Dynamic Visual Saliency Modeling Based on Spatiotemporal Analysis

Duan-Yu Chen, Hsiao-Rong Tyan, Dun-Yu Hsiao, Shen-Wen Shih, and Hong-Yuan Mark Liao
Outline

- Introduction
- Related Works
- Proposed Approach
  - Selective Visual Attention Modeling
    - Detect Spatiotemporal Salient Points
    - Compute Motion Attention Map
    - Determine the Extent of Attended Regions
- Experiment Results
- Conclusion
Introduction

- Methods for annotating video content as well as related video retrieval techniques have attracted a great deal of attention in recent years.
- Video content contains much richer information than other medias and thus can be annotated for applications, such as video surveillance and entertainment.
- Detecting visual saliency from a video is considered a good way of understanding or annotating the video content.
Introduction (cont.)

- The idea is from the father of American psychology, William James in 1980, who suggests that subjects selectively direct attention to objects in a scene using both bottom-up, image-based saliency cues and top-down, task-dependent cues.
A Typical Model for the Control of Bottom-up Attention

Itti’01
Introduction (c.1)

- Motion attention models can be classified into two categories
  - Motion Estimation
    - E.g. block-based matching for motion vector estimation
  - Structural tensor-based approach
Motivation – using *Structural Tensor*

- **Spatio-temporal image data** contains rich information
- Traditional methods for video analysis include
  - *optical flow estimation*
  - *tracking of features/models over time*
- Observation:
  - Events in video are frequently characterized by non-constant motion and non-constant appearance

⇒  Emphasize local structures by *spatio-temporal analysis over a set of consecutive video frames.*

- Temporal Attention Model
  - Use SIFT (Scale Invariant Feature Transformation) to compute correspondences between points in frame pairs.

(a) Image 1  (b) Image 2  (c) Motion Regions
Temporal Attention Model

Use RANSAC to determine $H_1, H_2, H_3$

H: Homography

$H_1: 1 \rightarrow 1'$

$H_2: 2 \rightarrow 2'$

$H_3: 3 \rightarrow 3'$
Temporal Salience Maps

Example

(a) Image 1
(b) Image 2
(c) Point Correspondences
(d) Temporal Saliency Map

- Spatiotemporal Attention Modeling
  - *Center-surround pyramids* are used in motion attention detection to produce salience maps with different precisions.

![Image of motion saliency estimation](image_url)

Fig. 3. An example of the motion saliency estimation. The first column shows the original images in the video sequence. Columns from 2 to 6 are the generated motion maps with center-surround scales of 6-3, 7-3, 7-4, 8-4, and 8-5 respectively. The last column shows the combined motion saliency map. In this example, the camera tracks the walking person. While the background has intensive motions, the person’s motion is greatly canceled by the camera motion. As shown in the maps, the person has a motion salient to his intermediate background. The brighter color indicates higher attention values in the attention map.
Center-Surround pyramids

- The weighting is realized by a Gaussian kernel.
- $\nabla g = (g_x, g_y, g_t)$ is the space-time gradients of the intensity.

Given the spatiotemporal structural tensors $M_1$ and $M_2$ of a local neighborhood $\Omega_1$ and a background $\Omega_2$ respectively,

$$M^* = \tilde{M}_{12}^T \tilde{M}_{12} = \begin{bmatrix} \tilde{M}_1^T \tilde{M}_1 + \tilde{M}_2^T \tilde{M}_2 \end{bmatrix}_{3x3}$$
Compute Salience Maps using Different Visual Cues

Fig. 5. Three examples of adaptive fusion with N=10. Column (a) is the key frame of the given sequence. (b) shows the temporal attention map generated from the two neighboring frames of the sequence. (c)–(e) are the spatial attention maps using color, intensity and orientation respectively. (f) is the combined spatiotemporal attention map. (g) shows the generated FOA in the original video frame.
T. Liu, et al., “Learning to Detect A Salient Object,” CVPR’07

- Detect A Salient Object in Still Images
Static Visual Saliency – color, intensity, orientation

shape from: SaliencyMap

1
2
3
Proposed Approach

☐ Goal

- To design a measure that better captures the extent of motion saliency without using the clue of spatial saliency, even in a dynamic cluttered background.

☐ Visual Saliency Maps Integrations

- Detect *Salient Points* to determine the rough boundary of a moving target
- Detect *Salient Regions* with consistent motion to determine the regions inside the estimated boundary
Salient points in space

- (Harris and Stephens 1988): image points with high variation of values in both image direction

⇒ High eigenvectors of the second-moment matrix integrated at the local neighbourhood

\[
\mu^s = g^s(\cdot; \sigma_i^2) * \begin{pmatrix}
(L_x^s)^2 & L_x^s L_y^s \\
L_x^s L_y^s & (L_y^s)^2
\end{pmatrix}
\]

where \( L_x, L_y \) are Gaussian derivatives

\[
L_x^s(\cdot; \sigma_i^2) = \partial_x (g^s * f^s),
L_y^s(\cdot; \sigma_i^2) = \partial_y (g^s * f^s).
\]

⇒ Select points with maxima of the corner function

\[
\text{ICME } H^s = \det(\mu^s) - k \text{ trace}^2(\mu^s) = \lambda_1 \lambda_2 - k(\lambda_1 + \lambda_2)^2.
\]
Interest points in \textit{space-time}

- High variation of image values in both \textit{space} and \textit{time}

⇒ extend Harris corner function into 3D spatio-temporal domain; compute the second moment matrix

\[ \mu = g(\cdot; \sigma_i^2, \tau_i^2) \ast \begin{pmatrix} L_x^2 & L_xL_y & L_xL_t \\ L_xL_y & L_y^2 & L_yL_t \\ L_xL_t & L_yL_t & L_t^2 \end{pmatrix} \]

where \( L_x, L_y, L_t \) are Gaussian derivatives in space-time obtained by spatio-temporal convolution:

\[ L_\xi(\cdot; \sigma_i^2, \tau_i^2) = \partial_\xi(g(\cdot; \sigma_i^2, \tau_i^2) \ast f) \]

with

\[ g(x, y, t; \sigma_i^2, \tau_i^2) = \frac{\exp\left(-\frac{x^2 + y^2}{2\sigma_i^2} - \frac{t^2}{2\tau_i^2}\right)}{\sqrt{(2\pi)^3\sigma_i^4\tau_i^2}} \]
Salient region in *space-time*

- Function of Visual Saliency

\[ H = \det(\mu) - k \text{trace}^3(\mu) = \lambda_1 \lambda_2 \lambda_3 - k(\lambda_1 + \lambda_2 + \lambda_3)^3, \]

where \( \lambda_1, \lambda_2, \lambda_3 \) are eigenvalues of \( \mu \).
Problem of only detecting salient regions based on 3D corner detection

- Broken parts of moving targets

- Why results in broken parts?
  - Regions inside a moving target are relatively smoothly changed within X YT video volumes.
Solution

- Smoothly changed regions physically mean that these regions are moving with consistent motions
- What is the motion cue?
  - The rank of structural tensor

\[
\mu = g(\cdot; \sigma_i^2, \tau_i^2) \ast \begin{pmatrix}
L_x^2 & L_xL_y & L_xL_t \\
L_yL_x & L_y^2 & L_yL_t \\
L_tL_x & L_tL_y & L_t^2
\end{pmatrix}
\]

- \( \text{Rank}(\mu) = 3 \): multiple motions
- \( \text{Rank}(\mu) = 2 \): one dominant motion
Solution (c.1)

- Noise and FG/BG movements in video sequences would result in multiple motions within a neighborhood.
  - \( \text{Rank}(\mu) = 3 \) in almost all video volumes
- A new measure of rank deficiency
  - \( d_{ST} = \frac{\lambda^2}{0.5 \cdot \lambda^2_1 + 0.5 \cdot \lambda^2_2 + \varepsilon} \)
Median Filter for $d_{ST}$

- Goal: to drop off the regions with multiple motions and only keep regions with consistent motion

- Use median filter to filter out two extreme cases
  - $\text{rank}(\mu) = 1$ and $\text{rank}(\mu) = 3$
Demo – 1WP with more complex background
The most appropriate scale $S_e$ for each region centered at the seed $(x_s, y_s)$ of a motion attention map is defined by

$$S_e = \arg \max_e \{ H_{\text{exp}}(e, (x_s, y_s), t) \times W_{\text{exp}}(e, (x_s, y_s), t) \}$$

$$H_{\text{exp}}(e, (x_s, y_s), t) = \sum_{v \in D} p_{v,e,(x_s,y_s),t} \times \exp(1 - p_{v,e,(x_s,y_s),t})$$

$$W_{\text{exp}}(e, (x_s, y_s), t) = |H_{\text{exp}}(e, (x_s, y_s), t) - H_{\text{exp}}(e - 1, (x_s, y_s), t)|$$
Experimental Results

- Qualitative Evaluation
- Quantitative Evaluation – Precision & Recall
Demo
Demo (c.1)
Demo (c.2)
Performance Comparison

Performance Comparison
Performance Comparison

Quantitative Evaluation
Quantitative Evaluation (cont.)

![Recall vs Frame Graph](image)

- **Ours**
- **Zhai**
- **Ours Mean**
- **Zhai Mean**
## Quantitative Evaluation (cont.)

<table>
<thead>
<tr>
<th></th>
<th>Our method</th>
<th>Zhai and Shah[3]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>71.8%</td>
<td>57.9%</td>
</tr>
<tr>
<td>Recall</td>
<td>80.6%</td>
<td>53.8%</td>
</tr>
</tbody>
</table>
Conclusion and Future Work

- Visual Salient Maps Integrations
  - Detect *Salient Points* to determine the boundary of a moving target
  - Detect the appropriate *Extent of Salient Regions* in the motion attention map and refine using the *exponential entropy-based center-surround approach*.

- Future Work
  - Annotate the extent of salient regions automatically in general video sequences.
Object Annotation

- Hierarchical Filtering
  - Orientation Filters

Examples
Thank You!
(b) Image 2

(d) Temporal Saliency Map