What we see is most likely to be what matters: Visual attention and applications

O. Le Meur
olemeur@irisa.fr

P. Le Callet
patrick.lecallet@univ-nantes.fr

http://www.irisa.fr/temics/staff/lemeur/

November 9, 2009
Yarbus [Yarbus, 1967] demonstrated how eye movements changed depending on the question asked of the subject.

1. No question asked
2. Judge economic status
3. What were they doing before the visitor arrived?
4. What clothes are they wearing?
5. Where are they?
6. How long is it since the visitor has seen the family?
7. Estimate how long the unexpected visitor had been away from the family.

Each recording lasted 3 minutes.
1 Introduction

2 Computational models: two schools of thought

3 A survey of computational models of bottom-up visual attention
   - Hierarchical models
   - Statistical models
   - Some examples

4 Performance assessments
   - Building a ground truth...
   - Metrics
   - Limitations

5 Applications
   - Retargeting
   - Compression
   - Others

6 Conclusion
Introduction

2 Computational models: two schools of thought

3 A survey of computational models of bottom-up visual attention

4 Performance assessments

5 Applications

6 Conclusion
For the computational modelling, two ‘schools’ can be considered:

- One based on the assumption that there is an unique saliency map [Koch et al., 1985][Li, 2002]:

**Definition (saliency map)**

A topographic representation that combines the information from the individual feature maps into one global measure of conspicuity. This map can be modulated by a higher-level feedback.

A comfortable view for the computational modelling...
For the computational modelling, two ‘schools’ can be considered:

- There exist **multiple saliency maps** (distributed throughout the visual areas) [Tsotsos et al., 1995]).

Many candidate locations for a saliency map:

- Primary visual cortex [Li, 2002]
- Lateral Intraparietal area (LIP) [Kusunoki et al., 2000]
- Medial Temporal cortex [Treue et al., 2006]

‘At each level, saliency can thus be used as a gain control mechanism to spatially gate relevant information for the next processing level. From [Van Rullen, 2003].
1 Introduction

2 Computational models: two schools of thought

3 A survey of computational models of bottom-up visual attention
   - Hierarchical models
   - Statistical models
   - Some examples

4 Performance assessments

5 Applications

6 Conclusion
Hierarchical models

Itti’s model [Itti et al., 1998], probably the most known...

- Based on the Koch and Ullman’s scheme
- Hierarchical decomposition (Gaussian)
- Early visual features extraction in a massively parallel manner
- Center-surround operations
- Pooling of the feature maps to form the saliency map
Hierarchical models

Le Meur’s model [Le Meur et al., 2006], an extension of Itti’s model...

- Based on the Koch and Ullman’s scheme
- Light adaptation and Contrast Sensitivity Function
- Hierarchical and oriented decomposition (Fourier spectrum)
- Early visual features extraction in a massively parallel manner
- Center-surround operations on each oriented subband
- Enhanced pooling [Le Meur et al., 2007] of the feature maps to form the saliency map

Other models in the same vein: [Marat et al., 2009], [Bur et al., 2007]...
Statistical models are based on a probabilistic framework taken their origin in the information theory.

**Definition (Self-information)**

Self-information is a measure of the amount information provided by an event. For a discrete $X$ r.v defined by $A = \{x_1, ..., x_N\}$ and by a pdf, the amount of information of the event $X = x_i$ is given by:

$$I(X = x_i) = -\log_2 p(X = x_i), \text{ bit/symbol}$$

**Properties**

- if $p(X = x_i) < p(X = x_j)$ then $I(X = x_i) > I(X = x_j)$
- $p(X = x_i) \rightarrow 0, I(X = x_i) \rightarrow +\infty$

The saliency of visual content could be deduced from the self-information measure.

**Self-information ≡ rareness, surprise, contrast...**
Statistical models

- First model resting on this approach has been proposed in 2003 [Oliva et al., 2003]: \[ S(x) = \frac{1}{P(v_f(x))} \], where \( v_f \) is a feature vector (48 dimensions), the pdf is computed over the whole image.

- Bruce in 2004 [Bruce, 2004] and 2009 [Bruce et al., 2009] modified the previous approach by using the self-information locally on independent coefficients (projection on a given basis).

Other models in the same vein: [Mancas et al., 2006], [Zhang, 2008].
Some examples

(a) Original    (b) Itti    (c) Le Meur    (d) Bruce    (e) Zhang
1 Introduction

2 Computational models: two schools of thought

3 A survey of computational models of bottom-up visual attention

4 Performance assessments
   - Building a ground truth...
   - Metrics
   - Limitations

5 Applications

6 Conclusion
Ground truth

A good review of parsing algorithm in [Salvucci et al., 2000].

On the web (for natural images):
- Bruce's database http://www-sop.inria.fr/members/Neil.Bruce/
- Le Meur's database (28 color pictures) http://www.irisa.fr/temics/staff/lemeur/
- Rajashekar's database http://live.ece.utexas.edu/research/doves/
Different methods are used to assess the degree of similarity between the ground truth and the prediction:

- **Saliency-map-based method:**
  
  \[\begin{align*}
    \text{(a) Original} & \quad \text{(b) Exp. SM} & \quad \text{(c) Predicted SM} \\
  \end{align*}\]

  \[\text{→ ROC (Receiver Operating Characteristic). Each pixel is labeled as fixated or not. Several thresholds are used (AUC (Area Under Curve)).}\]

  \[\text{The higher the AUC, the better is the prediction, with 0.50 indicating random performance and 1.00 denoting perfect performance.}\]

  \[\text{→ KL-Divergence, Linear correlation coefficient...}\]
Fixation point-based method:

→ NSS (Normalized Scanpath salience) gives the degree of correspondence between human fixation locations and predicted saliency maps [Parkhurst et al., 2002], [Peters et al., 2005].

1. Each saliency map is normalized to have zero mean and one unit standard deviation.

2. Extraction of the predicted saliency at a given human fixation point.

3. Average of the previous values.

NSS = 0: random performance; 
NSS >> 0: correspondence between human fixation locations and the predicted salient points; 
NSS << 0: anti-correspondence.
Mostly inherent to the building of the ground truth and to the experimental setting...

1. Several parameters can have a significant impact on the results:
   - the task subjects performed. What is the question we should asked?
   - the nature of the stimuli viewed
   - the apparatus used to record the eye movements
   - a significant central bias
   - cognitive constraints (higher-level goals, prior knowledge, expectations...)

2. Has the fixation the same meaning?
Limitations and perspectives

The assessment of computational model mainly rests on the analysis of individual fixations. But...

- Is there a focal-ambient dichotomy? [Unema et al., 2005][Follet et al., 2009].

- Does every fixation convey the same processing?

- Would it be possible to categorize fixations as attentional fixations, semantic fixations...?

A promising solution to disentangle different processes is called the EFRP (Eye-Fixation-Related Potentials) [Baccino et al., 2005].

Measuring ERP (Event-Related Potential) and EM conjointly to track the cognitive processes.
1 Introduction

2 Computational models: two schools of thought

3 A survey of computational models of bottom-up visual attention

4 Performance assessments

5 Applications
   • Retargeting
   • Compression
   • Others

6 Conclusion
Applications

- Retargeting (or reframing):
  → Principle [Fan et al., 2003] [Chamaret et al., 2008]:

→ Examples:

<table>
<thead>
<tr>
<th>A conventional thumbnail view</th>
<th>A saliency-based thumbnail view</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Image 1" /></td>
<td><img src="image2.png" alt="Image 2" /></td>
</tr>
<tr>
<td><img src="image3.png" alt="Image 3" /></td>
<td><img src="image4.png" alt="Image 4" /></td>
</tr>
<tr>
<td><img src="image5.png" alt="Image 5" /></td>
<td><img src="image6.png" alt="Image 6" /></td>
</tr>
<tr>
<td><img src="image7.png" alt="Image 7" /></td>
<td><img src="image8.png" alt="Image 8" /></td>
</tr>
<tr>
<td><img src="image9.png" alt="Image 9" /></td>
<td><img src="image10.png" alt="Image 10" /></td>
</tr>
<tr>
<td><img src="image11.png" alt="Image 11" /></td>
<td><img src="image12.png" alt="Image 12" /></td>
</tr>
<tr>
<td><img src="image13.png" alt="Image 13" /></td>
<td><img src="image14.png" alt="Image 14" /></td>
</tr>
<tr>
<td><img src="image15.png" alt="Image 15" /></td>
<td><img src="image16.png" alt="Image 16" /></td>
</tr>
<tr>
<td><img src="image17.png" alt="Image 17" /></td>
<td><img src="image18.png" alt="Image 18" /></td>
</tr>
<tr>
<td><img src="image19.png" alt="Image 19" /></td>
<td><img src="image20.png" alt="Image 20" /></td>
</tr>
<tr>
<td><img src="image21.png" alt="Image 21" /></td>
<td><img src="image22.png" alt="Image 22" /></td>
</tr>
<tr>
<td><img src="image23.png" alt="Image 23" /></td>
<td><img src="image24.png" alt="Image 24" /></td>
</tr>
<tr>
<td><img src="image25.png" alt="Image 25" /></td>
<td><img src="image26.png" alt="Image 26" /></td>
</tr>
<tr>
<td><img src="image27.png" alt="Image 27" /></td>
<td><img src="image28.png" alt="Image 28" /></td>
</tr>
<tr>
<td><img src="image29.png" alt="Image 29" /></td>
<td><img src="image30.png" alt="Image 30" /></td>
</tr>
<tr>
<td><img src="image31.png" alt="Image 31" /></td>
<td><img src="image32.png" alt="Image 32" /></td>
</tr>
<tr>
<td><img src="image33.png" alt="Image 33" /></td>
<td><img src="image34.png" alt="Image 34" /></td>
</tr>
<tr>
<td><img src="image35.png" alt="Image 35" /></td>
<td><img src="image36.png" alt="Image 36" /></td>
</tr>
</tbody>
</table>

From [Le Meur et al., 2006].
Applications

- Video compression:
  - To allocate more bit rate to the salient areas than to others (adaptive quantization). Below, the macroblock cost distribution for a H.264 coding and for a saliency-based H.264 coding:

  ![Image of video compression examples]

  - A spatial blur is applied to the input frames such that the non regions of interest are strongly blurred [Itti, 2004].

  ![Image of spatial blur examples]

  From [Itti, 2004].
Applications

- Quality assessment: to weight the distortion of an area by its level of interest to have a better prediction of the quality score [Ninassi et al., 2007].

- Structured document evaluation:
  
  (a) Original images from websites
  
  (b) Mouse-tracking map (From [Mancas, 2009])

- Others: robot navigation, super-resolution, advertising...
Introduction

Computational models: two schools of thought

A survey of computational models of bottom-up visual attention

Performance assessments

Applications

Conclusion
Eye movements are the results of a multiple source guidance [Henderson, 2003].


