Biologically Inspired Attention

Daniel Harris

MS Thesis Defense

February 14, 2008

Committee

Chair: Dr. Roger Gaborski

Reader: Prof. Paul Tymann

Observer: Prof. Thomas Borrelli

What is attention?

"Attention is the cognitive process of selectively concentrating on one aspect of the environment while ignoring other things"

Does anything catch your attention here?

Perhaps the sign caught your attention?

A little bit more about attention...

- Is attention like a mental spotlight?
  - Theory proposed by Anne Treisman
  - Spotlight illuminates small areas of the visual field
  - Spotlight can jump around to different areas of interest
- Why would we evolve something like this?
  - Enormous amount of information being sent from the senses to the brain every second
  - By ignoring the majority of information received we can actually understand some of what is going on around us as it happens!

Great... how does this relate to computer vision?

- Object recognition gets all the love?
  - Many object recognition models have been built over the years...
    - Thomas Serre (MIT)
    - Edmund Rolls (Oxford University)
    - Roger Gaborski (RIT)
  - Most recognition models perform with over 95% accuracy, but....
  - What kind of input images are they using to recognize objects?

Object Recognition Input

- Input is fairly simple
- Focused on the target
- Very little in the background
- Objects takes up most of the input

What about “real world” input?

Perhaps visual attention is in order?

- "Real world" input is almost always too complex for a recognition model to accurately process
  - Can contain many different objects
  - Can contain many similar objects
  - Usually contains a lot of non-object noise or clutter
- Possible solutions
  - Build a more robust recognition model? No!
  - Use attention to reduce the size of the input? Yes!

Research Goals

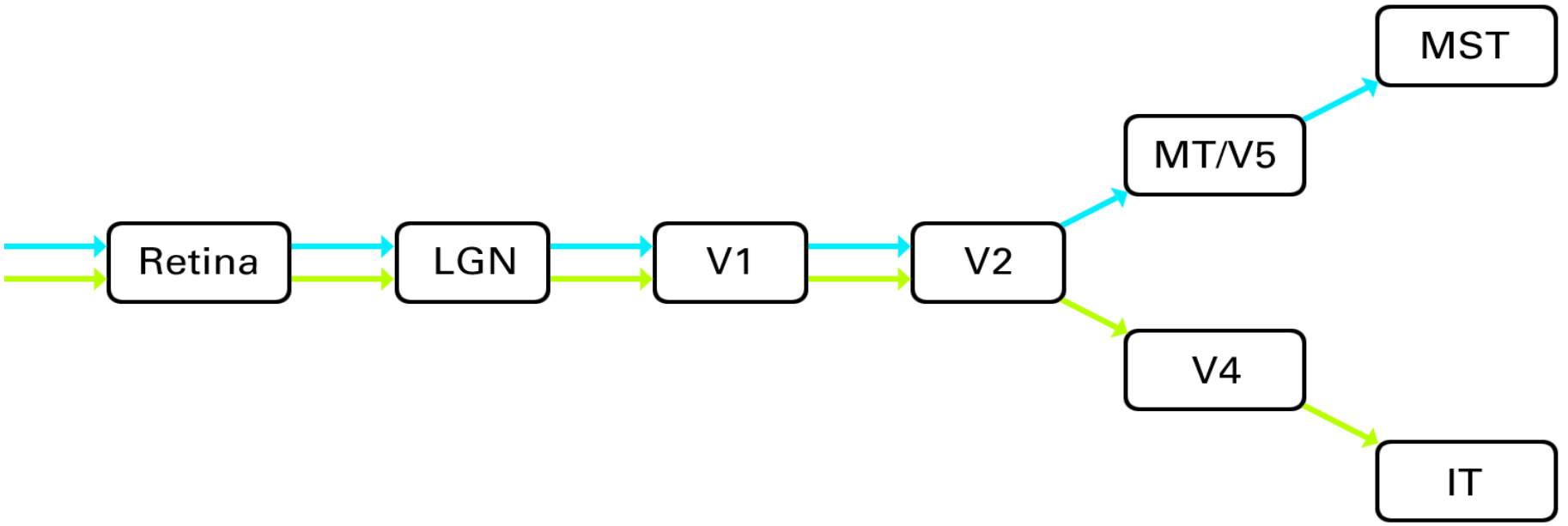
- Understand some of the biological and psychological concepts underlying visual attention
- Build a biologically inspired focus of attention model
  - Extract different “interesting” regions for each frame from a complex video for recognition
- Use attention model in conjunction with a recognition model to improve overall performance on complex input

Why build a biologically inspired model?

- The human visual system has been “in development” for thousands of years
- Many object recognition systems are biologically based and work very well
- Most of the information about how attention works is biologically or psychologically based
- Biological systems can teach us something about ourselves
- And well... it's darn cool!

Basic Biology

- Visual system is broken into two sections of cortical structures that visual information goes through to be processed.



Feature Integration Theory of Attention

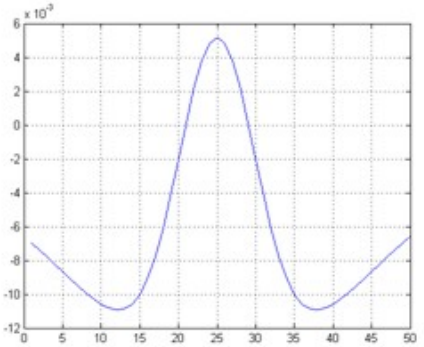
- Developed by Tresiman and Gelade in the 1980's
- Many different low level features are computed in parallel by the brain for a given input
- Individual features are combined later by the brain to help focus attention and processes a smaller area in more detail.
- Their work gave way to the concept of a saliency map
  - A map created by combining low level features which shows areas of high interest

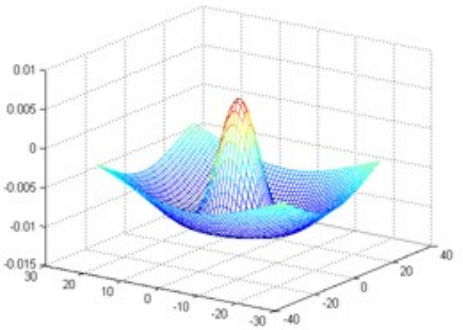
Building a saliency model

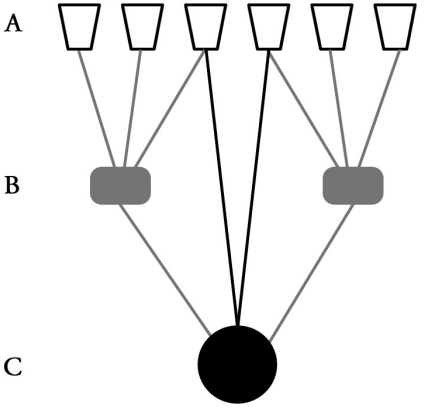
- Basic saliency model first described by Koch and Ullman in 1985
- Extract multiple contrast, orientation, and color-difference features from an input image.
  - Use biologically inspired filters to find low level features
- Make a feature map for each feature type
- Later...
  - Combine all of the feature maps into a saliency map
  - Find and shift focus of attention

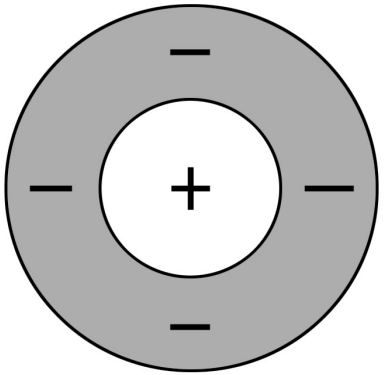
Contrast Feature Map

- Use Difference of Gaussian (DoG) filters
  - Derived from center/surround receptive fields found in brian
- 3 differently sized filters find areas of different contrast
- Combine 3 maps into single feature map









Contrast Feature Map Cont.

Original Image

contrast8

contrast16

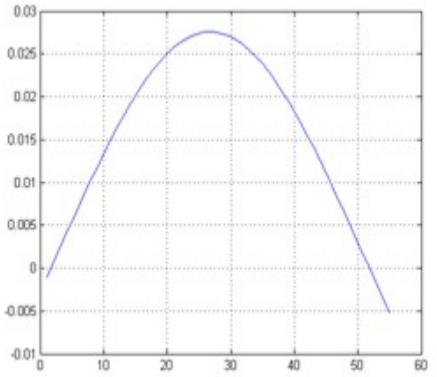
contrast32

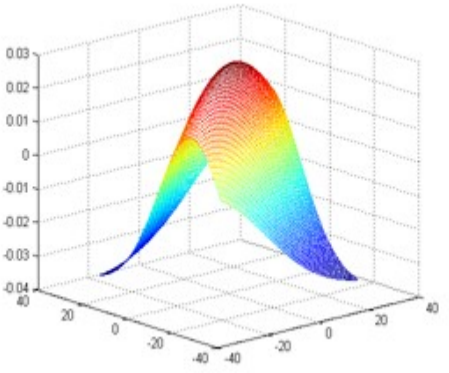
The Problem of Dynamic Range

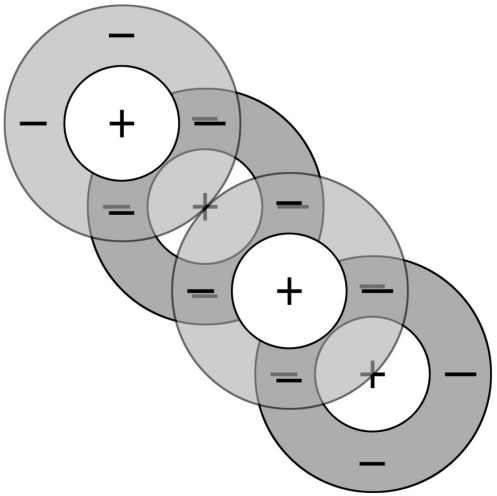
- Each sub-feature map is created with a different filter
  - Each filter's optimal response will vary
  - Filter 'X' may have a maximum response of 5
  - Filter 'Y' may have a maximum response of 500
- Must scale to the same dynamic range before combining sub-feature maps into a feature map
- Scale each sub-feature map by the maximum response of the filter used to create it
- Once scaled – Combine and normalize sub-feature maps

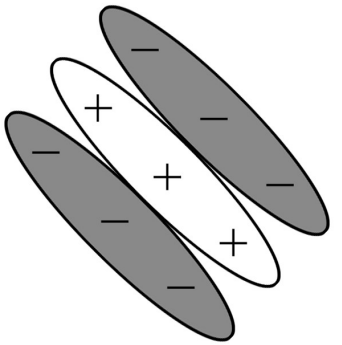
Orientation Feature Map

- Use Gabor filters
  - Derived from aligning center surround fields in LGN/V1
- Find edges of different orientation and size
  - 4 orientations, 4 sizes
- Combine into a single orientation feature map
  - Scale to same dynamic range first!

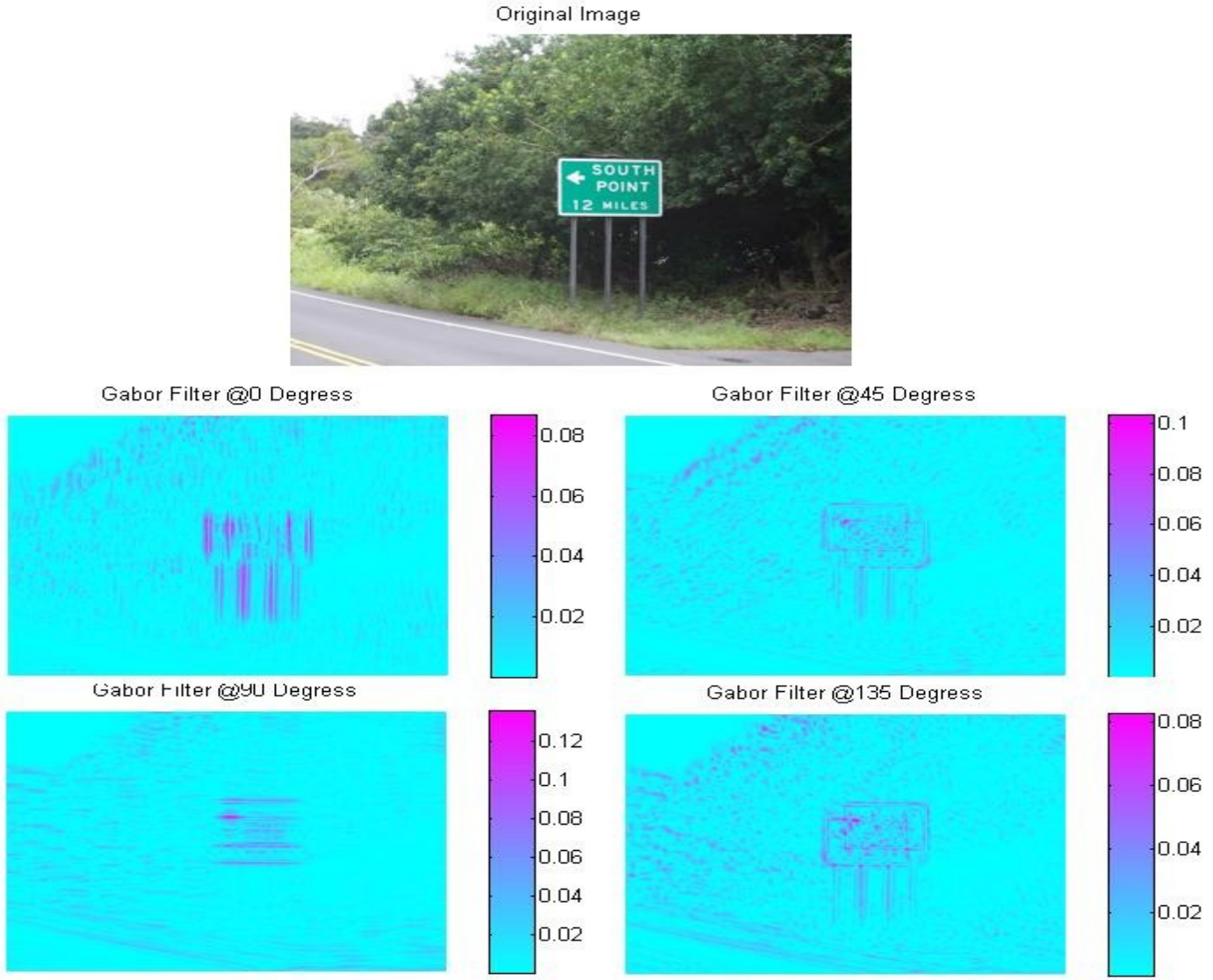








Orientation Feature Map Cont.



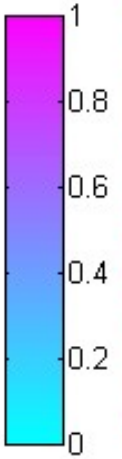
Color-Difference Feature Map

- Red-Green and Blue-Yellow color differences are calculated
  - Research has shown that mammals have sensitivity in V1 to these color differences
- Red-Green vs Green-Red does not matter
  - Only the difference matters, not which order
- Once a difference is taken, take the contrast of the color difference maps to find regions of high color difference contrast
- Scale to same dynamic range and combine

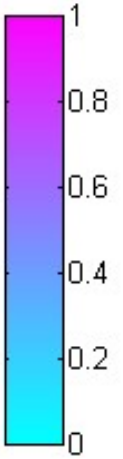
Color-Difference Feature Map Cont.

Original Image

Red-Green Difference



Blue-Yellow Difference

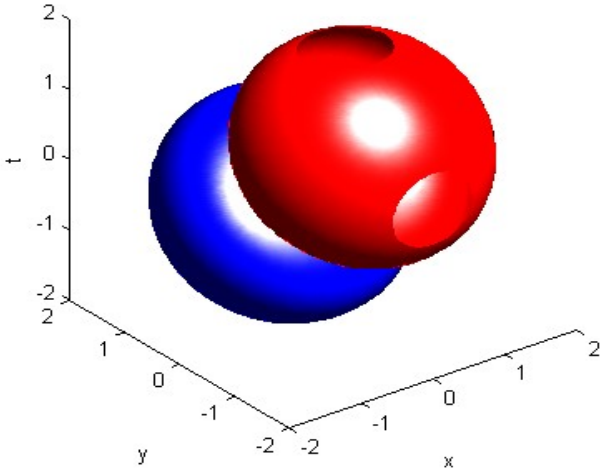


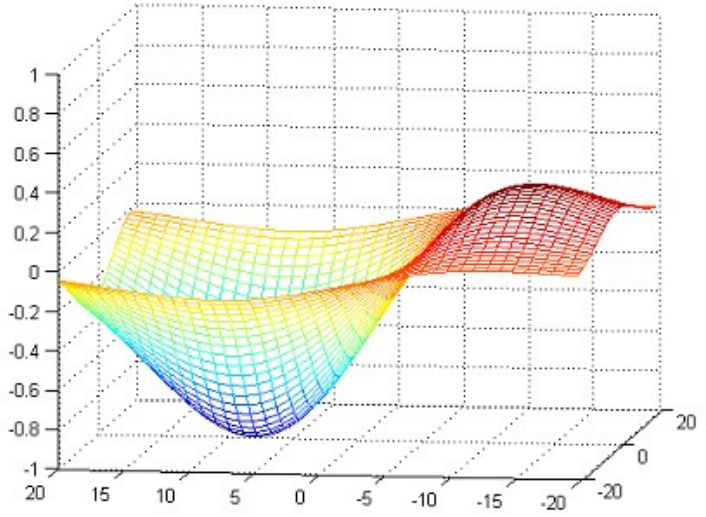
Advanced Saliency: Motion

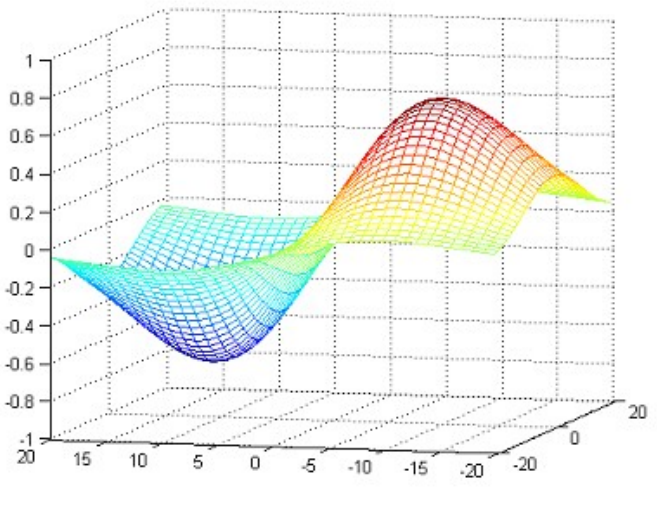
- Koch and Ullman were concerned mainly with static images
- Goal is to process a video
  - Motion is very “eye catching”
- Add a motion feature map into the saliency mix
- Biologically inspired spatio-temporal motion filters developed by Young and Lesperance
  - Essentially Gabor filters with an added aspect of time
  - Detect moving bars / edges

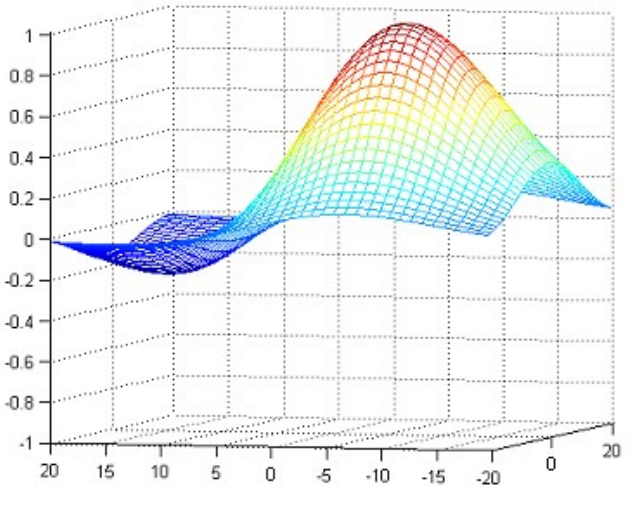
Motion Cont.

- Created by multiplying 3 derivative of Gaussian functions together
- Takes slices over the time axis to make multiple Gabor filters









Motion Cont.

- Remove all non-moving objects from input
  - Use blink filter
    - Special case of spatio-temporal filter
    - Detects only moving objects, direction of motion doesn't matter
- Detect motion in 4 primary directions
  - 0, 45, 90, 135 degrees
  - Approx. 1 pixel per frame of motion
- Combine motion sub-feature maps into single motion feature map
  - Remember to fix dynamic range!

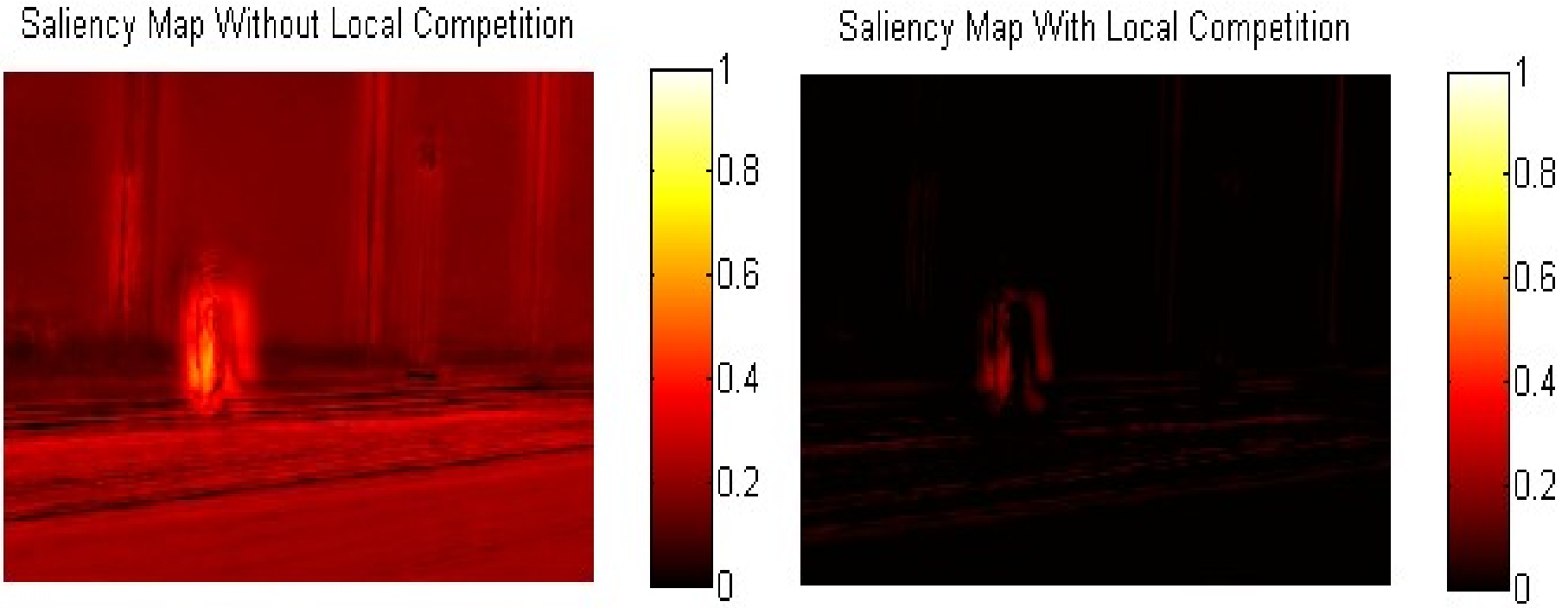
Saliency Map

- Combine all 4 feature maps to create a final saliency map
  - Dynamic range problem again, each of the feature maps has different meaning
- Itti and Koch experimented with combination strategies
  - Naive
  - Training
  - Global Amplification
  - Local Competition

Saliency Map Cont.

- Use local competition
  - Most biologically based approach
  - Otherwise known as “lateral inhibition”
- Single pass a wide DoG filter over each feature map
  - Single pass for speed (Itti and Koch use 12)
- Use naïve approach to combine each inhibited feature map.
  - After local competition they each map can be considered to be in the same dynamic range

Saliency Map Cont.

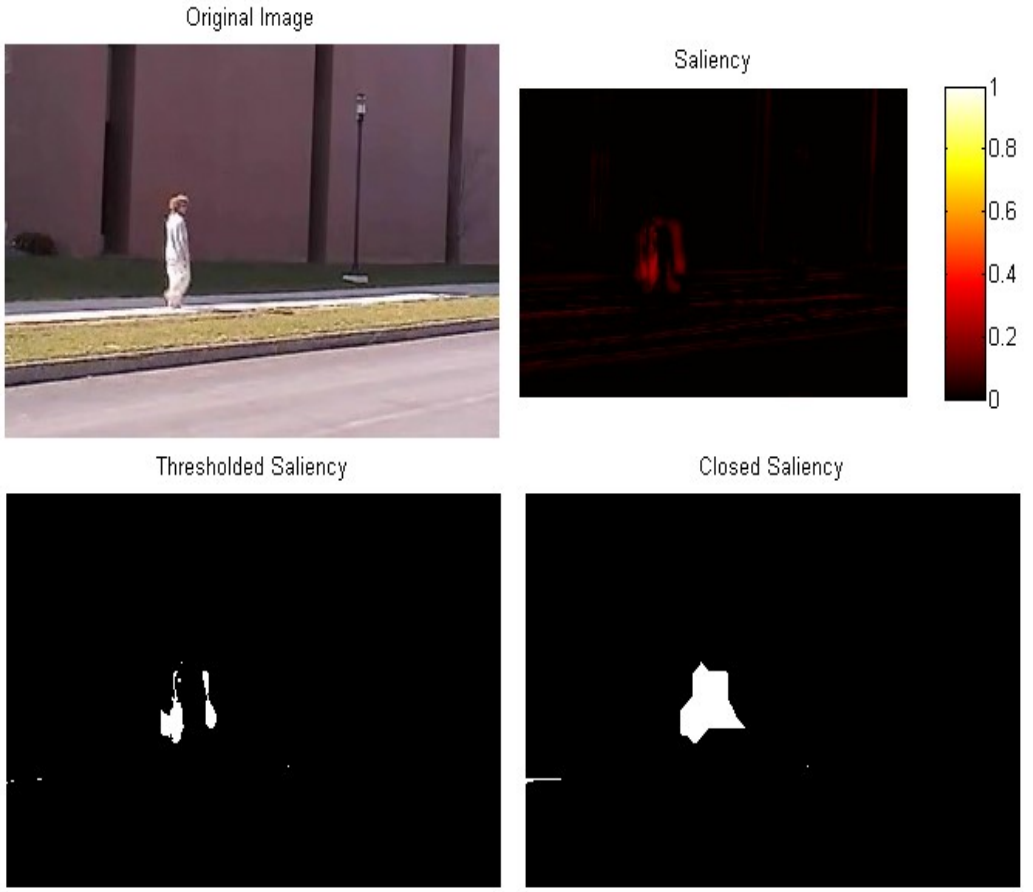


Focusing Attention

- Two tasks need to be completed
  - Find a region of interest
  - Find the right approximate size of the region
    - Goal is to extract whole objects
- Use some basic computer vision techniques to extract the information we need from the saliency map
  - thresholding
  - image closures
  - region labeling

Focusing Attention Cont.

- Threshold saliency map (> 0.5)
- Perform Image Closure
- Label Regions
- Find sufficiently large region with highest average salience
- Lower threshold and repeat if no objects are found



Shifts in Attention

- Finding a single region of attention is relatively easy
- How can we find a region for the next frame?
  - Do the exact same thing on the next frame?
    - No. If the video doesn't change much (or all) from frame to frame the same region will be selected over and over.
- Use Habituation and Dishabituation to force shifts from frame to frame
- Use proximity and similarity preference to make shifts more psychologically plausible

Habituation, Dishabituation, and Proximity Preference

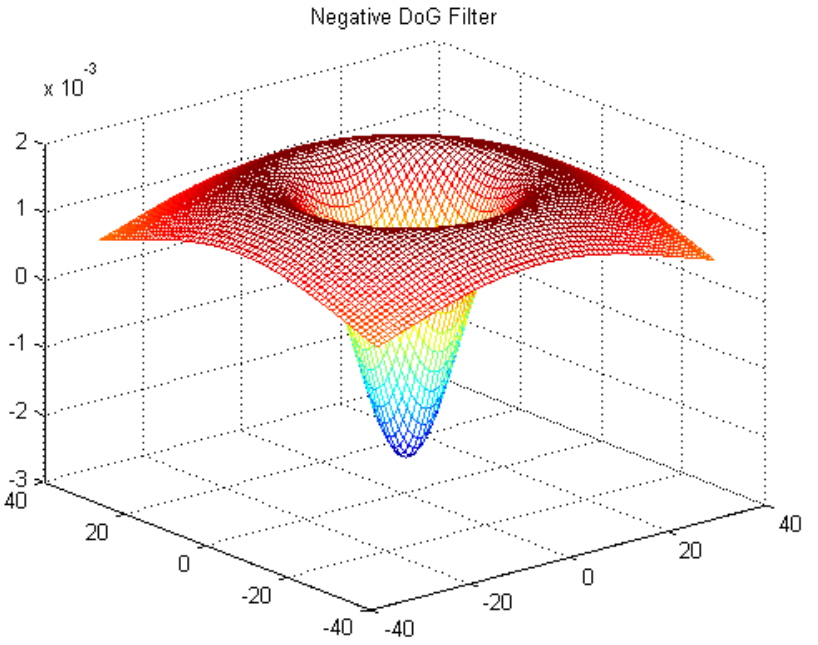
- Habituation is the lowering of a response when presented with the same stimulus
- Dishabituation is the strengthening of a response that was previously subject to habituation
- Proximity preference is the concept that attentional shifts will occur between regions that are spatially close together
  - Rather than jumping all over the input

Habituation, Dishabituation, and Proximity Preference Cont.

- Can kill three birds with one stone!
- Create a short-term memory of recently attended regions
  - Before selecting a new region each frame, inhibit all of the previously seen regions (habituation)
  - Linearly scale the amount each region is inhibited based on how long has passed since it has been focused on.
    - Over time the region will dishabituate back to “normal”
- Use a negative DoG filter to accomplish the inhibition

Habituation, Dishabituion, and Proximity Preference Cont.

- Make the entire negative region of the negative DoG fill the selected area
  - This will inhibit the previously seen region
- The positive area of the negative DoG filter will excite regions spatially close to the selected region

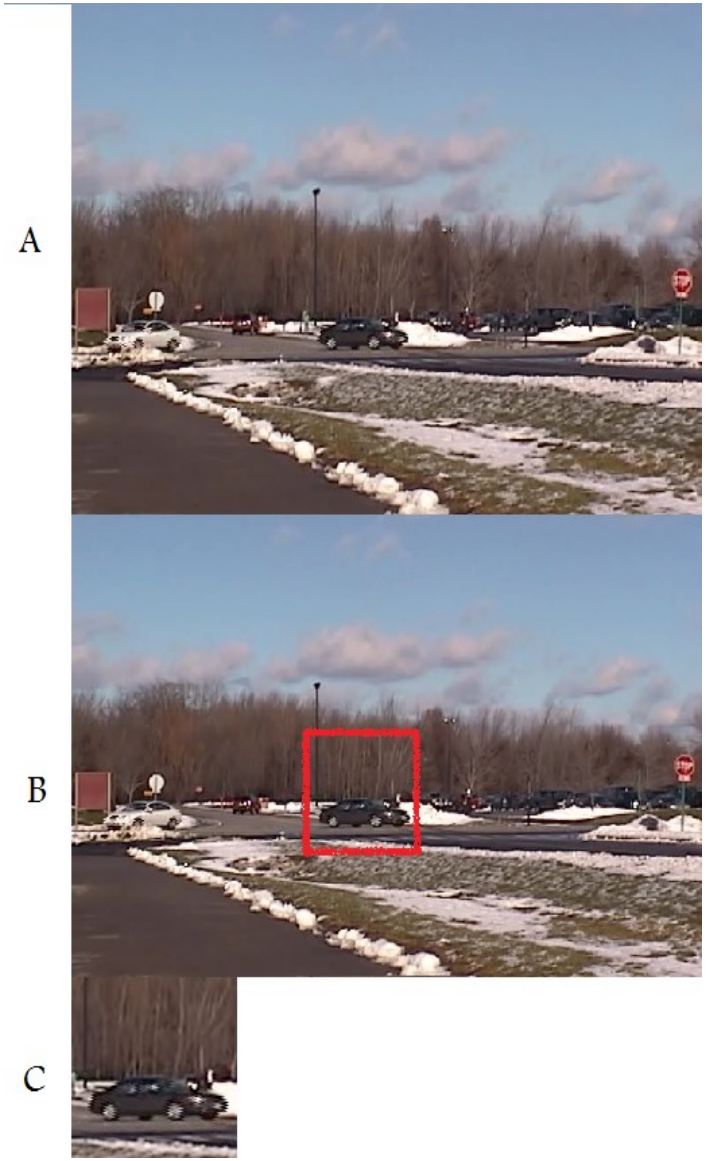


Similarity Preference

- Shifts in attention are more likely to occur between objects/areas that have similar features
  - Anne Treisman's concept of perceptual grouping
- At present, saliency map is created by equally weighting each feature map
  - Alter weights of each feature map according to the feature composition of the previously selected region
  - Increase the effect of each feature map by up to 25% depending on composition of selected region

Model Output

- Three 4 Types of Output
  - Raw video frame
  - Boxed region of attention
  - Extracted region of attention
  - Textual output
    - X/Y coordinate
    - Width/Height
    - Frame #



Output Classifications

- Simulations and Results

Single Target Input

Single Target Results

Table 8: Single Target Simulation Results

	Partial Acquisition		Optimal Acquisition	
	Time Elapsed	Total Frames	Time Elapsed	Total Frames
Simulation 1	0.5 sec	26	1.8 sec	3
Simulation 2	-	-	0.3	6

Double Target Input

Double Target Results

Table 9: Double Target Simulation Results

	Partial Acquisition		**Optimal Acquisition**	
	Time Elapsed	Total Frames	Time Elapsed	Total Frames
Simulation 1, Target 1	0.1 sec	18	-	-
Simulation 1, Target 2	N/A sec	N/A	1.6 sec	3
Simulation 2, Target 1	0.0 sec	28	0.76 sec	10
Simulation 2, Target 2	0.3 sec	12	2.6 sec	3

Performance

- Videos are not processed in real time
  - Pre-recorded then processed frame by frame
- How long does it take to process each frame
  - What part of the model is taking the most amount of time?

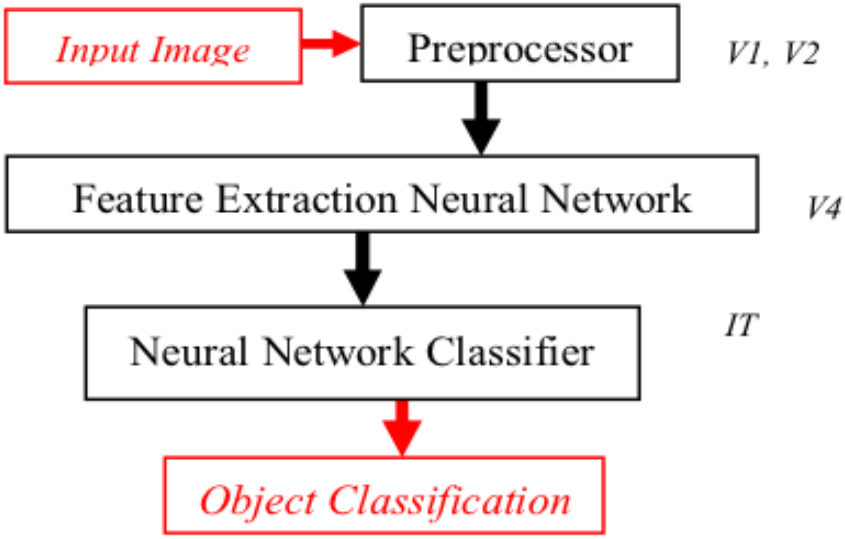
Table 11: Attention Model Benchmarks

Category	Computer 1	Computer 2
Processor	2x 1.4Ghz	2x 2.2Ghz
Intensity Contrast	0.49 sec	0.16 sec
Orientation	4.75 sec	1.77 sec
Color Difference	0.88 sec	0.34 sec
Motion	34.10 sec	12.46 sec
Attention	8.08 sec	3.38 sec
Overall Time	48.3 sec	18.11

Integration with an Object Recognition Model

Object Categorization Model

- Developed at RIT by Dr. Gaborski, Myung Woo, and Theparit Peerasathien
- Neural networked based model that can categorize different objects in an input.
- On simple input, over 95% accurate!



Sample Input Used

Results

- Object categorization model outputs a 3 unit vector
  - Car Present: [0 1 0]
  - No Car: [1 0 1]
- Model was trained to output [1 0 1] or [0 1 0]
  - Results are actually decimals which can be thresholded

Table 10: Car Categorization Results

	Low Resolution	
	Extracted Region	Full Frame
Trial 1	[0.005 0.995 0.005]	[0.833 0.161 0.855]
Trial 2	[0.085 0.932 0.077]	[0.675 0.300 0.681]
Trial 3	[0.022 0.978 0.024]	[0.620 0.385 0.641]
Trial 4	[0.018 0.980 0.019]	[0.058 0.959 0.049]
Trial 5	[0.042 0.954 0.047]	[0.340 0.646 0.371]

Future Work

Integration with Recognition Models

- Remember...
  - The goal is to create an attention model to use in conjunction with a recognition model
- Many recognition models use feature extraction to recognize objects
  - Thomas Serre model from MIT
  - Gaborski model from RIT
- Both models can and should share the feature maps extracted so they do not have to recalculate
  - Will help build a real-time system

Real-Time Performance

- As seen from benchmarks, attention model is not real time
  - Developed in MATLAB, moving to a compiled language will increase performance
  - Replace biologically plausible filters
    - Biologically plausible filters require many convolution operations (computationally expensive)
    - Replace with non-biological operations
    - Frame differencing instead of spatio-temporal motion filters
    - global amplification instead of local competition
  - Compute each feature or sub-feature map in parallel

Pre-Cueing

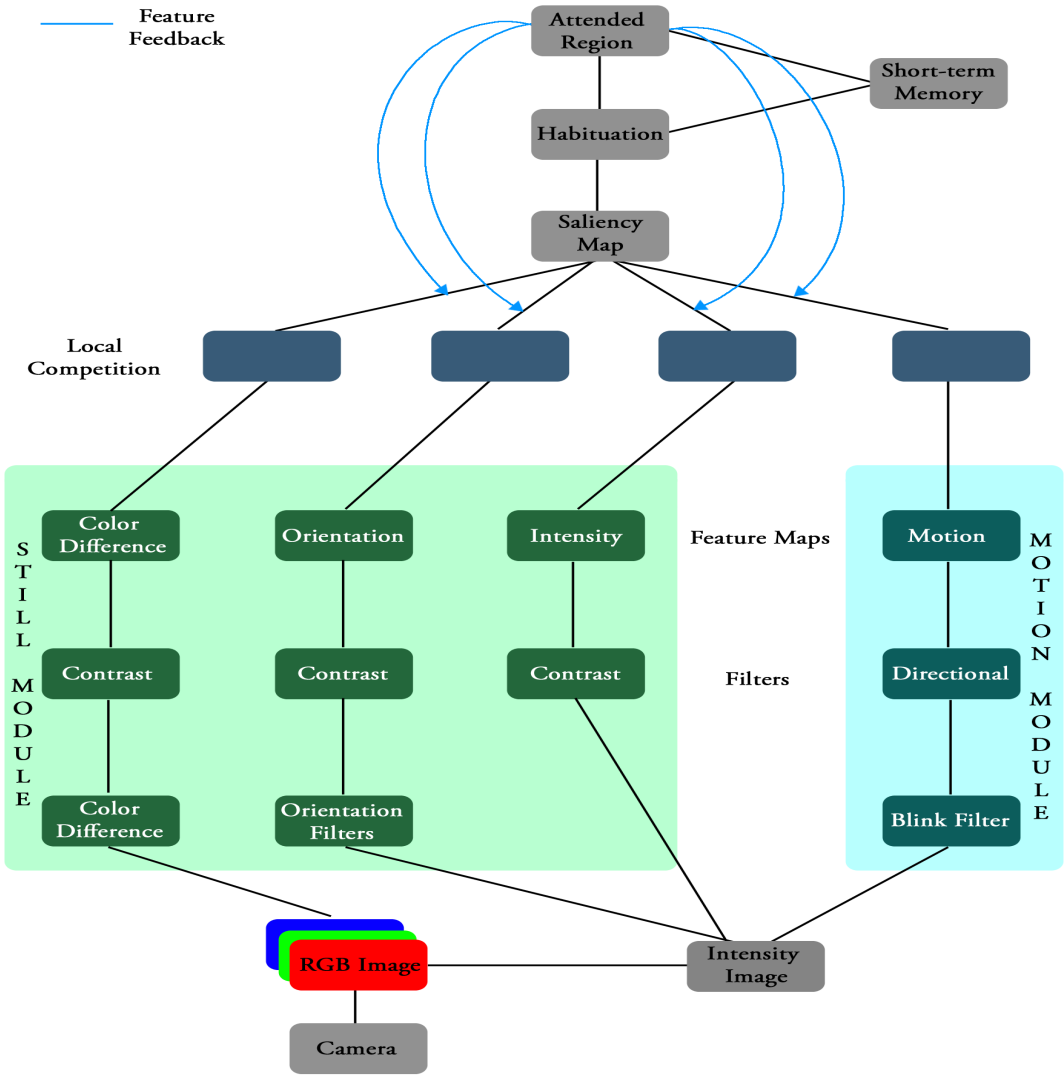
- It is possible to pre-cue the attention model
  - Make it more likely to find an object with a particular feature
  - Make it more likely to find an object in a particular location in the input field
- Accomplish this by altering the weights of how feature maps are combined into the saliency map after local competition
  - Just like the implementation for similarity preference

Object Tracking

- Attention model can be modified to track objects once identified by a recognition model
  - Suspend any habituation/dishabituation
    - Makes it likely the same object will be selected again
  - Make the feature feedback stronger than usual
    - Tweaks the saliency map to respond strongly to areas which are very similar to the previously selected area
- This will not ensure that the same object is re-selected, but it will be much more likely

Conclusions

- Accurately finds the location interesting objects
  - Does not always extract them optimally
- Easy integration with recognition models
- Improves performance of recognition model
- Potential for expandability



That's All Folks!

Questions / Comments