Video Saliency Detection via Dynamic Consistent Spatio-Temporal Attention Modelling

Sheng-hua ZHONG¹, Yan LIU¹, Feifei REN¹,², Jinhuan ZHANG², Tongwei REN³

¹Department of Computing, The Hong Kong Polytechnic University
²School of Psychology, Shandong Normal University
³Software Institute, Nanjing University
Outline

- Introduction to video saliency detection
- Spatio-temporal attention technique
- Experiments and results
- Conclusion and future work
Outline

- Introduction to video saliency detection
- Spatio-temporal attention technique
- Experiments and results
- Conclusion and future work
Introduction to Saliency Map

- Definition of saliency map
  - The most famous attention model, referred to as the allocation of processing resources
  - Measure of conspicuity and calculate the likelihood of a location to attract attention [Koch et. al, Hum Neurobiol, 1985]

- Motivation of constructing saliency map
  - Provide predictions about which regions are likely to attract observers’ attention
  - Be useful to image/video representation (Wang et al. ICME, 2007), object detection and recognition (Yu et al. ACM, 2010), object tracking (Yilmaz et al. CSUR, 2006), and robotics controls (Jiang & Crookes, AAAI, 2012)
Video Saliency Detection

- Definition of video saliency map
  - Calculate the salient degree of each location both in spatial and in temporal areas [Li et al. AAAI, 2012]
  - Not much work has been extended to video sequences where motion plays an important role

- Two pathways simulation
  - Video saliency detection procedure are divided into spatial and temporal channels [Marat et al., IJCV, 2009] corresponding to the magnocellular and parvocellular pathways
  - Classical optical flow model is the most widely used motion detection approaches in video saliency detection

- Classical optical flow model in saliency detection
  - The independent calculation of each frame pair leads to high computational complexity
  - The continuous motion of the prominent object cannot be popped out
Outline

- Introduction to video saliency detection
- Spatio-temporal attention technique
- Experiments and results
- Conclusion and future work
Video Saliency Detection via Dynamic Consistent Spatio-Temporal Attention Modelling
Temporal Saliency Map Construction

- Basic idea
  - Emphasize the dynamic continuity of neighbor locations in the same frame
  - Emphasize the dynamic continuity of same locations in the temporal domain

$$\arg\min_{u, v, n} E(u, v, \hat{u}, \hat{v})$$

$$= \sum_{i,j} \left\{ f_D(\sum_{k \leq n} (I_m(i, j) - I_{m+k}(i + ku_{i,j}, j + kv_{i,j})) \right\}$$

$$+ \lambda_1[f_S(u_{i,j} - u'_{i+1,j}) + f_S(u_{i,j} - u'_{i,j+1}) + f_S(v_{i,j} - v'_{i+1,j}) + f_S(v_{i,j} - v'_{i,j+1})]$$

$$+ \lambda_2(\|u - \hat{u}\| + \|v - \hat{v}\|) + \sum_{i,j} \sum_{(i^*, j^*) \in N_{i,j}} \lambda_3(\left|\hat{u}_{i,j} - \hat{u}'_{i^*, j^*}\right| + \left|\hat{v}_{i,j} - \hat{v}'_{i^*, j^*}\right|)$$

s.t. \( u_{i,j}^2 + v_{i,j}^2 \leq \sigma_o(i^*, j^*, i^*, j^*) = \arg\max_{i,j} (u_{i,j}^2 + v_{i,j}^2) \)
Dynamic Consistent Saliency Detection

Video Saliency Detection via Dynamic Consistent Spatio-Temporal Attention Modelling
Spatial Saliency Map and Spatio-Temporal Saliency Fusion

- Spatial saliency map construction
  - Extraction: multiple low-level visual features are extracted at multiple scales
  - Activation: activation maps are built based on multiple low-level feature maps
  - Normalization: saliency map is constructed by a normalized combination of the activation map

- Spatio-temporal saliency fusion
  - Different fusions methods can be utilized, such as “mean” fusion, “max” fusion, and “multiplicative” fusion
  - “Max” integration method has best performance [Marat et al., IJCV, 2009]

\[
FSmap = \max(SSmap, TSmap)
\]

**FSmap**: Spatio-temporal saliency map
**SSmap**: Spatial saliency map
**TSmap**: Temporal saliency map
Outline

- Introduction to video saliency detection
- Spatio-temporal attention technique
- Experiments and results
- Conclusion and future work
Experiment Setting

Datasets
- Hollywood2 natural dynamic human scene videos dataset [Marszallek et al., CVPR, 2009]
  - Ten different natural environments, including: house, road, bedroom and so on
- Three typical CNN Headline news videos
  - Each video clip is approximately 30 seconds and the frame rate is 30 frames/second
  - Resolution is $640 \times 360$
- Subset of the largest real world actions video dataset with human fixations
  - 12 categories, 884 videos clips, including: answering phone, driving car, eating and so on
  - 16 subjects’ fixations
  - First 5 video clips from every category

Compared algorithms
- Temporal saliency detection models
  - Classical optical flow model (COF) [Horn & Schunck, AI, 1981] [Black & Anandan, CVIU, 1996]
  - Spatial continuous optical flow model (SOF) [Sun et al., CVPR, 2010]
- Spatio saliency detection models
  - Itti saliency model (Itti) [Itti et al., PAMI, 1998], graph based saliency map (GBVS) [Harel et al., NIPS, 2007]
Experiments on Natural Dynamic Scene Videos

- Dataset
  - Hollywood2 natural dynamic human scene videos dataset

- Experiments on face detection
  - Higher level visual cortex regions influence the human’s attention in a top-down manner;
  - Humans often fixate on people and face; Face detection region is often added into saliency map as a high level feature [Judd et al., NIPS 2009] [Mathe & Sminchisescu, ECCV, 2012]
Face Saliency Detection on Natural Dynamic Scene Videos

Table 1. Face saliency detection on natural dynamic scene videos

<table>
<thead>
<tr>
<th>Model</th>
<th>Average Saliency Value</th>
<th>Average Detection Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>DCOF</td>
<td>0.6501</td>
<td>0.8252</td>
</tr>
<tr>
<td>COF</td>
<td>0.6018</td>
<td>0.7537</td>
</tr>
<tr>
<td>SOF</td>
<td>0.6393</td>
<td>0.7782</td>
</tr>
</tbody>
</table>

(a) Original frame image  
(b) Face detection result

(c) Saliency map of DCOF  
(d) Saliency map visualization

(e) Saliency map of SOF  
(f) Saliency map visualization

Video Saliency Detection via Dynamic Consistent Spatio-Temporal Attention Modelling
Experiments on News Headline Videos

- Dataset
  - Three typical CNN Headline news videos

- Compared algorithms
  - Temporal saliency detection models COF, SOF

- Experiments
  - Efficiency comparison
  - Effectiveness comparison

Table. Efficiency comparison on the news headline videos

<table>
<thead>
<tr>
<th>Model</th>
<th>DCOF</th>
<th>COF</th>
<th>SOF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Running Time per Frame (s)</td>
<td>33.12</td>
<td>46.24</td>
<td>53.88</td>
</tr>
<tr>
<td>Output Frame Ratio</td>
<td>0.4</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Experiments on News Headline Videos

Figure. Temporal saliency detection result.
Experiments on Eye-tracking Action Videos

- Dataset: Largest real world actions video dataset with human fixations

- Compared algorithms
  - Temporal saliency detection models COF, SOF
  - Spatio-temporal saliency detection models (Itti, GBVS)+ (COF, SOF)

- Two experiments
  - Average receiver operating characteristic (ROC) areas
  - Average receiver operating characteristic (ROC) curves

Sample video with eye-tracking fixations
The area under the ROC curve to demonstrate the performance of a saliency model

<table>
<thead>
<tr>
<th>ROC Area</th>
<th>DCOF</th>
<th>COF</th>
<th>SOF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Answer phone</td>
<td>0.6098</td>
<td>0.5303</td>
<td>0.5910</td>
</tr>
<tr>
<td>Drive car</td>
<td>0.5233</td>
<td>0.4817</td>
<td>0.5195</td>
</tr>
<tr>
<td>Eat</td>
<td>0.6902</td>
<td>0.6598</td>
<td>0.6644</td>
</tr>
<tr>
<td>Fight</td>
<td>0.6045</td>
<td>0.5535</td>
<td>0.6005</td>
</tr>
<tr>
<td>Get out car</td>
<td>0.5260</td>
<td>0.4874</td>
<td>0.5212</td>
</tr>
<tr>
<td>Hand shake</td>
<td>0.6993</td>
<td>0.6485</td>
<td>0.6934</td>
</tr>
<tr>
<td>Hug</td>
<td>0.6402</td>
<td>0.5602</td>
<td>0.5996</td>
</tr>
<tr>
<td>Kiss</td>
<td>0.5833</td>
<td>0.5120</td>
<td>0.5503</td>
</tr>
<tr>
<td>Run</td>
<td>0.5535</td>
<td>0.5104</td>
<td>0.5496</td>
</tr>
<tr>
<td>Sit down</td>
<td>0.5183</td>
<td>0.4761</td>
<td>0.5074</td>
</tr>
<tr>
<td>Sit up</td>
<td>0.5171</td>
<td>0.4871</td>
<td>0.5006</td>
</tr>
<tr>
<td>Stand up</td>
<td>0.5602</td>
<td>0.5269</td>
<td>0.5601</td>
</tr>
</tbody>
</table>
ROC Curve Comparison

- ROC curve is plotted as the False Positive Rate vs. Hit Rate
Outline

- Introduction to video saliency detection
- Spatio-temporal attention technique
- Experiments and results
- Conclusion and future work
Conclusion and Future Work

Conclusion

- Emphasize the dynamic consistency of neighbor locations in the same frame and same locations in the temporal domain
- Effective prominent object detection and coverage
- Better efficiency and less storage space

Future work

- Jointly optimize the spatial and temporal saliency detection together
Reference

Reference

Q & A

Thank You!