Visual Perception and Attention

Tanja Schultz
Felix Putze
Dominic Heger

26.5.2011
Lecture „Cognitive Modeling“ (SS 2011)
Outline

- Introduction Perception
- Human Senses and Computer sensors
- David Marr’s model of visual perception
- Bottom up vs Top down processing

- Introduction Attention
- Donald Broadband’s Filter Theory
- Anne Triesman’s Feature Integration
- iCub’s Attention Model
- Jeremy Wolfe’s Guided Search Model
Perception

- Encyclopedia Britannica: *Perception is the process of registering sensory stimuli as meaningful experiences*

- Some aspects of Perception
  - Perception is not a passive receipt of sensory information or sensation
  - Not all stimuli are perceptions, only those that are cognitively processed and belong to the subject’s mental orientation
  - Includes reception, selection, processing, and interpretation of sensory information
  - Perception involves the "top-down" mental state, as well as the "bottom-up" process of processing sensory input.
  - Perception is basis for the construction of mental models of the environment
Computer Science Approach to Perception

- Large number of active and passive computer sensors
  - Cameras: Stereo, Time-of-flight, infrared, thermal,…
  - Microphones: Close talking, cardiod, arrays, …
  - Tactile sensors: Laser, Ultrasound, inertial sensors, …
  - Biosensors: EEG, EMG, EDA, …
  - Etc.

- Russell & Norvig\(^1\) propose two types of perception
  1. Function features
     - Derive features from sensor and directly use it for motor control
     (E.g. flies control muscles by features derived from optical flow)
  2. Model based perception
     - Construct a model of the world from sensory data
     - Idea: sensory stimuli \( S \) are a function \( f \) of the state of the world \( W \)
       (\( f \) is not completely invertible)

\[ S = f(W) \]
Human Senses

- Human perceive their world mainly by five senses
  - Visual
  - Acoustic
  - Tactile
  - Olfactory
  - Gustatory

- Specialized sensory areas in the Brain
  - Very simplified illustration
  - E.g. more details on visual system on following slides

- However, large parts of the cortex are multisensory (Ghanzanfar & Schroeder, 2006)
Human Visual System

- Light arrives
- Projected on the retina
- Photoreceptor cells (rods, cones) connected through intermediate cells to the optic nerve
- Via optic nerve to optic chiasm
  - Information from the left visual field of both eyes (right side of the retina) go to the right hemisphere
  - Information from right go to the left hemisphere
- Thalamus
  - 6 layers of Lateral Geniculate Nucleus (LGN)
Human Visual System

- Two visual pathway streams to each hemisphere
  - Dorsal stream: where?/how?
  - Ventral stream: what?
- Processing works parallel
- Some of the areas process mainly color, form, motion, ....

- Some known connections between cortical areas:

- But also subcortical areas are involved (e.g. superior colliculus)
David Marr’s Perception Model


- Modular processing stages
  1. Retinal image
     - Spatial distribution of light intensities on the retina
  2. Primal sketch
     - Basic features of the scene (e.g. contours, edges, ...)
  3. 2½-D sketch
     - Grouping of basic features to represent areas, rough depth characteristics, ...
  4. 3-D model representation
     - Combination to form 3-D environmental objects
     - Representation of the existing world
     - Subjective perception of the observer
David Marr’s Perception Model

- Marr proposed an algorithm and implementation of the model
  - Retinal image
    - Input intensity image on retina (2-D)
  - Primal sketch
    - Basic features of the scene (e.g. contours, edges, regions ...)
    - Detection of intensity changes in image
    - Convolution with edge detection operator (e.g. Marr-Wavelets or Difference-of-Gaussians filter)
David Marr’s Perception Model

• 2½-D sketch
  – Viewer-centered combination of information from both eyes
  – Rough depth and spatial information
  – Optical flow, rough reconstruction of distances

• 3-D model representation
  – Object centered hierarchical model of the world
  – Generalized Prototypes consisting of cones and cubes
  – Scene visualized as 3-D map
David Marr’s Perception Model

- **Discussion**
  - Model combines insights from psychology, neurophysiology and artificial intelligence (computer vision)
  - Proposed idea: One must understand 3 levels to analyze an information processing systems
    - Computational level (goals of the system)
    - Algorithmic level (how is input transformed to output?)
    - Implementational level (physical realization)
  - Some physiological validation (e.g. ganglion cells in retina are very similar to Difference-of-Gaussian filters)
  - Ignores important aspects perception
    - Color information
    - Multimodality (e.g. McGurk effect)
    - High level cognitive processes (bottom-up approach)

Perception is pure transformation from visual sensory input
Bottom-up / Top-down Processing

- Perception and recognition of stimuli can be guided from different directions

- **Bottom-up processing**
  - Originates from sensory information (pattern parts and their context)
  - From the individual pattern parts to the whole
  - Data driven

- **Top-down processing**
  - Originates from cognitive processes (high level knowledge/expectations)
  - From the whole to the individual pattern parts
  - Goal driven

- Human perception works in both directions
  - Some brain areas do bottom-up others top-down (both interconnected)
  - Experimental results
    - Bottom-up: Long presentation of stimulus, visually clear
    - Top-down: Short presentation of stimulus, vague visual clarity
Top-down Bottom-up in ASR

- Example: Automatic Speech Recognition exploits top-down as well as bottom-up processing

- Bottom-up: Acoustic information
  - Contributes sensory information to the recognition process
  - Extract features based on frequency characteristics of phones from audio data
  - Classification by Gaussian Mixture Models

- Top-down: Language information
  - Contributes information on syntax and semantic
  - Integrates prior knowledge about speech into recognition process
  - Statistical model learned on large text databases
  - Guides recognition process, i.e. which acoustic models will be evaluated
An Experiment
Inattentional Blindness

- Only a small region of the scene is analyzed (i.e. attended) in detail at each moment
- Famous invisible gorilla test (Simons & Chabris, 1999)
  - More than 50% of subjects do not see the gorilla
  - Effect is dependent on difficulty of the distraction task
  - → Relationship between visual impression and perception is strongly dependent on attention

- Selective attention
  - We perceive events selectively
  - Only when we have attention on them

- Shows how important attention is for human cognition
Other Famous Effects of Attention

- **Cocktail party effect**
  - Humans are able to select one of different acoustic sources from a mixed incoming signal
  - But if somebody says your name it immediately catches your auditory attention
  - Binaureal effect related to the auditory localization

- **Pop-out effect**
  - Some properties of stimuli make them stand out and immediately catch attention
  - Visual search has very low reaction time
  - Mostly occurs when distractors are homogeneous
  - Scenes are called odd-man-out scenes
What is Attention?

- Definitions
  - Cognitive Psychology, Solso (2005 p. 83): The concentration of mental effort on sensory or mental events
  - Corbetta (1990): Attention defines the mental ability to select stimuli, responses, memories, or thoughts that are behaviorally relevant among the many others that are behaviorally irrelevant

- Limited resources for cognitive processing and behavioral actions
  - Selection of information
  - Unbounded visual search is NP-complete (Garey & Johnson, 1979)

- Goal of attention is control of behavior (“selection for action”)
  - Frintrop et al. (2010): “In the broadest sense, any pre-processing method might be called attentional, because it focuses subsequent processing to parts of the data which seem to be relevant.”
Overt and Covert Attention

• Overt attention
  • Directing of senses towards a stimulus source
  • Example: Focus an object with the eyes

• Covert attention
  • Mental focusing on one of several sensory stimuli
  • Covert visual attention is linked to saccadic eye movement
  • Examples: How does your left big toe feel?
    Suddenly spot your name in a list
Broadbent’s Filter Theory

- *Perception and Communication* (Broadbent, 1958)
- Simple classical descriptive model of attention

- Attention is the result of an information processing system with restricted capacity
- Channel capacity is limited (in information theory sense)

- Brain is not capable to process all incoming sensory stimuli
- Humans focuses of selectively on few stimuli and ignores a major part of the other stimuli
Broadbent’s Filter Theory

- Pipeline model (serial processing)

- All incoming perception gets into sensory buffer
- Filter directly after sensory buffer
  - based on physical characteristics (e.g. pitch, intensity)
  - not based of analyses of information
- Filtered information gets into single serial limited capacity channel
Discussion of Broadbent’s Model

- **Model assumption**
  - Early filters based on physical characteristics (e.g., Location, size, frequencies, etc.)
    - Filtering occurs before stimulus information is processed
  - Split attention requires fast multiplexing of the filter

- **Extensions of Broadband’s model**
  - Attenuation of unattended information (Triesman)
  - Late selection (e.g., Deutsch & Deutsch 1963): All sensory information are preliminarily analyzed
  - Physiological evidence that humans use early filtering and late selection (Pashler, 1996)
Triesman’s Feature Integration Theory (1980)

- **Preattentive processing**
  - *Features* are processed as first step of visual processing instead of pure physical characteristics
  - Detectors for each feature
  - Level of analysis depends on available processing resources
  - Representation in feature maps
    - For each feature
    - Topographical map (correspond to locations on retina)

- **Attentive processing**
  - Fusion of all features into master map of locations
  - Attentional spotlight is focus of attention
  - Unattended information is no completely filtered but attenuated -> cocktail party effect
Features

- FIT adds basic features into Broadband’s filter theory and is the basic architecture found in many modern computational attention systems.

- Neuroscientific findings (fMRI, intracortical EEG) that brain processes basic visual features.

- Experimental evidence:
  - Example: Visual search experiments
    - Target is single feature
    - Distractors are homogeneous and different in the feature

- Primitive features of human perception:
  - Color and intensity, Motion, Orientation, Size
  - Probably: flicker, luminance, small break in line, depth, etc.
Saliency and Bottom-up attention

- Calculate those regions in an image interesting/relevant for further processing

- Basic structure of most bottom-up attention systems
  - Extract several features in parallel from input images
  - Fuse into saliency map
  - Find the most salient region
  - Set focus application’s attention

- Other architectures not discussed here
  - Connectionist
  - Dynamic systems
Saliency Map based attention

- Roots in Treisman’s Feature Integration Theory
- Basic process of Saliency map based attention
  1. Feature maps representing conspicuity of stimuli in a particular feature
  2. Topographic saliency map (combination of feature maps)
  3. Selective mapping into central non-topographic representation
  4. Winner-take-all selection on conspicuity of location and proximity
  5. Inhibition of selected location causing shift to next most conspicuous location
Empirical Evaluation of Attention models

- Detection of known effects usually on artificial images (e.g. odd-man-out scenes)

- Large databases primarily for evaluation of saliency based visual attention exist
  - Real world images
  - Groundtruth based on eye tracking data

- Application performance of attention systems
  - Support subsequent information processing (less computational time, accuracy of results)
  - Quantitative measurement of task performance
    E.g. success rate and speed of robot performing putting plate on a table
  - Qualitative measurement of natural behavior of a humanoid
iCub’s Attention Framework

- iCub (open) Humanoid Robot Platform
- Visual and Acoustic saliency maps
- 2 eyes and 2 spiral formed ears
- Real-time control 6 degrees of freedom for neck and eyes
iCub’s multimodal Attention Framework

Multimodal Saliency-Based Bottom-Up Attention
A Framework for the Humanoid Robot iCub

Jonas Ruesch¹,², Manuel Lopes¹, Alexandre Bernardino¹, Jonas Hörnstein¹, José Santos-Victor¹, Rolf Pfeifer²

Presented at ICRA’08, May 21, Pasadena, US

1) Instituto Superior Técnico Lisboa, VisLab, Portugal
2) University of Zurich, AILab, Switzerland
iCub’s Attention Framework

- Bottom-up attention system to control head and eye movements

- Visual Saliency features
  - Intensity
    - Grayscale conversion
    - Filtering with Mexican hat wavelets
  - Hue (color)
    - Calculate opponent-color channels
    - Maximum value of \(\{r', g', b', y'\}\) smoothed to single output saliency feature map

\[
\begin{align*}
    r' &= r_n - (g_n + b_n)/2 \\
    g' &= g_n - (r_n + b_n)/2 \\
    b' &= b_n - (r_n + g_n)/2 \\
    y' &= \frac{r_n + g_n}{2} - b_n - \|r_n - g_n\|
\end{align*}
\]
iCub’s Attention Framework

- Directional features
  - Gabor filters at 3 different scales and 4 different orientations (rotation)
- Motion
  - Reichard correlation (spatio temporal correlation model)

- Auditory Saliency features
  - Interaural time difference
    - Calculation of interaureal time difference by crosscorrelation gives azimuth angle (left-right) of the sound source
iCub’s Attention Framework

- Interaural spectral difference
  - The form of the ears changes the frequencies as a function $H$ of the location of the sound source
  - Spectral notches: Certain frequencies are canceled by wave reflected at the pinna
  - Spectral notch occurs when microphone-pinna distance is $n \times \frac{\lambda}{2} + \frac{\lambda}{4} = d \ (n = 0, 1, 2, ...)$ with wavelength $\lambda$
  - Calculate difference between spectra of the ears
  - Train classifier with minima of the difference spectrum as feature using different angles and distances
Multimodal integration of Saliency Maps

- Ego-sphere coherent representation of multimodal saliency
- Project stimulus intensity onto modality specific ego-sphere (polar coordinates)
- Fusion into single saliency ego-sphere by simply taking the maximum at each location
- Egocentric saliency map corresponds to robot’s short-term memory
  - Robot can shift attention to previously recognized salient regions
  - Decay at each timestep with a forgetting factor
Visual Exploration

• Inhibition-of-return: seek regions that have not been attended in the past → exploration
  • Habituation map H
    – Encodes getting used to persistent or repetitive stimuli and drawing attention towards novel salient regions
    – Gaussian around Focus of attention \( G_h(\theta - \theta_0, \varphi - \varphi_0) \)
  • Inhibition map A
    – Inhibition of regions close to the focus of attention for some time
    – Add Gaussian to A, when H exceeds threshold at specific location
  • Multiplication of Saliency map with A
  • Location with highest overall saliency is attended

\[
H(t) = H(t-1) + d_h (G_h(\theta - \theta_0, \varphi - \varphi_0) - H(t-1))
\]
\[
A(t) = A(t-1) + d_a (1.0 - A(t-1))
\]

33/40
Top Down Cues

• Most implementations of attention models are purely bottom-up, but

• Do you look into the sky when you are searching a person in a street scene?

• Knowledge of the outer world integrated in top-down fashion
  • Semantic category of the scene (aka gist)
  • Prior knowledge about the target
  • Current task
  • Experiences from prior tasks
  • Emotions, desires, motivations

• Implementations
  • Selection/weighting of feature detectors or regions
  • Tuning of conspicuity maps
Guided Visual Search Theory

- Jeremy Wolfe (currently version 4.0)
- Based on Treisman’s Feature Integration Theory
  - Feature maps describing difference to other stimuli at a particular position
  - Overall map of activations, that corresponds to a saliency map
  - Focus attention at location with highest activation
- Visual Search: Find target among distractor stimuli in a scene
- Activation map:
  - Representation of visual space where the activation at a location is the likelihood that the location contains a target
  - Weighted sum of the various sources of top-down and bottom-up guidance
  - Attention will be directed to the item with the highest priority (winner-take-all)
Guided Visual Search Theory

Basic Components of Guided Search

The Stimulus is filtered through broadly-tuned "categorical" channels.

The output produces feature maps with activation based on local differences (bottom-up) and task demands (top-down).

A weighted sum of these activations forms the Activation Map. In visual search, attention deploys limited capacity resources in order of decreasing activation.
Guided Visual Search Theory

- **Bottom-up in simple feature search**
  - Search for a vertical stimulus
- **The more an item differs from the others the higher the saliency**

- **Top-down in conjunction search**
  - Search for a black vertical stimulus
- **No special activation in saliency maps for target**
Guided Visual Search Theory

- Saliency calculation is assumed to be noisy
- Distractors may have higher activation than target
- → Serial search
  - Non target item is rejected
  - Attention moves on to the next item according to activation ranking of items
- Model can simulate many empirical findings very accurately
- Continuously extended by Wolfe
Modeling Attention

- Perception and attention are classic topics of cognitive psychology and have been strongly researched.
- This lecture can only give you a very basic understanding of basic models of perception and attention.
- There is a very large number of models of visual attention.
- Plus a number of models of other sensory modalities.
Modeling Attention

• Why should we model attention?
  • Understand human perceptual capabilities
  • Reduce the quantity of information to process
  • Experimental interests
    – Robotic vision and human-like behavior
    – Prediction of eye movements and active vision
    – Explanations for biological vision, e.g. neural activity
    – Etc.

• Which is the best model of attention?
• Comparing of cognitive models is often not possible or useful
  • Different focus
  • Level of granularity
  • Descriptive / computational
  • Theoretical / empirical