# Video Saliency Detection via Dynamic Consistent Spatio-Temporal Attention Modelling

*Sheng-hua ZHONG[1], Yan LIU[1], Feifei REN[1,2], Jinghuan ZHANG[2], Tongwei REN[3]*

[1]*Department of Computing, The Hong Kong Polytechnic University*

[2]*School of Psychology, Shandong Normal University*

[3]*Software Institute, Nanjing University*

# Outline

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

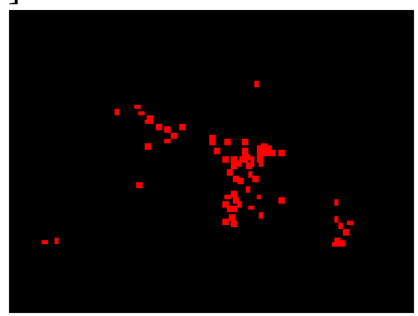
# Outline

- 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]

(a) original image



(b) Eye- fixation locations

(c) Ground truth of saliency map

- 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. ACMMM, 2010), object tracking (Yilmaz et al. CSUR, 2006), and robotics controls (Jiang & Crookes, AAAI, 2012)

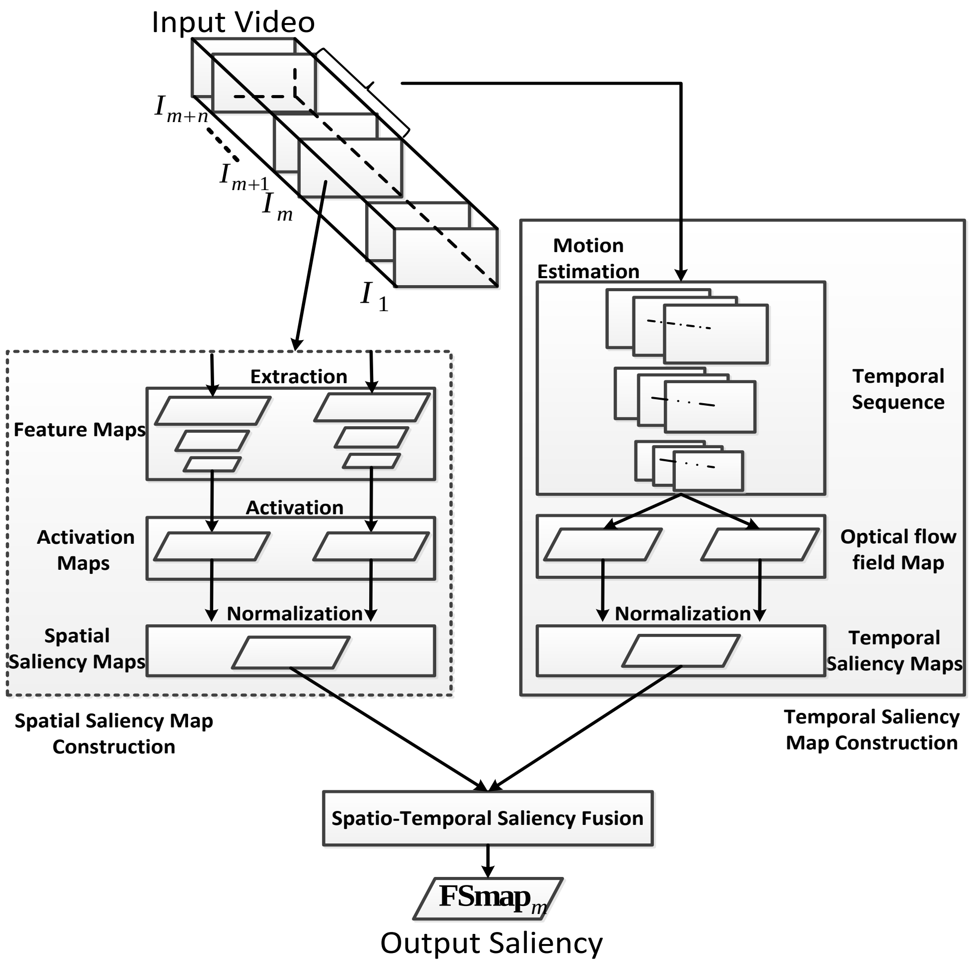
# Video Saliency Detection

- 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

# Framework of Spatio-temporal Attention Technique



# 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

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

$$=\sum_{i,j}\{ f_D[\sum_{k\le n}(I_m(i,j)-I_{m+k}(i+ku_{i,j}, j+kv_{i,j}))]$$

**Dynamic continuity of same locations in temporal domain**

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

**Dynamic continuity of neighbor locations in same frame**

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

**Smoothness within a region in the auxiliary flow field**

$$s.t.\ \sqrt{u_{i^*,j^*}{}^2+v_{i^*,j^*}{}^2}\le\sigma_o(i^*,j^*),(i^*,j^*)=\arg\max_{i,j}(\sqrt{u_{i,j}{}^2+v_{i,j}{}^2})$$

**Constraint based on the observation standard deviation of human visual system**

# Dynamic Consistent Saliency Detection

**Algorithm 1:** Dynamic Consistent Saliency Detection

**Input:** Video sequence data **X** ; Number of frames $N_f$; Pyramid level $N_p$.

**Output:** Temporal saliency map **TSmap**.

1. **while** $m < N_f - n$
2. $p = 1$;
3. **while** $p < N_p + 1$ **do**
4. $(\mathbf{u}, \mathbf{v}) = \arg\min\ E(\mathbf{u}, \mathbf{v}, \hat{\mathbf{u}}, \hat{\mathbf{v}})$ ;
5. $(i^*, j^*) = \arg\max_{i,j}(\sqrt{u_{i,j}{}^2 + v_{i,j}{}^2})$ ;
6. **if** $\sqrt{u_{i^*,j^*}{}^2 + v_{i^*,j^*}{}^2} > \sigma_o(i^*, j^*) / 2^{p-1}$ **and** $n > 1$
7. $n = \max(n - 1, 1); p = 1$;
8. **else**
9. $p = p + 1$;
10. **end if**
11. **end while**
12. $\mathbf{TSmap}_m(i, j) = \text{normalize}(\sqrt{u_{i,j}{}^2 + v_{i,j}{}^2})$
13. $m = m + n$;
14. **end while**

# 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]

$$\mathbf{FSmap} = \max(\mathbf{SSmap}, \mathbf{TSmap})$$

**SSmap**: Spatial saliency map ; **TSmap**: Temporal saliency map;
**FSmap**: Spatio-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 optial 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

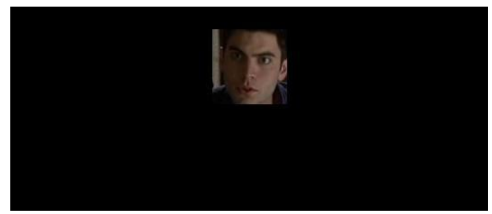
- Dataset
  - Hollywood2 natural dynamic human scene videos dataset
- Experiments on face detection
  - Higher level visual cortex regions influence the human's attention in a top-down manner;
  - Humans often fixate on people and face; Face detection region is often added into saliency map as a high level feature [Judd et al., NIPS 2009] [Mathe & Sminchisescu, ECCV, 2012]

# Face Saliency Detection on Natural Dynamic Scene Videos

Table . Face saliency detection on natural dynamic scene videos

| Model | Average Saliency Value | Average Detection Accuracy |
|---|---|---|
| **DCOF** | **0.6501** | **0.8252** |
| **COF** | 0.6018 | 0.7537 |
| **SOF** | 0.6393 | 0.7782 |

(a) Original frame image

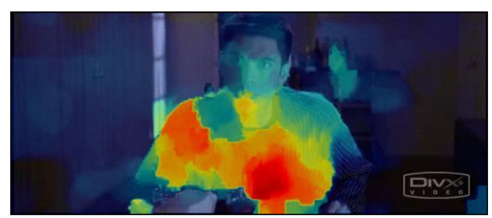


(b) Face detection result

(c) Saliency map of DCOF

(d) Saliency map visualization

(e) Saliency map of SOF



(f) Saliency map visualization

# 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

| **Model** | DCOF | COF | SOF |
|---|---|---|---|
| **Running Time per Frame (s)** | **33.12** | 46.24 | 53.88 |
| **Output Frame Ratio** | **0.4** | 1 | 1 |

# Experiments on News Headline Videos

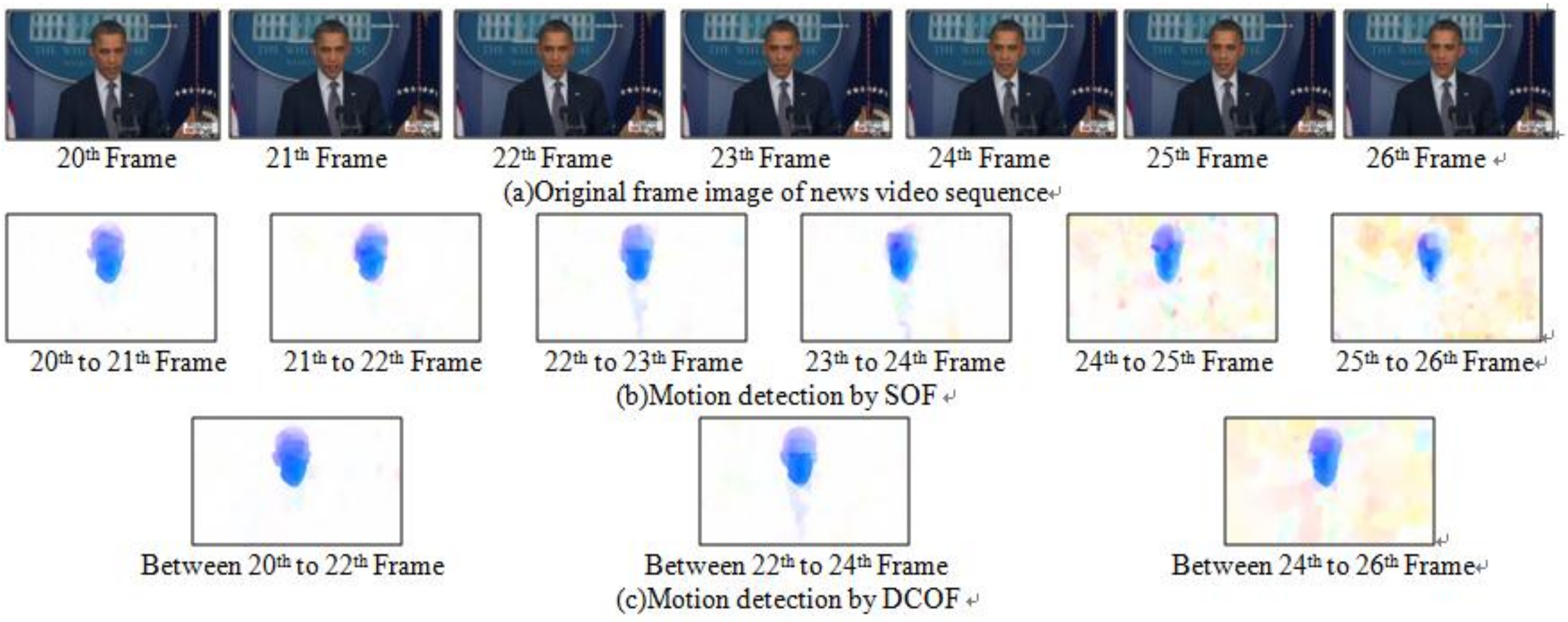



Figure. Temporal saliency detection result.

# Experiments on Eye-tracking Action Videos

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

**Sample video with eye-tracking 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

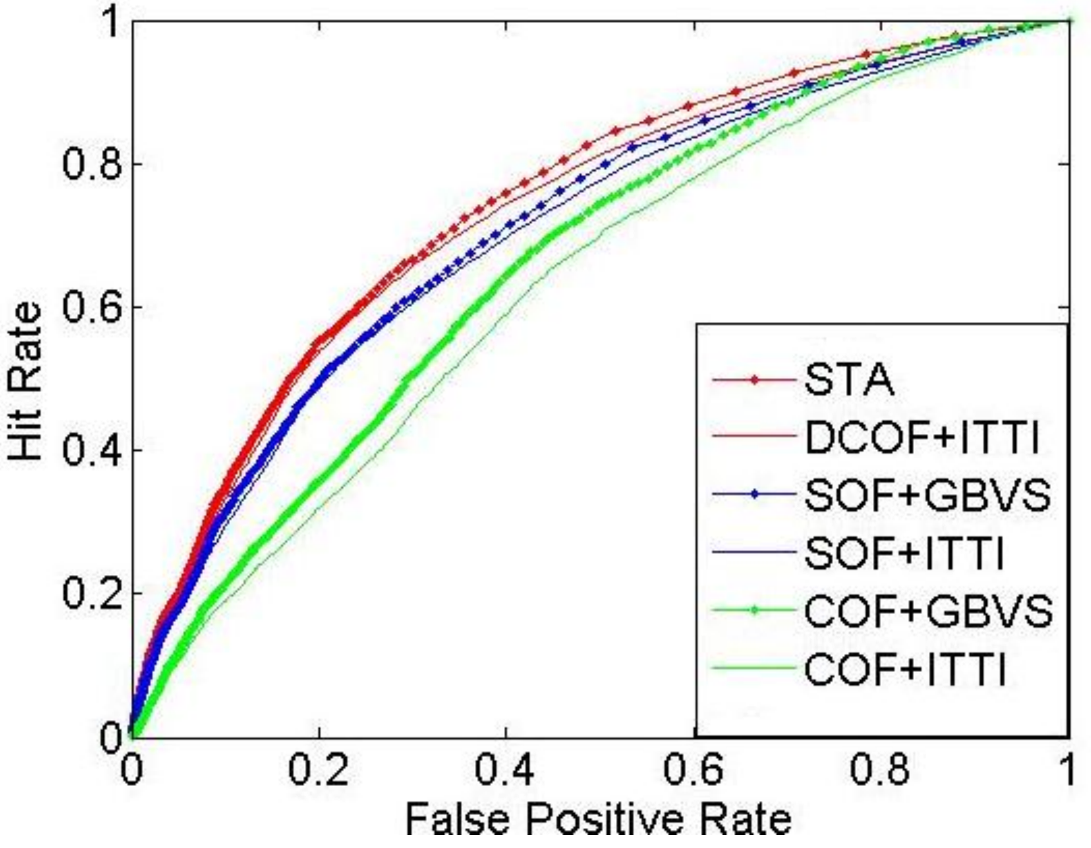
# ROC Area Comparison

- The area under the ROC curve to demonstrate the performance of a saliency model

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

# 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


- [Black, M. J. & Anandan, P. 1996. The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields. In *Computer Vision and Image Understanding*, 63, pages 75–104.
- Harel, J., Koch, C., and Perona, P.. 2007. Graph-based visual saliency. In Proceedings of the 20th Annual Conference on Neural Information Processing Systems (NIPS), pages 545-552.
- Horn, B. & Schunck, B. 1981. Determining optical flow. In *Artificial Intelligence*, 16, pages 185–203.
- Jiang, J.R. & Crookes, D.. 2012. Visual saliency estimation through manifold learning. In Proceedings of the 26th AAAI Conference on Artificial Intelligence, pages 2003-2010.
- Judd, T., Ehinger, K., Durand, F. and Torralba, A.. 2009. Learning to predict where humans look. In Proceedings of the IEEE 12th International Conference on Computer Vision (ICCV), pages 2106-2113.
- Itti, L., Koch, C. and Niebur, E.. 1998. A model of saliency-based visual attention for rapid scene analysis. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11), pages 1254-1259.
- Koch, C. & Ullman, S.. 1985. Shifts in selective visual attention: Towards the Underlying Neural Circuitry. In *Human Neurobiology*, pages 219-227.
- Li, B., Xiong, W.H., Hu, W.M.. 2012. Visual saliency map from tensor analysis. In Proceedings of the 26th AAAI Conference on Artificial Intelligence, pages 1585-1591.
- Marszałek, M., Laptev, I., and Schmid, C.. 2009. Actions in Context. In Proceedings of the 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR).
- Marat, S., Ho Phuoc, T., Granjon, L., Guyader, N., Pellerin, D., Guérin-Dugué A.. 2009. Modelling spatio-temporal saliency to predict gaze direction for short videos. In *International Journal of Computer Vision*, 82(3), pages 231–243.
- Mathe, S. & Sminchisescu, C.. 2012. Dynamic Eye Movement Datasets and Learnt Saliency Models for Visual Action Recognition, In Proceedings of the 9th European Conference on Computer Vision (ECCV), pages 842-856.
- Sun, D., Roth, S., Black, M. J.. 2010. Secrets of Optical Flow Estimation and Their Principles. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR).

# Reference


- Wang, T., Mei, T., Hua, X.-S., Liu, X., and Zhou, H.-Q. 2007. Video collage: A novel presentation of video sequence. In Proceedings of the 2009 IEEE International Conference on Multimedia and Expo (ICME), pages 1479--1482.
- Yilmaz, A., Javed, O., and Shah, M.. 2006. Object Tracking: A Survey, In *ACM Computing Surveys*, 38(4), pages 13.
- Yu, H., Li, J., Tian, Y., Huang, T., 2010. Automatic interesting object extraction from images using complementary saliency maps, In Proceedings of the 18th ACM International Conference on Multimedia (ACMMM), pages 891-894.

# Q & A

*Thank You !*