Collaborators

Chris Kanan

Visual Salience

- *Visual Salience* is some notion of what is interesting in the world - it captures our attention.
- Visual salience is important because it drives a decision we make a **couple of hundred thousand times a day** - where to look.

Data We Want to Explain

- Visual search:
  - Search asymmetry: A search for one object among a set of distractors is faster than vice versa.
  - Parallel vs. serial search (and the continuum in between): An item “pops out” of the display no matter how many distractors vs. reaction time increasing with the number of distractors (not emphasized in this talk...)
- Eye movements when viewing images and videos.
Audience participation!

Look for the unique item

Clap when you find it
What just happened?

- This phenomenon is called the visual search asymmetry:
  - Tilted bars are more easily found among vertical bars than vice-versa.
  - Backwards “s”s are more easily found among normal “s”s than vice-versa.
  - Upside-down elephants are more easily found among right-side up ones than vice-versa.

Why is there an asymmetry?

- There are not too many computational explanations:
  - “Prototypes do not pop out”
  - “Novelty attracts attention”
- Our model of visual salience will naturally account for this.

Saliency Maps

- Koch and Ullman, 1985: the brain calculates an explicit saliency map of the visual world
- Their definition of saliency relied on center-surround principles
  - Points in the visual scene are salient if they differ from their neighbors
- In more recent years, there have been a multitude of definitions of saliency
Saliency Maps

- There are a number of candidates for the saliency map: there is at least one in LIP, the Lateral Intraparietal Sulcus, a region of the parietal lobe, also in the frontal eye fields, the superior colliculus,... but there may be representations of salience much earlier in the visual pathway - some even suggest in V1.
- But we won’t be talking about the brain today...

Probabilistic Saliency

- Our basic assumption:
  - The main goal of the visual system is to find potential targets that are important for survival, such as prey and predators.
  - The visual system should direct attention to locations in the visual field with a high probability of the target class or classes.
  - We will lump all of the potential targets together in one random variable, \( T \).
  - For ease of exposition, we will leave out our location random variable, \( L \).

Notation:

- \( x \) denotes a point in the visual field
- \( T_x \): binary variable signifying whether point \( x \) belongs to a target class
- \( F_x \): the visual features at point \( x \)
- The task is to find the point \( x \) that maximizes
  
  \[ p(T_x | F_x) \]

  the probability of a target given the features at point \( x \)
- This quantity is the saliency of a point \( x \)
- Note: This is what most classifiers compute!

Taking the log and applying Bayes’ Rule results in:

\[
\log p(T_x | F_x) = \log \frac{p(F_x | T_x) p(T_x)}{p(F_x)} \\
= \log p(F_x | T_x) + \log p(T_x) + \log \frac{1}{p(F_x)}
\]
Log $p(F_x | T_x)$
- Probabilistic description of the features of the target
- Provides a form of top-down (endogenous, intrinsic) saliency
- Some similarity to Iconic Search (Rao et al., 1995) and Guided Search (Wolfe, 1989)

-log $p(F_x)$
- This is called the self-information of this variable
- It says that rare feature values attract attention
- Independent of task
- Provides notion of bottom-up (exogenous, extrinsic) saliency

Now we have two terms:
- Top-down saliency
- Bottom-up saliency
- Taken together, this is the pointwise mutual information between the features and the target
For most of what I will be telling you about next, we use only the \(-\log p(F)\) term, or bottom up salience.

Remember, this means rare feature values attract attention.

This is a computational instantiation of the idea that "novelty attracts attention".

Remember, this means rare feature values attract attention.

This means two things:

- We need some features (that have values)! What should we use?
- We need to know when the values are unusual: So we need \textit{experience}.

Experience, in this case, means collecting statistics of how the features respond to natural images.

We will use two kinds of features:

- Difference of Gaussians (DOGs)
- Independent Components Analysis (ICA) derived features

These respond to differences in brightness between the center and the surround. We apply them to three different color channels separately (intensity, Red-Green and Blue-Yellow) at four scales: 12 features total.
Feature Space 1: Differences of Gaussians

- Now, we run these over Lingyun’s vacation photos, and record how frequently they respond.

Learning the Distribution

We fit a generalized Gaussian distribution to the histogram of each feature.

\[
p\left(F_i; \sigma_i, \theta_i\right) = \frac{\theta_i}{2\sigma_i \Gamma\left(\frac{1}{\theta_i}\right)} \exp\left(-\frac{\left|F_i\right|^\theta_i}{\sigma_i}\right)
\]

where \(F_i\) is the \(i^{th}\) filter response, \(\theta_i\) is the shape parameter and \(\sigma_i\) is the scale parameter.

\[
\log p\left(F_i\right) = \text{const.} - \frac{\left|F_i\right|^\theta_i}{\sigma_i}
\]

The Learned Distribution (DOGs)

- This is \(P(F)\) for four different features.
- Note these features are \textit{sparse} - I.e., their most frequent response is near 0.
- When there is a big response (positive or negative), it is interesting!
The Learned Distribution (ICA)

- For example, here’s a feature:
- Here’s a frequency count of how often it matches a patch of image:
- Most of the time, it doesn’t match at all - a response of “0”
- Very infrequently, it matches very well - a response of “200”

Bottom-up Saliency

- We have to estimate the joint probability from the features.
- If all filter responses are independent:
  \[ -\log p(F) = -\sum_i \log p(F_i) \]
- They’re not independent, but we proceed as if they are. (ICA features are “pretty independent”)
- Note: No weighting of features is necessary!

Qualitative Results: BU Saliency

Original Image | Human fixations | DOG Salience | ICA Salience
--- | --- | --- | ---

Original Image | Human fixations | DOG Salience | ICA Salience
--- | --- | --- | ---

Original Image | Human fixations | DOG Salience | ICA Salience
--- | --- | --- | ---

Original Image | Human fixations | DOG Salience | ICA Salience
--- | --- | --- | ---
Qualitative Results: BU Saliency

Quantitative Results: BU Saliency

- These are quantitative measures of how well the salience map predicts human fixations in static images.
- We are best in the KL distance measure, and second best in the ROC measure.
- Our main competition is Bruce & Tsotsos, who have essentially the same idea we have, except they compute novelty in the current image.

Related Work

- Torralba et al. (2003) derives a similar probabilistic account of saliency, but:
  - Uses current image’s statistics
  - Emphasizes effects of global features and scene gist
- Bruce and Tsotsos (2006) also use self-information as bottom-up saliency
  - Uses current image’s statistics

Related Work

- The use of the current image’s statistics means:
  - These models follow a very different principle: finds rare feature values in the current image instead of unusual feature values in general: novelty.
- As we’ll see, novelty helps explain several search asymmetries
- Models using the current image’s statistics are unlikely to be neurally computable in the necessary timeframe, as the system must collect statistics from entire image to calculate local saliency at each point
Search Asymmetry

- Our definition of bottom-up saliency leads to a clean explanation of several search asymmetries (Zhang, Tong, and Cottrell, 2007)
  - All else being equal, targets with uncommon feature values are easier to find
- Examples:
  - Treisman and Gormican, 1988 - A tilted bar is more easily found among vertical bars than vice versa
  - Levin, 2000 - For Caucasian subjects, finding an African-American face in Caucasian faces is faster due to its relative rarity in our experience (basketball fans who have to identify the players do not show this effect).

Search Asymmetry Results

Top-down salience in Visual Search

- Suppose we actually have a target in mind - e.g., find pictures, or mugs, or people in scenes.
- As I mentioned previously, the original (stripped down) salience model can be implemented as a classifier applied to each point in the image.
- When we include location, we get (after a large number of completely unwarranted assumptions):

\[
\log \text{salience}_T = - \log p(F = f_x) + \log p(F = f_x | T_x = 1) + \log p(T_x = 1 | I = l)
\]
Qualitative Results (mug search)

- Where we disagree the most with Torralba et al. (2006)
  - GIST
  - SUN

Qualitative Results (picture search)

- Where we disagree the most with Torralba et al. (2006)
  - GIST
  - SUN

Qualitative Results (people search)

- Where we agree the most with Torralba et al. (2006)
  - GIST
  - SUN

Qualitative Results (painting search)

- This is an example where SUN and humans make the same mistake due to the similar appearance of TV’s and pictures (the black square in the upper left is a TV!).
Quantitative Results

Area Under the ROC Curve (AUC) gives basically identical results.

Saliency of Dynamic Scenes

- Created spatiotemporal filters
  - Temporal filters: Difference of exponentials (DoE)
    - Highly active if change
    - If features stay constant, goes to zero response
    - Resembles responses of some neurons (cells in LGN)
    - Easy to compute
  - Convolve with spatial filters to create spatiotemporal filters

Bayesian Saliency (Itti and Baldi, 2006):

- Saliency is Bayesian “surprise” (different from self-information)
- Maintain distribution over a set of models attempting to explain the data, P(M)
- As new data comes in, calculate saliency of a point as the degree to which it makes you alter your models
  - Total surprise: \( S(D, M) = KL(P(M|D); P(M)) \)
  - Better predictor than standard spatial salience
  - Much more complicated (~500,000 different distributions being modeled) than SUN dynamic saliency (days to run vs. hours or real-time)

In the process of evaluating and comparing, we discovered how much the center-bias of human fixations was affecting results.

Most human fixations are towards the center of the screen (Reinagel, 1999)

Accumulated human fixations from three experiments
Results varied widely depending on how edges were handled

- How is the invalid portion of the convolution handled?

 accumulated saliency of three models

Measures of Dynamic Saliency

- Typically, the algorithm is compared to the human fixations within a frame
  - I.e., how salient is the human-fixated point according to the model versus all other points in the frame
  - This measure is subject to the center bias - if the borders are down-weighted, the score goes up

Measures of Dynamic Saliency

- An alternative is to compare the salience of the human-fixated point to the same point across frames
  - Underestimates performance, since often locations are genuinely more salient at all time points (e.g., an anchor's face during a news broadcast)
  - Gives any static measure (e.g., centered-Gaussian) a baseline score of 0.
  - This is equivalent to sampling from the distribution of human fixations, rather than uniformly
  - On this set of measures, we perform comparably with (Itti and Baldi, 2006)
Saliency of Dynamic Scenes

<table>
<thead>
<tr>
<th>Method</th>
<th>KL</th>
<th>ROC area</th>
<th>% above median</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0</td>
<td>0.5</td>
<td>50%</td>
</tr>
<tr>
<td>Bayesian Surprise</td>
<td>0.0344</td>
<td>0.5808</td>
<td>61.66%</td>
</tr>
<tr>
<td>Dynamic Saliency</td>
<td>0.0409</td>
<td>0.5818</td>
<td>62.39%</td>
</tr>
</tbody>
</table>

Results using non-center-biased metrics on the human fixation data on videos from Itti (2005) - 4 subjects/movie, 50 movies, ~25 minutes of video.
Summary of this part of the talk

- It is a good idea to start from first principles.
- Often the simplest model is best
- Our model of salience rocks.
  - It does bottom up
  - It does top down
  - It does video (fast!)
  - It naturally accounts for search asymmetries

Summary and Conclusions

- But, as is usually the case with grad students, Lingyun didn’t do everything I asked…
- We are beginning to explore models based on utility: Some targets are more useful than others, depending on the state of the animal
- We are also looking at using our hierarchical ICA model, to get higher-level features
Summary and Conclusions

- And a foveated retina,
- And updating the salience based on where the model looks (as is actually seen in LIP).

Motivation

- Now we have a model of salience - but what can it be used for?
- Here, we show that we can use it to recognize objects.

One reason why this might be a good idea...

- Our attention is automatically drawn to interesting regions in images.
- Our salience algorithm is automatically drawn to interesting regions in images.
- These are useful locations for discriminating one object (face, butterfly) from another.
Main Idea

- **Training Phase (learning object appearances):**
  - Use the salience map to decide where to look. (We use the ICA salience map)
  - Memorize these samples of the image, with labels (Bob, Carol, Ted, or Alice) (We store the ICA feature values)

Christopher Kanan

Main Idea

- **Testing Phase (recognizing objects we have learned):**
  - Now, given a new face, use the salience map to decide where to look.
  - Compare new image samples to stored ones - the closest ones in memory get to vote for their label.

Christopher Kanan

Voting

- The voting process is actually based on Bayesian updating (and the Naïve Bayes assumption).
- The size of the vote depends on the distance from the stored sample, using kernel density estimation.
- Hence NIMBLE: NIM with Bayesian Likelihood Estimation.

Christopher Kanan
Overview of the system

- The ICA features do double-duty:
  - They are combined to make the salience map - which is used to decide where to look
  - They are stored to represent the object at that location

NIMBLE vs. Computer Vision

- Compare this to standard computer vision systems:
  - One pass over the image, and global features.

NIMBLE Example

Belief After 1 Fixation
Belief After 10 Fixations
Robust Vision

- Human vision works in multiple environments - our basic features (neurons!) don’t change from one problem to the next.
- We tune our parameters so that the system works well on Bird and Butterfly datasets - and then apply the system *unchanged* to faces, flowers, and objects
- This is very different from standard computer vision systems, that are tuned to particular sets

Datasets

- Cal Tech 101: 101 Different Categories
- AR dataset: 120 Different People with different lighting, expression, and accessories

Data: Flowers: 102 Different Flower Species

- ~7 fixations required to achieve at least 90% of maximum performance
That’s nice

- So, we created a simple cognitive model that uses simulated fixations to recognize things.
  - But it isn’t that complicated.
- How does it compare to approaches in computer vision?

Percent Increase Over Some of the Best Methods in Computer Vision

- Caveats:
  - As of mid-2010.
  - Only comparing to single feature type approaches (no “Multiple Kernel Learning” (MKL) approaches).
  - Still superior to MKL with very few training examples per category.

Caveats:

- As of mid-2010.
- Only comparing to single feature type approaches (no “Multiple Kernel Learning” (MKL) approaches).
- Still superior to MKL with very few training examples per category.
More neurally and behaviorally relevant gaze control and fixation integration.
- People don’t randomly sample images.
- A foveated retina
- Comparison with human eye movement data during recognition/classification of faces, objects, etc.

Summary
- A fixation-based approach can work well for image classification.
- Fixation-based models can achieve, and even exceed, some of the best models in computer vision.
...Especially when you don’t have a lot of training images.

Thanks!
- Software and Paper Available at
  [www.chriskanan.com](http://www.chriskanan.com)
- ckanan@ucsd.edu

This work was supported by the NSF (grant #SBE-0542013) to the Temporal Dynamics of Learning Center.
Thanks!