# Computer Vision

James L. Crowley

M2R MoSIG Fall Semester

12 Nov 2020

Lesson 5

## Attention and Cognition for Computer Vision

**Lesson Outline**:



**Bibliography**


- N. Cowan, Working memory underpins cognitive development, learning, and education. *Educational Psychology Review*, 26(2), 197–223, 2014
- J. F. Sowa, *Knowledge representation: logical, philosophical, and computational foundations* (Vol. 13). Pacific Grove, CA: Brooks/Cole, 2000.
- W. Kintsch, *Comprehension: A paradigm for cognition*. Cambridge university press, 1998.
- P. N. Johnson-Laird, *Mental models*, MIT Press Cambridge, MA, USA, 1989.
- J. R. Anderson, A spreading activation theory of memory. *Journal of Verbal Learning and Verbal Behavior*, *22*(3), 261-295, 1983.
- M. Minsky, A Framework for Representing Knowledge, in P. H. Winston (ed.), *The Psychology of Computer Vision*. McGraw-Hill, New York (U.S.A.), 1975.
- G. A. Miller, The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63(2), 81-97, 1956.

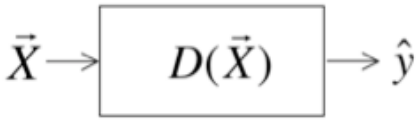
# 1 Cognitive Vision

## 1.1 The Hour-Glass model in machine learning

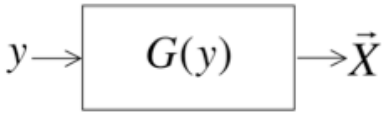
An increasing trend in Machine Learning is to construct an "end-to-end" system that maps an input signal (such as an image) directly onto an output signal (such as spoken word or an action). This is used, for example for speech translation and for chat bots. This is done by combining a discriminative network with a generative Network.

**Discriminative and Generative Networks**

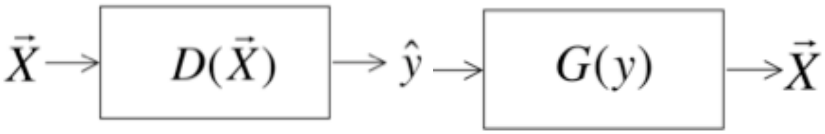
The Discriminative networks takes a signal as an input and gives the likelihood the one or more target classes are in the signal as an output.



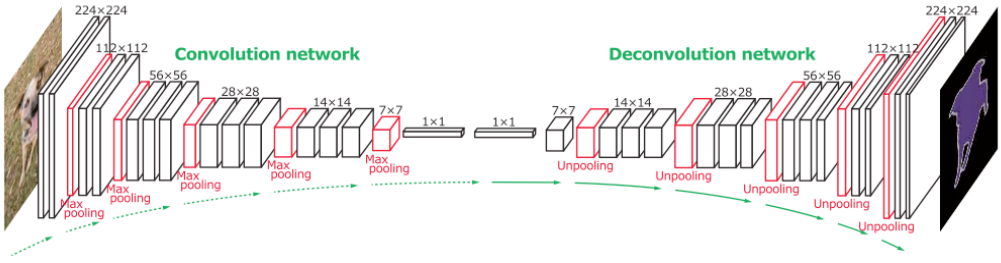

A Generative network runs the other way, generating a realistic signal from an input code.




Both the Discriminative and the Generative process can be learned from training data using back-propagation. We can put two such networks together to generate an output in one domain from an input in another.




This called an "Hourglass model". An autoencoder is an example.




However, this model generates a simple Stimlus->Response.
There is no "understanding". There is no intelligence.

What would it take to add "intelligence" to such a model?

We can find inspiration in Cognitive Science.

## 1.2 Long Term Memory and Working Memory

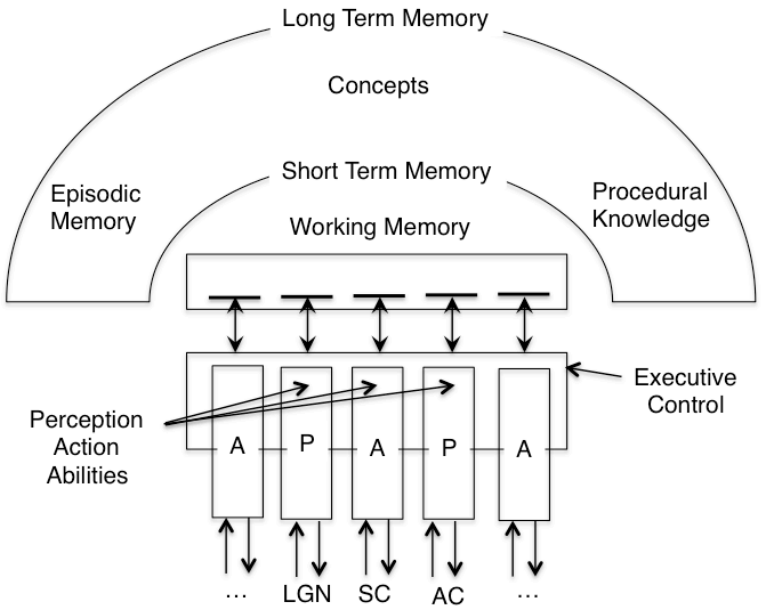
Most models of human cognition share a number of common elements:

- Perception: Transforms and combines sensory stimuli to Phenomena
- Action: Transforms intention to motor commands.
- Perceptual Memory: A temporary buffer holding recent perceptual stimuli
- Motor Memory: A state model that generates a temporal sequence of actions.
- Working Memory: 7+/-2 memory entities (perceived or remembered)
- Long Term Memory: Episodic, Procedural, Spatial and Conceptual

Long-term memory (LTM) refers to memory structures used in several different cognitive abilities:

- Episodic Memories: recordings of significant sensory experiences
- Conceptual Memory: Abstract representations for sensory experiences
- Procedural Memory: Sequences of operations to accomplish goals
- Operational memory: (also called skills) reactive perception-action abilities.
- Spatial memory: A network of spatial relations between places or entities
- Auditory memory (Sounds, words, music)

## 1.3 Working memory

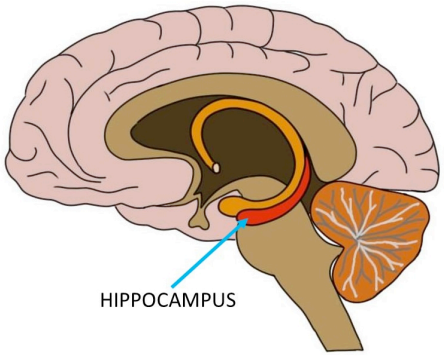



The core element for a cognitive system is Working Memory (WM) where sensory information (phenomena) are associated with real or synthetic memories (episodes), actions, procedures (procedural knowledge), concept and other perceptions. Working memory is a limited number of storage units that are used to associate perceptual phenomena with episodic memory, learned concepts, spatial memory, procedural knowledge and reactive skills (actions).

In humans, working memory is thought to reside in the hippocampus with dense associations to the visual cortex, auditory cortex, and many regions of the superior cortex. It is easily demonstrated that Working Memory for humans is limited to an association of 7 +/- 2 entities (Miller 56),

Working memory is demonstrated by asking people to retain a series of random letters or numbers while distracting them with a different task. The average person can retain no more than 7 letters in working memory. Some individuals can hold 8, or even 9. Others are limited to 6 or 5. Unless the person refreshes working memory by internal repetition (rehearsal), elements in working memory tend to decay within 30 seconds to a minute.

There is evidence that that working memory is located in an organ at the center of the brain known as the hippocampus, although this controversial, with some authors arguing that working memory is distributed. The hippocampus is known to be involved in relational reasoning, particularly with relations of space and time. For example, the hippocampus of rats can be shown to hold a network of places, and there are demonstrations in which researchers drive a rat through a maze by stimulating different regions of the hippocampus with electrodes.



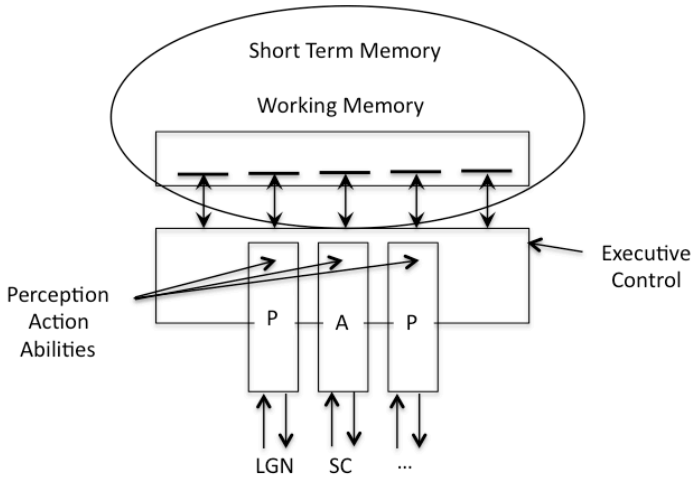

Information from the various senses is processed by specialized regions of the brain and integrated in working memory. In most cases, sensory information takes the form of a spatio-temporal map that is processed and combined by a series of neural layers to produce interpretations.

Perception and action are sub-symbolic processes that are activated by working memory. Visual information that passes through an attention filter triggers recognition processes that express visual patterns as entities recalled from concepts. Entities in working memories are instantiations of Concepts from Long-Term memory. Visual concepts represent perceptual phenomena and are associated with episodic memories, procedures, and other entities in working memory. Visual concepts are learned from experience and do not necessarily have a name.

Actions are learned processes for locomotion, manipulation, vocal expression etc. Actions include "perceptual actions" that can enhance perception by directing fixation, tuning auditory perception, or focusing attention.

## 1.4 Perception is active, Action is perceptive

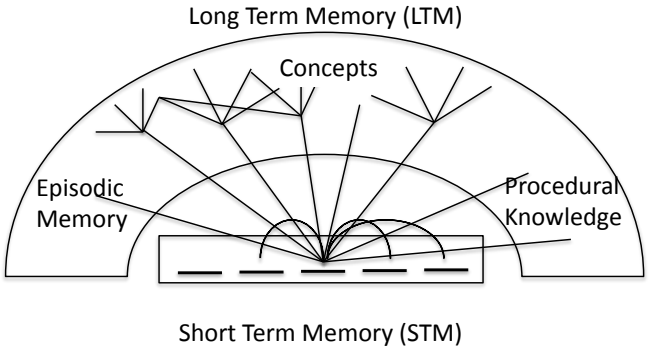
The perception and action channels have state (memory), similar to what is currently provided by Recurrent Neural networks such as LSTM.




Recognition is an active process, with hypotheses in working memory driving a generative process that generates a synthetic image that is compared (top-down) to the sensor image. When the synthetic image and the perceived image match, the concept is reinforced. Errors in matching reduce the activation and cause the process to examine other hypotheses. Similar active processes are involved in perception from all of the senses including hearing (auditory perception), touch (tactile perception), Taste (gustatory perception) and smell (olfactory perception). A similar phenomena occurs with action, as the results of muscle movements are felt through a phenomena called "proprioception". Internal organs are sensed through the somatic system.

## 1.5 Spreading Activation

Spreading activation (Anderson 83) is a mechanism for associating entities in Working Memory (WM) with Long Term Memory (LTM) and controlling the contents of the working memory. Most cognitive theories assume a form of Hebbian memory, in which activation energy from working memory propagates through a network of cognitive "units" in Long Term Memory. LTM Units that receive energy from several other units can become "activated" and can replace one of the 7±2 active units in working memory.




For example, Kintch (Kintsch 98) explains that when we read a word, we use frequency of occurrence of word sequences (N-Grams of words that frequently co-occur with that word) to activate many possible interpretations. As additional words are perceived, the associations are correlated and the most activated association provides meaning.

Activation energy spreads from working memory to other elements of working memory and to long-term memory including concept memory, episodic memory, procedural knowledge, etc. Activated units then spread their energy to other units where it can arrive from multiple paths and accumulate. At the same time the energy decays with time, disappearing within 30 seconds. Theories differ in describing how activation energy propagates and how this propagation can be controlled by emotions and physiological state.

Units that receive energy from several other units can become "activated" and can replace one of the active units in working memory. Theories differ in describing how activation energy propagates and how this propagation can be controlled by emotions and physiological state.

The limited size of working memory is the primary bottleneck for cognition. This limit is used to explain and predict many phenomena in human Factors (Ergonomy). This limit is NOT because of the cost of memory. The limitation is the algorithm complexity caused by spreading activation. This limitation is actually an advantage, because it forces abstraction and formation of new concepts, in a process called "chunking".

## 1.6 Chunking

Chunking is a process of hierarchically grouping cognitive units into larger composite units. Chunking allows multiple cognitive units to be held as a single element in short term memory, overcoming the 7±2 limit. However, associations for a chunk in WM are with the chunk and not its individual elements.

To say more, we need to define what we mean by a cognitive "unit". In Cognitive systems, spreading activation is represented as relations that are encoded as predicates in Rules or Frames.

# 2 Conceptual Knowledge

Concepts are mental constructs that represent abstract or generic ideas generalized from particular instances. Concepts are the basic elements of cognition. Concepts provide abstractions for reasoning by associations with different forms of LTM, as well as symbols for communications .

Concepts arise as abstractions or generalizations from experience or from the transformation of existing concepts. For human's, concepts can be learned from experience, or by communication with other humans. Concepts can represent words, actions, perceived phenomena, experiences, feelings, etc. Many concepts do not have an identifiable name, but are, none-the-less used for reasoning.

A concept is instantiated as an association of memories and other concepts. These memories may be images, sounds, image sequences, feelings, or any other perceived phenomena (e.g. taste, smell, etc).

Chunking is a process of grouping individual cognitive units into larger composed units. Chunking allows multiple cognitive units to be held in working memory at the same time, overcoming the limits to working memory. However, associations to LTM and STM are with the chunk and not its individual elements. Some theories postulate that frequently encountered situations are recorded as "chunks" giving rise to new concepts.

In both computer science and cognitive science concepts are formalized as Frames using Schema.

## 2.1 Schema

Schema are declarative structures for representing concepts. The term Schema and was originally proposed by Emmanual Kant in "Critique of Pure Reason (1781). Schema have been used in philosophy and cognition psychology since the 19th centurey to represent concepts for reasoning, perception, problem solving and natural language interaction.

A schema (plural schemata or schemas) describes a pattern of thought that organizes information. A key property of Schema is the association of concepts with procedures for perception, action and reasoning.

Schemas represent concepts as data structures with slots that define the properties of the concept and associate the concept with other concepts.

A typical Schema for a concept has

1) A name.
2) A definition (test for inclusion)
3) Meanings: memories of examples of the concept
4) Roles: Operations or procedures that are enabled or prevented by the concept.
5) Relations to other concepts and other elements in LTM.
(In many schema systems, meaning and roles are part of the list of associations).

Meaning denotes memories that serve as examples. Meanings can be from actual examples or can be imagined. Meanings can be visual, episodic, auditory, olfactory, emotional or examples of feelings.

Roles are operations or procedures (procedural knowledge) that are enabled or prevented by the concept. Roles can also refer to uses that the concept can have.

For Example: Consider the number 5.

The number 5 has
1) a name: (five in english, cinq in french etc).
2) a definition (the name of a set of all sets with 5 elements)

This is an intentional definition that may be implemented either by counting the elements (5 comes after 4) (procedural Knowledge) or by direct recognition (learned perceptual ability).

3) Meanings: Experiences with examples of the concept 5. (visual pattern, sounds).

4) Roles: Operations such as addition, subtraction, division, etc that are made possible. (Example 5 can not be directly divided by 2).

5) Relations: Associations with other concepts, episodic memories, or actions.
Multiple kinds of relations are possible:
ISA and AKO: Identifies the concept as a member of a larger class.
(AKO = A Kind Of). Examples: (5 ISA number) (5 ISA integer) (5 ISA odd)
Part-Of: Identifies the concept as a component of a larger concept
(5 is a part of the number 15, 5 is a part of the formula 15/3)
Order Relations : $(5 < 6)$, $(3 < 5)$, Time relations 5h is before 6h.

## 2.2 Relations

Relations organize concepts and associate concepts with perception, action, LTM and STM. Examples include temporal relations, spatial relations, Part-whole relations, family relations, social relations, administrative organizations, military hierarchies, and class hierarchies.

**Kinds of Relations**

A non-exhaustive list of relations between concepts includes:

1) Class membership (ISA, AKO) relations
2) Structural (Part-of) Relations
3) Ordinal relations (bigger-than, smaller than)
4) Temporal relations (Allen's 13 relations between intervals).
5) Spatial Relations (right-of, left-of, above, below, in-front-of, behind, etc)
6) Organizational relations (team member, leader, etc)
7) Family (parents, brothers, sisters, etc)
8) Causal (action A caused phenomena P)

This list is NON-EXHAUSTIVE. Relations can be defined as needed by a domain.

## 2.3 Predicates

Relations are formalized as Predicates (Truth functions).

A predicates is function that associates concepts. Traditionally, predicates are assumed to be Boolean functions, but probabilistic predicates are increasingly used to represent relations.

A predicates is a function that tells whether or not a relation is valid for a set of entities. Classically, predicates are treated as Boolean functions that can only return a value of TRUE or FALSE. As we have seen, in probabilistic reasoning, predicates represent the likelihood that the relation holds, with a value between 0 and 1.

## 2.4 Implicit vs Explicit representations for Relations

Relations can be represented "implicitly" or "explicitly".

**Implicit representation**

Most programming systems for structured knowledge representation provide data structures that represent cognitive units as "objects" or instances of a class.

In such a system, each object is a list of named slots. Slots can have types, default values, possible ranges and other defining information. Slots can contain values for properties, contain pointers to other units or pointers to code that can be executed. Cognitive units are organised in class hierarchies, providing inheritance of structures and procedures.

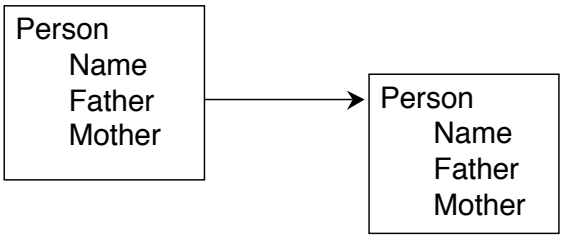
Such structures are typically accompanied by a number of procedures (methods), and are implemented as a form of object-oriented programming system.

In such systems, the cognitive unit associates properties, code and other units "implicitly". With an implicit representation, the relation is represented as a pointer in a slot.

For example, the following is a Class for family relations expressed with a simplified form of LISP.

(defclass PERSON (slot NAME) (slot FATHER) (slot MOTHER) ).
John <- (make-instance PERSON (NAME John) (FATHER Joe) (MOTHER Jane))

The slot FATHER contains a pointer to an object of the class PERSON that represents the father of the person. The pointer is the object address.




Implicit representations are simple and more efficient in computing and memory.

However, with an implicit (slot based) representation for a concept, the set of relations is fixed and cannot change dynamically.

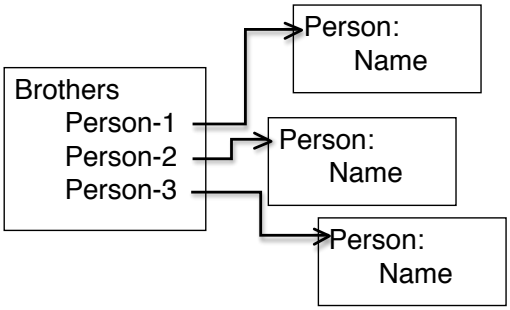
An implicit (slot based) representation for relations is not easily completed with meta information. Some forms of reasoning are much easier with an explicit representation.

**Explicit representation**

With an explicit representation, relations are represented by a schema whose arguments are concepts. Explicit Representations for relations can be changed dynamically without changing the underlying concept.

(defclass PERSON (slot NAME) )
(defclass BrothersP (slot person1), (slot person2), (slot person2))

A slot holds a pointer to the object that represents the relation. This object can then provide additional information about the object, such as what, where, why, when, who and how.




With an explicit representation for relations, it is possible to write a set of general procedures for acquiring (learning), reasoning, and explaining that apply to all relations.

With an implicit representation, such procedures would be specific to each class.
Thus explicit relations support generalized methods acquiring (learning), reasoning, and explaining about relations.

For example, Allen's temporal reasoning is much easier to program using explicit models of relations because the set of relations between intervals changes dynamically.

# 3 Structured Knowledge Representations

Structured knowledge representations were invented in the 1970's as a programming tool for intelligent systems. This approach suffers from a number of problems:

1) Top down reasoning: Frames (and most schema systems) are designed for top-down reasoning. Most human reasoning is both top-down and bottom-up (active), with associations flowing both ways.

2) Knowledge Acquisition: Building a Frame system by hand is long, tedious, and ad hoc process. There is a temptation to overload the system with useless information, "just in case". Automatic acquisitions (learning) of frame systems for recognition and reasoning is a notoriously hard research problem for classical AI.

3) Context Recognition (The Frame problem): Many problems are easily solved once the appropriate frame is known. Recognizing the correct context can be very difficult.

4) Semantic Alignment: Two Frame systems describing the same concepts, may not have the same relations. Meanings of similar concepts might be slightly different. However, communication and integration of conceptual knowledge from different sources requires a shared ontology.

However, recent progress in machine learning have provided possible solutions to most of these problems, and combining structured knowledge representation with machine learning has recently emerged as a very hot area for research in artificial intelligence.

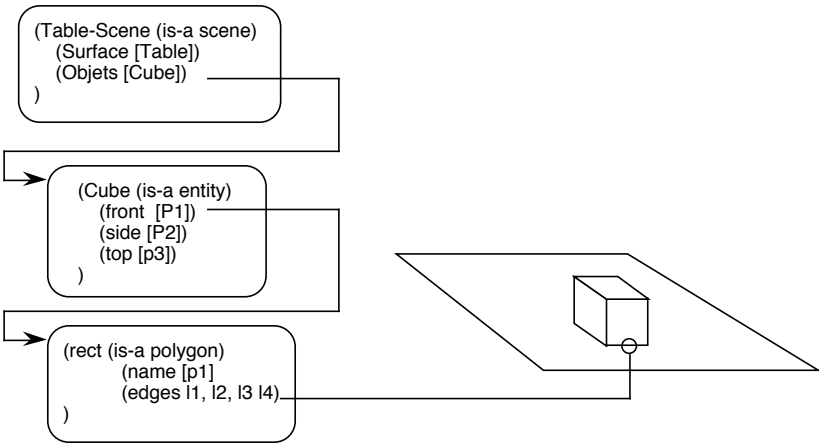
## 3.1 Frames

Frames were proposed by Marvin Minsky in 1976 as a data structure that could be used to that guide visual perception. Frames represent perceived entities as examples of concepts.

Frames were popular for many years as a programming method to organize perceptions in Computer Vision, Linguistics and Cognitive Systems. Frames provide a control structure to guide visual interpretation in a top down manner. Frames tell a vision system where to look and what to look for. A frame describes a perceived entity with a set of properties and relations, represented by slots. The frame includes a collection of procedures for perceiving, reasoning and acting with the concept.

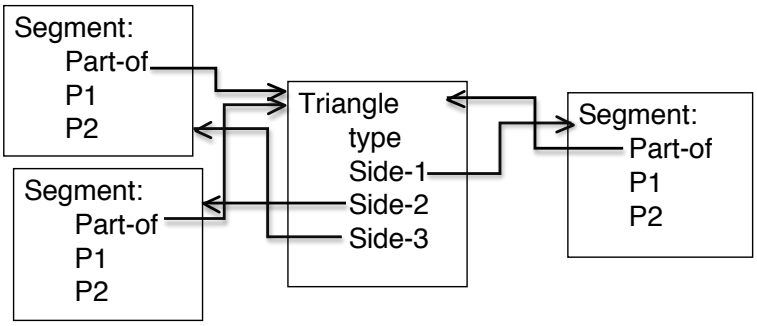
Frames provide procedures or operations to detect entities. Frames also provide default values for properties when perception is not possible or fails.

Frames can be composed hierarchically to describe complex entities.

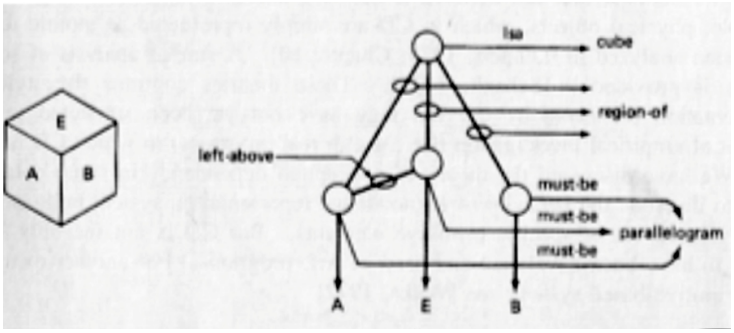



Frames are implemented as a form of Schema. They are composed of relations, represented by slots that contain pointers to other frames. Relations represents information about the object, such as part relations (composed of, part-of), Position relations (above, below, beside, inside, contains), Time relations (before, after, during), as well as specific properties of the entity (size, position, color, orientation).

For example, the concept Triangle has the part relation "composed of" with three segments. The triangle can also have an is-a relation (category membership) with different triangle types such as equilateral, isosceles, right angle, etc.




Ultimately, some slots point to raw perceptions (visual phenomena).




A Frame for a Cube
(from E. Rich "Artificial Intelligence", Fig 7-13, p231

Frames typically come with methods (procedures) for searching for the entities that can plays roles in the frame. Typically a slot-filling procedure will apply a set of acceptance tests to an entity to see if it satisfies the requirements for the slot.

Frames generally include prototypes that can serve as examples in reasoning, and default values that are used if no entity has been found to fill the slot. Thus frames can be used for abstract reasoning or for reasoning when perception is not possible.

## 3.2 Scripts

A script is a schema structure used to represent a stereotypical sequence of events.

Scripts are used for interpreting stories. For examples, scripts used to construct systems that interpret and extract information from Newspaper Stories. Scripts are used in natural language understanding systems to organize a knowledge base in terms of the situations that the system should understand. Scripts are also used to observe an actor and to describe (or recognize) what the actor is doing. This includes plan-recognition as well as activity description. Scripts are also be used to represent procedural knowledge for plans.

Scripts are schema much like Frames, except that the slots point to a sequence of situations. Classically, a script is composed of

1) A Scene: situation in which the script takes place
2) Props: Entities (objects) involved in the script.
3) Roles: Actors (agents) that can provoke changes in the scenes.
   Actors are typically people, but may be artificial.
4) Events (acts): A sequence of events that lead to changes in situations and make up the script.

The script can be represented as a tree or network of scenes, driven by actions of the actors. In each scene, one or more actors perform actions. The actors act with the props. The script can be represented as a tree or network of states, driven by events. As with Frames, scripts drive interpretation by telling the system what to look for and where to look next. The script can predict events.

***Example of a script: Restaurant Script.***

The classic example is the restaurant script:

Props: A restaurant with an entrance, tables, chairs, plates, eating utensils, glasses, menu, etc

Actors: The host (Maitre d'Hotel), clients, servers, chef, bus-boy, etc.

Scenes: Entry, seating, reading the menu, ordering, serving, requesting the check, paying, leaving, etc.

***Scripts provide context for default reasoning.***
As with Frames, scripts drive interpretation by providing procedures that tell a system what to look for and where to look for it. Scripts also provide default knowledge for reasoning about stories or actions.
For example, for story understanding, the story will typically only provide sparse detail of what happened. The reader is expected to fill in the missing knowledge with default knowledge.

### 3.3 Situation Models

Situations models are used in cognitive psychology and ergonomics to express the mental models that people use to understand and reason. Situations models describe the contents of Working Memory.

A situation model is composed of entities with properties, and relations. Relations may be with other entities, memory episodes, concepts, actions or procedures.

| | |
|---|---|
| Entities: | Anything that can be named or designated; People, things, etc.<br>(entitles are defined using schema or frames) |
| Properties: | Descriptions of entities such as position, size, color, etc |
| Relations: | N-ary predicates (N=1,2,3 …) that relate entities.<br>(relations are defined as tests on the properties of entities). |
| Situation: | A set of relations between entities |

# 4 Situation Models as a Programming Method

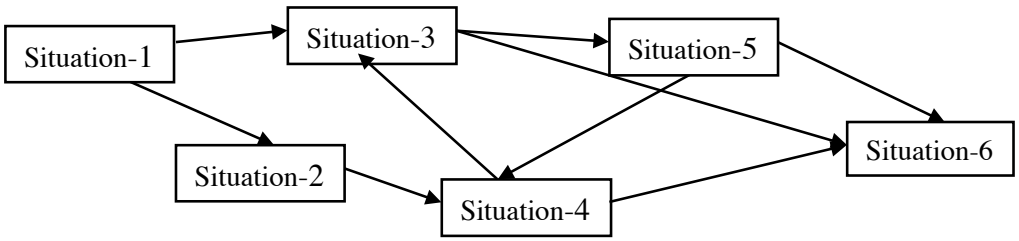
Situation models are also used as a programming method to construct context aware systems. Situation models
(1) Providing meaning and explanations for observed phenomena
(2) Describe phenomena that cannot be observed
(3) Predict phenomena that have not yet occurred.

A Situation is a set of relations over entities (a state).

A situation, S, is defined as a : {X},{R}
1) A set of Entities {X}: Instances of concepts that represent phenomena.
2) A set of Relations {R}: Predicates over subsets of Entities.

A Context is be defined as
1) A set of entities, {X}, with their properties.
2) A set of relations {R}
3) A set of situations {S} composed of relations over entities, such that each situation includes
– A list of adjacent situations, possibly with transition probabilities.
– A list of expected events that can occur when the situation is active, with possible reactions to the events
– A list of system behaviors that are allowed or forbidden,

Situations can be organized into a state space referred to as a situation network.
A situation model is a directed graph where the nodes represent situations and the arcs represent events. Each situation (or state) corresponds to a specific configuration of relations between entities. A change in relation results in a change in situation (or state).




Each situation can prescribe and proscribe behaviors.
1) Behaviors: List of actions and reactions that are allowed or forbidden for each situation. Behaviors are commonly encoded as Condition-Action rules.
2) Attention: entities and relations for the system to observe, with methods to observe the entities
3) Default values: Expectations for entities, relations, and properties
4) Possible situations: Adjacent neighbors in the situation graph.

Each situation indicates:

Transition probabilities for next situations

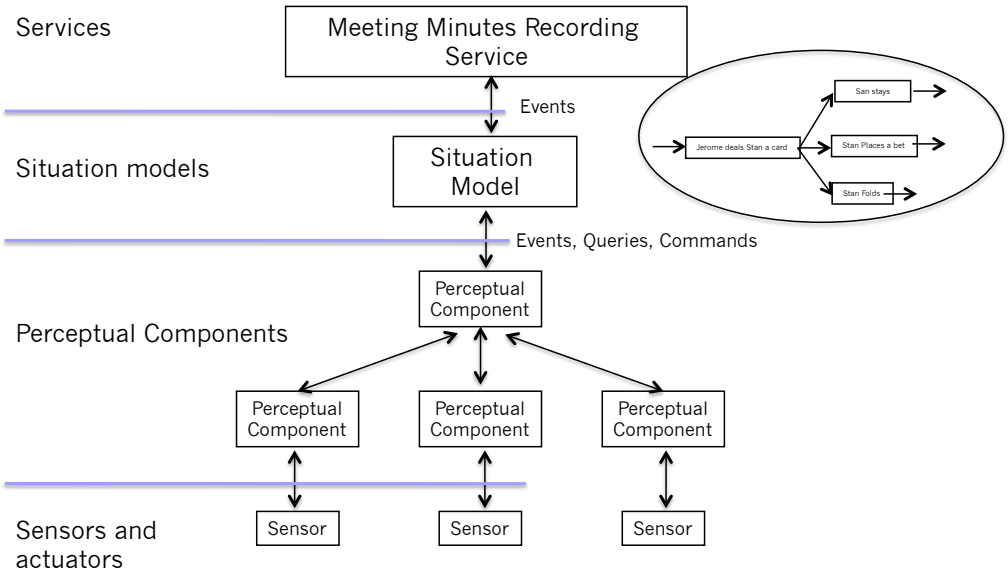
The appropriateness or inappropriateness of behaviors

Behaviors include

1) methods for sensing and perception, and
2) appropriateness of actions
3) changes in state in reaction to events.

The sets of entities, relations, behaviors, and situations define a "Context".

Example: Multimodal Meeting Recording System constructed at INRIA around 2008. (the Rocamaroll smart multi-camera system currently installed in Amphi D is a commercial version of this system).




In this example, the situation model acts as a form of Working Memory, telling the system where to look, and what to look for in order to follow the meeting.

<Fame Intelligent Camera Man Demo>