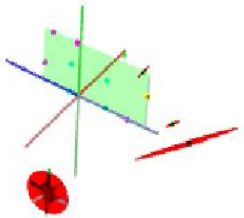


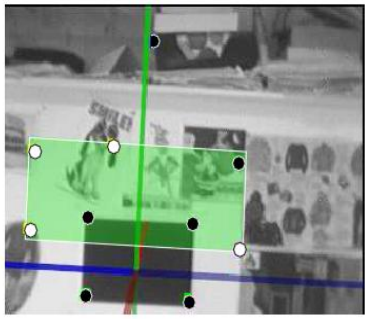


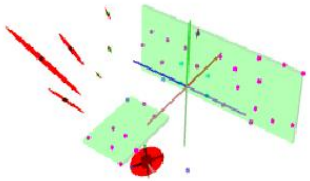
Discovering Higher Level Structure in Visual SLAM

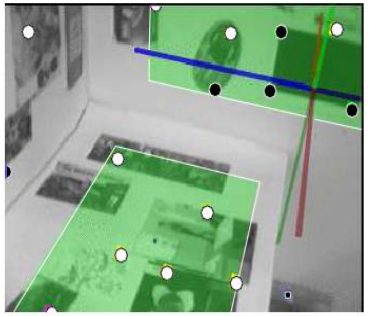
Andrew P. Gee et alii – University of Bristol

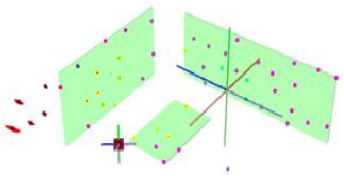
- Discovery of planes and lines
- Predictive filtering (EKF)
- Dynamically changes in state size
- Combines points (or edglets) with planes and lines

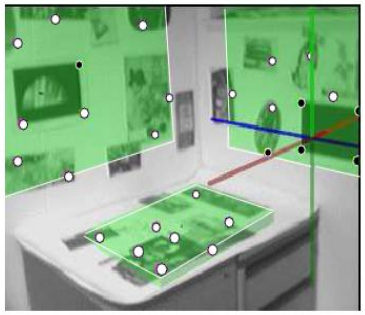












FAB-MAP: Appearance-Based Place Recognition and Mapping using a Learned Visual Vocabulary Model

Mark Cummins and Paul Newman – University of Oxford

- Uses bag-of-words (BoW)
- Tree-Structured Bayesian Network and Recursive Bayesian Filtering in place of Term Frequency – Inverse Document Frequency (TF-IDF)
- Chow Liu algorithm
- Tackles perceptual Aliasing and Variability using a Probabilistic Model on top of BoW

(a) Perceptual Variability

(b) Perceptual Aliasing

Same Place

Different Place

Green: Common words
Red: Others

Visual Synset: Towards a Higher-level Visual Representation

Yan-Tao Zheng et alii – National University of Singapore and Google Inc. USA

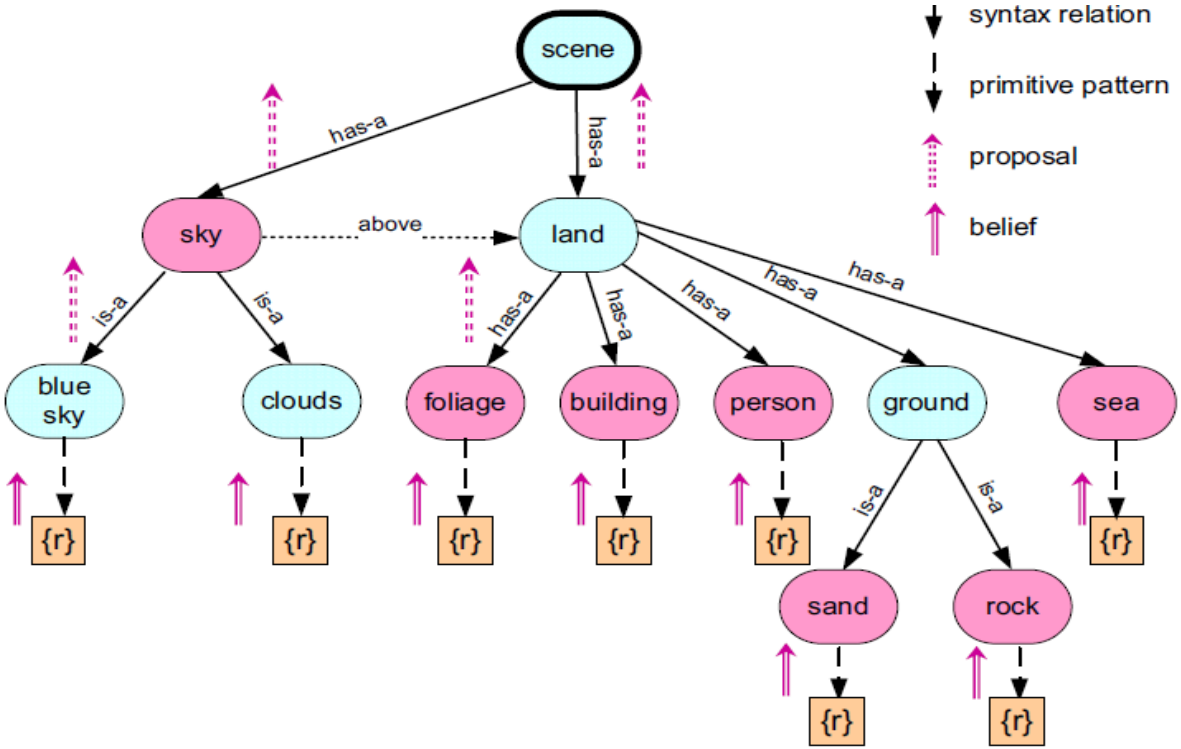
- Extends bag-of-words
- Polysemy: delta visual phrases (co-occurrance + spatial inf.) - Frequent item-set mining (FIM)
- Synonymy: visual synsets (synonymy sets) – supervised learning – labeled image classes (“semantic”) – Information Bootleneck Principle
- Visual lexicon = delta visual phrases and visual words

Syntactic Image Parsing using Ontology and Semantic Description

Ifeoma Nwogu et alii – University of Rochester and University of Buffalo

---

- Formal ontology for natural outdoor scenes (Mereological Relationships and First Order Logic)
- Image grammar based in F. S. Ku proposal (1970)
- Does not present the connexion between the Grammar formalism and First Order Logic or between the Grammar and the Parsing Strategy
- Low-level: MRF Graph of Superpixels (over segmentation)



Multi-modal Semantic Place Classification

A. Pronobis et alii – Royal Inst. of Technology, Sweden

- Combination of clues: local (SIFT), global (CRFH) visual features and laser scan
- SVM in two levels (each clue, resulting scores)
- Odometry information to help “Semantic Labeling” (give a name to the place where the robot is ?)

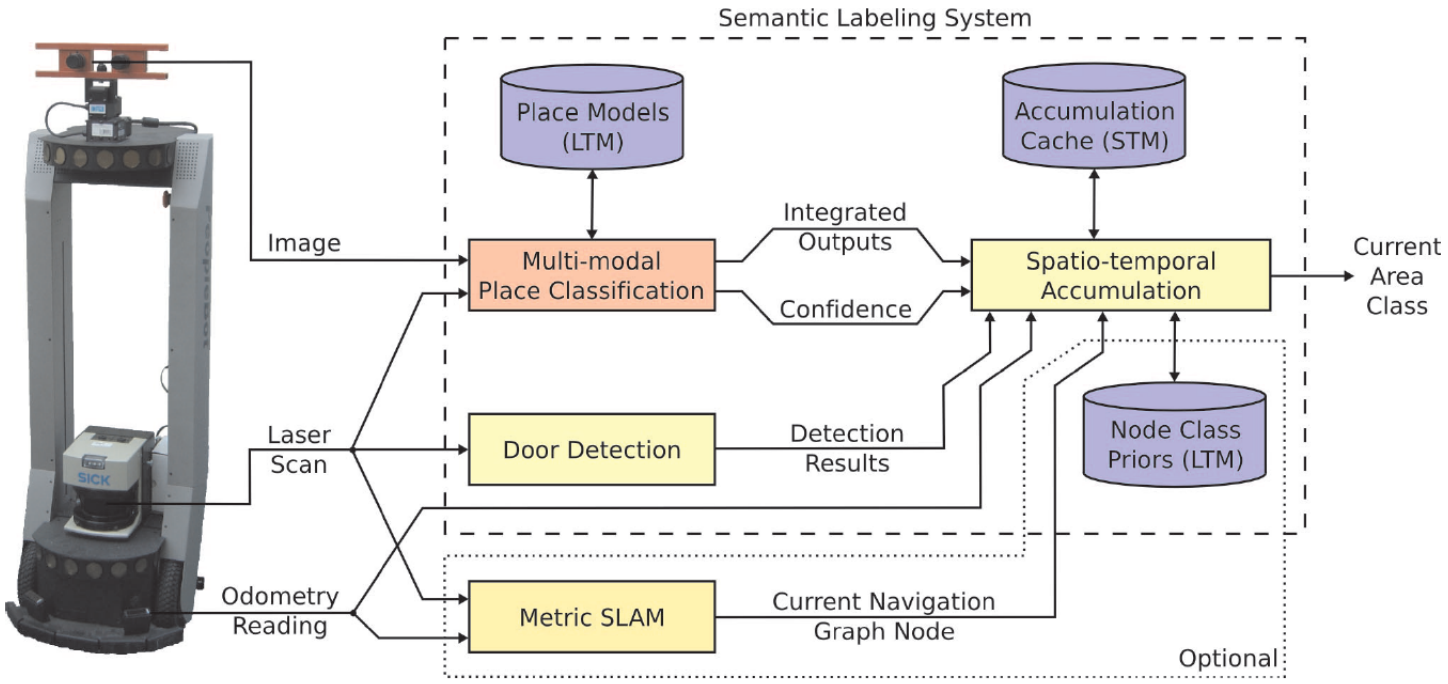
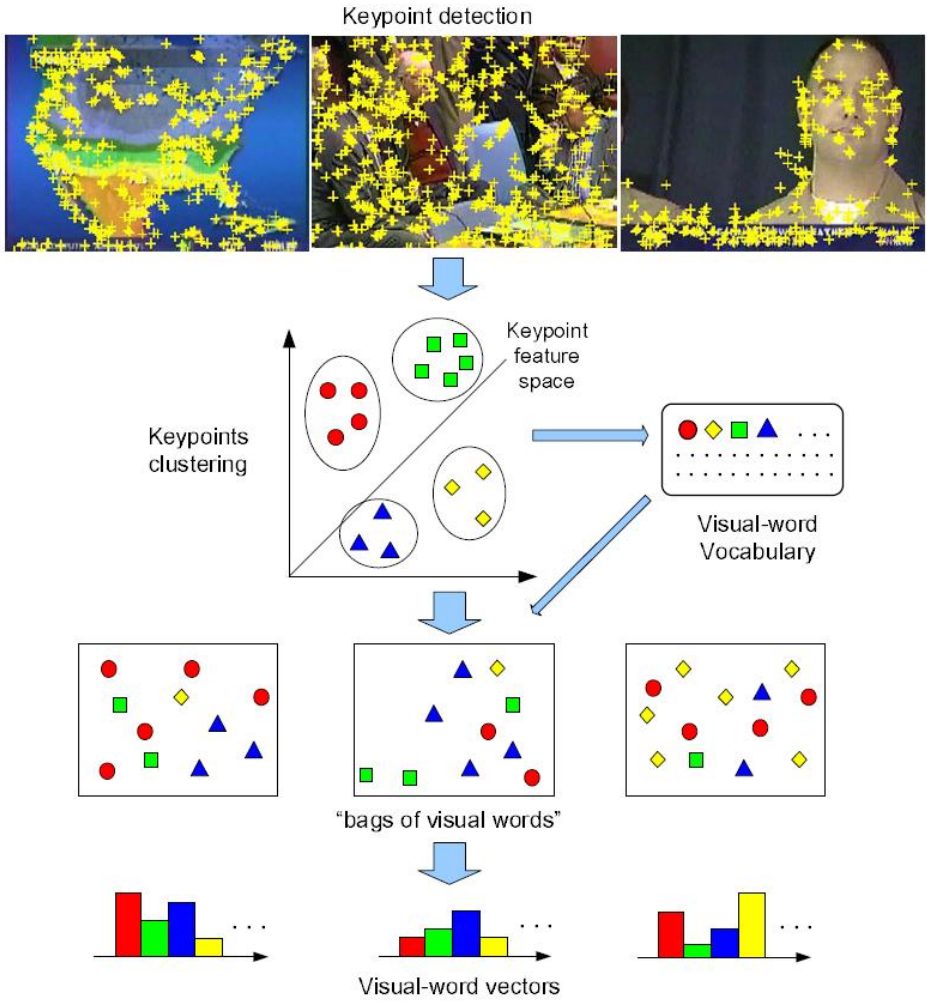


Fig. 3. Architecture of the semantic space labeling system based on place classification (LTM: Long-Term Memory; STM: Short-Term Memory).

Evaluating Bag-of-Visual-Words Representations in Scene Classification
Jun Yang et alii – CMU and City University of Hong Kong

- Evaluate "text classification analogies" to produce different Bag-of-Visual Words (term weighting and normalization, stop word removal, feature selection, ...) and other BoVW "parameters "(vector size, use of spacial information, ...)
- Datasets: TRECVID 2005 (20 most frequent concepts: outdoor, indoor, objects and people activities) and PASCAL 2005 (only 4 categories: motorcycles, bicycles, people and cars)



How could we use, improve and merge these works

- Xiao (SUN Dataset)
  - Benchmark to evaluate our proposal
- Dickson (Evolution of object categorization)
  - Support our quest for new structural representation and inference mechanism
- Choi (Spatial perception and language)
  - We must take into account during our research
- Santini (Definition of categories)
  - We must take into account when building our grammar
- D`Ulizia (Survey on grammar inference)
  - Maybe we can use/adapt one of these method for visual grammar learning
- Li (Object Bank)
  - Maybe we can improve this method introducing higher level structure using grammars
- Shet (Predicate Logic based Visual Grammar)
  - Can be used as our representation mechanism
- Le (Large scale unsupervised learning)
  - Maybe we can also benefit from paralelism and grid computing
- Zhu (Stochatisc grammar of images)
  - Maybe we can use some of the general rules they have proposed

How could we use, improve and merge these works

- Gee (Higher level structure for SLAM)
  - Maybe we can get even higher level structures and extend the states used by the Kalman Filter
- Newman (FAB-MAP)
  - Not sure … compare against ?
- Zheng (Visual Synset)
  - Can be used as grammar atoms … facts …
- Nwogu (Ontology)
  - Maybe the superpixel concept is usuful
- Pronobis (Multi-model)
  - We can also use multi-model …
- Yang (Evaluation bag-of-words)
  - Not sure if we must use bag-of-words at the low level feature level. Maybe we can compare the alternatives